Pointing out the Shortcomings of Relation Extraction Models with Semantically Motivated Adversarials

Gennaro Nolano, [1]**Moritz Blum,** [1,2]**Basil Ell,** [1]**Philipp Cimiano**
[1]Bielefeld University, [2]University of Oslo
Germany, Norway
{mblum, bell, cimiano}@techfak.uni-bielefeld.de, nolanogenn@gmail.com

Abstract

In recent years, large language models have achieved state-of-the-art performance across various NLP tasks. However, investigations have shown that these models tend to rely on shortcut features, leading to inaccurate predictions and causing the models to be unreliable at generalization to out-of-distribution (OOD) samples. For instance, in the context of relation extraction (RE), we would expect a model to identify the same relation independently of the entities involved in it. For example, consider the sentence ”*Leonardo da Vinci painted the Mona Lisa*” expressing the `created(Leonardo_da_Vinci, Mona_Lisa)` relation. If we substiute ”*Leonardo da Vinci*” with ”*Barack Obama*”, then the sentence still expresses the `created` relation. A robust model is supposed to detect the same relation in both cases.

In this work, we describe several semantically-motivated strategies to generate adversarial examples by replacing entity mentions and investigate how state-of-the-art RE models perform under pressure. Our analyses show that the performance of these models significantly deteriorates on the modified datasets (avg. of -48.5% in $F1$), which indicates that these models rely to a great extent on shortcuts, such as surface forms (or patterns therein) of entities, without making full use of the information present in the sentences.

Keywords: Shortcut Learning, Adversarial Evaluation, Relation Extraction

1. Introduction

Recent large language models have achieved state-of-the-art performance on many NLP tasks, such as Relation Extraction (RE), and are generally considered as the go-to methodology to tackle this sort of task (Ji et al., 2022). These models exploit contextual information extracted from natural language sentences in the form of embeddings to label relations between entity mentions (Baldini Soares et al., 2019; Wu and He, 2019).

Despite achieving generally good results, closer investigations show that these modes' inner workings have yet to be fully understood, and that they are less robust than expected.For instance, the work by Rosenman et al. (2020) highlights that SOTA models for RE seem to exploit simple heuristics. Furthermore, 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 behavior of models fine-tuned for RE in an adversarial setting, showing that they act in unexpected ways when encountering unforeseen situations. In particular, we test the models using several adversarial datasets generated by substituting the entities mentioned in a sentence realizing a specific relation, and by following specific strategies. In doing so, the contextual information between and around the entities, and the relation expressed by the sentence, is left untouched.

For instance, considering the sentence expressing a relation between the entities *Leonardo Da Vinci* (e_{SUBJ}) and the *Mona Lisa* (e_{OBJ}):

> ”<u>*Leonardo Da Vinci*</u> *painted the* <u>*Mona Lisa*</u>.”

Replacing the entity *Leonardo Da Vinci* with another entity, such as *Michelangelo*, *Barack Obama* or *Stratolaunch*, will inevitably change the semantics of the sentence, but a robust RE system should still recognize the original relation to be expressed given that, beyond the entities, the actual context is unchanged.

We hypothesize that, in case a model is not able to predict the correct relation type given such adversarial examples, it is because the model itself has learned entity- and relation-specific heuristics at training time, rather than actually learning how relations are expressed in natural language.

More specifically, given a text example expressing a relation r between entities e_{SUBJ} and e_{OBJ}, we replace e_{SUBJ}, e_{OBJ}, or both, with other entities, according to specific substitution operations.

Our experiments with different types of substitution operations show that the models are significantly misled by such adversarials, reducing performance by an average of 48.5% in *F1*, depending on the model involved and the substitution operations applied.

12809

2. Related Work

Shortcut learning in NLP refers to the process where a model relies on data biases to make predictions. More precisely, the model learns a co-occurrence of certain features and labels, without achieving true language understanding. These features are called shortcut features. Du et al. (2021) demonstrate that NLP models have a strong preference for features located at the head of the long-tailed local mutual information (LMU) (Schuster et al., 2019) distribution of words and labels. If a model learns such shortcut features, it might suffer from low generalization abilities and low adversarial robustness on data that do not share the same shortcut features as the training data.

To explore an NLP model's inclination towards shortcut learning, one approach is to examine how the model behaves when confronted with adversarials (Morris et al., 2020). Other approaches analyze syntactic heuristics (McCoy et al., 2019) or do randomization ablation studies (Sinha et al., 2021).

The core idea behind the use of adversarials to test ML models has its roots in the field of 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 employing different kinds of permutations (Goyal et al., 2022).

Generally speaking, adversarial attacks can be divided into two categories according to Gaiński and Bałazy (2023): white-box methods, which make use of gradients to generate perturbations (Ebrahimi et al., 2018; Cheng et al., 2019; Xu and Du, 2020; Yuan et al., 2023), and black-box methods, which define heuristics to implement input data permutations, e.g. character or word replacements (Yoo et al., 2020; Morris et al., 2020). The present work implements a black-box methodology with heuristics based on the semantics of relations and entity types. For a list of adversarial attack methodologies and techniques for NLP, see (Morris et al., 2020).

Adversarials have been widely used to highlight specific points of weakness in NLP models, 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 points of failure in machine comprehension, visual QA, and sentiment analysis.

Similarly, Hosseini et al. (2017) have shown the effect of symbol addition and typo insertion on the task of toxic language detection, while Kaushik and Lipton (2018) show that reading comprehen-

sion datasets are not heavily affected by the removal of questions. Finally, Belinkov and Bisk (2018) applied such permutations to investigate the robustness of machine translation.

Recently, Formento et al. (2023) proved the effectiveness of character insertion as a form of adversarial attack against several tasks, such as text classification, entailment, and question answering. While these models prove the effects of using adversarials to investigate NLP models, they generally exploit simple surface patterns to generate adversarials (Wallace et al., 2019). A more complex framework in this regard is proposed by Li et al. (2021), who make use of entity-altering permutations to investigate the robustness of *BERT*-based models for RE. However, they only investigate the substitution of entities with entities of the same semantic type, and entity masking. Moreover, they limit their evaluation to 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 more complex types of entity substitutions, while also taking into account the performance of various state-of-the-art models for RE.

3. Methodology

We construct 12 adversarial datasets by changing the semantics of sentences while preserving the underlying relation between entities expressed in the sentence.

We argue that, while information about the entities involved in the relation is relevant to determine whether that relation can exist between the two entities, a robust model should not rely only on information about the entities, such as their semantic types or the gender of a mentioned person, but should make use of the intra-sentence context in which the entities are mentioned.

In general, human readers can identify the relation between two entities as expressed in a natural language sentence by following simple lexico-syntactic patterns (Hearst, 1992), and the same should be expected from a RE model.

We obtain adversarial examples from a sentence that expresses a specific relation (e.g., `created` in the example about the *Mona Lisa* shown in Sec. 1) between a subject entity and an object entity by replacing the mention of the subject or object entity (or both) with the surface form of another entity. Thereby, from a sentence that expresses a relation that is factually true (e.g., *"Leonardo Da Vinci painted the Mona Lisa."*) we may obtain a sentence that is factually incorrect (e.g., *"Leonardo Da Vinci painted the Scream."*). However, the relation expressed (`created`) stays the same, albeit between different entities.

We introduce four substitution strategies which

we apply to create 12 adversarial datasets to test how models fare under different semantically-motivated permutations.

We define our corpus D as a set of quadruples of the form $(d, r, e_{SUBJ}, e_{OBJ})$ where d is a sentence expressing the relation r between the subject entity e_{SUBJ} and the object entity e_{OBJ}. By λ_{SUBJ} and λ_{OBJ} we denote the mention of the subject and object entity in the sentence d, respectively, and by t_{SUBJ} and t_{OBJ} we denote the type of the subject and object entity, respectively.

Given a quadruple $(d, r, e_{SUBJ}, e_{OBJ})$, each substitution strategy creates a new quadruple $(d', r, e'_{SUBJ}, e'_{OBJ})$ where either the original subject entity is replaced by another entity, the original object entity is replaced by another entity, or both are replaced, and where a new sentence d' is obtained by replacing the original surface form of the previously mentioned entity(ies) with the surface form(s) of the new entity(ies).

The substitution strategies are as follows:

- **same-role substitution**

  e'_{SUBJ} is randomly chosen from the set $\{x \mid \exists (d'', r, x, e''_{OBJ}) \in D : x \neq e_{SUBJ}\}$ and $e_{OBJ} = e'_{OBJ}$. e'_{OBJ} can be selected analogously.

  Thus, an entity is replaced with another entity that occurs in the same role (as subject or object) in another sentence expressing the same relation. Either the subject or the object or both are replaced.

- **same-type substitution**

  e'_{SUBJ} is randomly chosen from the set $\{x \mid \exists (d'', r', x, e''_{OBJ}) \in D : r \neq r' \land x \neq e_{SUBJ} \land t_{SUBJ} = t'_{SUBJ}\}$. e'_{OBJ} can be selected analogously.

  Thus, an entity is replaced with another entity of the same type that occurs in the same role (as subject or object) in another sentence expressing another relation. Either the subject or the object or both are replaced.

- **different-type substitution** e'_{SUBJ} is randomly chosen from the set $\{x \mid \exists (d'', r', x, e''_{OBJ}) \in D : r \neq r' \land x \neq e_{SUBJ} \land t_{SUBJ} \neq t'_{SUBJ}\}$. e'_{OBJ} can be selected analogously.

  Thus, an entity is replaced with another entity of a different type that never occurs in the same role (as subject or object) in another sentence expressing the same relation. Either the subject or the object or both are replaced.

- **masking**

  e'_{SUBJ} is replaced with the [MASK] token. We set t'_{SUBJ} to NONE. e'_{OBJ} can be selected anal-

ogously. Either the subject or the object or both are replaced.

We apply each of the four strategies to replace either the subject, the object or both entities to each sentence of the testing set, thus obtaining 12 adversarial datasets from the original data. An original sentence and one adversarial example for each substitution strategy is shown in Table 1.

	Original Sentence		
	Leonardo da Vinci	painted the	Mona Lisa
	Same-role		
subj mod.	Michelangelo	painted the	Mona Lisa
obj mod.	Leonardo da Vinci	painted the	Scream
subj+obj mod.	Michelangelo	painted the	Scream
	Same-type		
subj mod.	Barack Obama	painted the	Mona Lisa
obj mod.	Leonardo da Vinci	painted the	Baloon Girl
subj+obj mod.	Barack Obama	painted the	Baloon Girl
	Diff.-type		
subj mod.	Stratolaunch	painted the	Mona Lisa
obj mod.	Leonardo da Vinci	painted the	Berlin Wall
subj+obj mod.	Stratolaunch	painted the	Berlin Wall
	Masking		
subj mod.	[MASK]	painted the	Mona Lisa
obj mod.	Leonardo da Vinci	painted the	[MASK]
subj+obj mod.	[MASK]	painted the	[MASK]

Table 1: Examples of the adversarial strategies

4. Experiments

4.1. Data

In order to build adversarial datasets and to test the models, the starting point is the human-labelled TACRED dataset (Zhang et al., 2017). The dataset contains 68,124 samples for training, 22,631 samples for evaluation, and 15,509 samples for testing, expressing 42 relation types. The dataset only contains English sentences.

In each sample, both entities are labelled with their semantic type,[1] and linked, when possible, to their corresponding entry in Wikidata.[2]

We use the training split where needed to finetune the models and use the test split to generate the adversarials examples used to test the models.[3]

Some sentences in TACRED contain arguments of relations that cannot be linked to Wikidata. This is, for example, the case when the mention is a personal pronoun. By this process, we focus on sentences where entities are mentioned explicitly and ignore sentences where the entities are only indirectly mentioned / referred to. We discard those sentences in the test split in which at least on argument of the relation in not linked to Wikidata.

---

[1] From the Stanford NER system, see `https://stanfordnlp.github.io/CoreNLP/ner.html`.

[2] Wikidata version accessed *Wikidata 2022-01-03*

[3] The script to create the proposed adversarial datasets from TACRED can be found in our GitHub repository. See `https://github.com/Nolanogenn/adv_re`.

This process removes 9,232 sentences from the test split, thus leaving us with 6,277 sentences that were then used to generate adversarial examples. One of the most peculiar aspects of TACRED is that the relation classes are not equally distributed. In fact, the training set is, by design, heavily skewed towards the no_relation label,[4] which heavily affects models as they seem to default to this class when facing unexpected pairs of entities, as described in Section 5.2.

4.2. Models

The adversarial examples are tested using different SOTA models on TACRED. In particular, the investigated models are: **LUKE** (Yamada et al., 2020), **SpanBERT** (Joshi et al., 2020), **UniST** (Huang et al., 2022), **SuRE** (Lu et al., 2022), **TYP-Marker** (Zhou and Chen, 2022), and **NLI** (Sainz and Rigau, 2021). For LUKE and Span-BERT, the checkpoints are loaded from the HuggingfaceHub,[5,6] while the checkpoints for NLI are downloaded following the instructions found in the repository.[7] Since no weights are publicly available for SuRE and UniST, we finetune these models using the TACRED training set by following the procedures described in the official repositories[8,9] before testing them in the adversarial setting.

For the NLI model, we test both the original model (**NLI_w**), which implements heuristics-based constraints on the possible entity types that can appear as subject/object for a given relation, and a version of the model where these constraints are not implemented (**NLI_w/o**).

5. Results

5.1. General Results

We evaluate the performance of the models under pressure using the *F1* score computed with the official TACRED evaluation script.[10] The *F1* values for every model and each of our adversarial strategies/datasets are shown in Table 2.

The results show that the models are affected by the adversarials, with an average loss of 48.5% in *F1* score. As it turns out, not all adversarial strategies have the same impact on a model's performance: in particular, the *same-role* and

the *same-type* substitution are the least impactful ones, while the *masking* substitution has the strongest impact on the results, which proves that models strongly rely on entity surface forms to predict relation classes, and are led astray when such entity surface forms are absent.

It seems that models learn that they can rely on the surface forms instead of paying attention to how the relation is expressed in the context of the surface forms with which the entities are mentioned. Thus, the models perform weakly in case of unexpected pairs of entities (see Section 5.2).

We also observe that, on average, model performance follows a similar pattern with regards to the adversarial strategy applied. In particular, the models generally fare better when the subject entity, rather than the object entity, is substituted. The only exception is SpanBERT, where replacing either the subject entity or the object entity results in a comparable decrease in performance.

We also observe that substituting both entities has the strongest negative impact for the first three types of substitutions. Among the models tested, only the UniST model exhibits a marginal performance improvement when both entities are replaced in a *same-role* substitution compared to when only object is replaced. On average, the *F1* score is 9.67 points lower when objects are substituted, than when subjects are substituted, with difference as high as 17.2 *F1* score in the case of *diff.-type* substitution.

The most robust model, i. e., LUKE, which reaches 72.0 *F1* in the standard evaluation, loses an average of 24.7% of *F1* in the adversarial evaluation. Even in cases of adversarials that are semantically closer to the original sentence, i. e., *same-role* and *same-type*, this model reaches as low as 64.9 and 57.3 of *F1*, respectively, and in case of *masking* entities, it reaches as low as 27.7 *F1* score.

Some models fare particularly bad in the *masking* strategy. Here, NLI (w/o) reaches an *F1* score between 0.0 and 0.27 when entities are masked. NLI (w/) always reaches 0.0 *F1* score in the *masking* substitutions.

5.2. Relations predicted

The models, when put under pressure, tend to default to the no_relation label, which turns out to be the most frequently predicted relation. While this is to be expected given the unbalanced nature of TACRED, the models still predict no_relation more frequently in the adversarial setting, compared to those actually present in the data, as shown in Table 3. In this case, as well, object substitutions have a stronger impact than subject substitutions. In fact, across all cases except for the *same-type* substitution (SpanBERT, NLI (w/)) and the *masking* substitution (SpanBERT, NLI (w/o)),

---

[4]See the official dataset statistics at https://nlp.stanford.edu/projects/tacred/#stats

[5]https://huggingface.co/studio-ousia/luke-large

[6]https://huggingface.co/mrm8488/spanbert-base-finetuned-tacred

[7]http://github.com/osainz59/Ask2Transformers

[8]https://github.com/luka-group/unist

[9]https://github.com/luka-group/sure

[10]https://github.com/yuhaozhang/tacred-relation/tree/master

Model	std.	adv.	diff.	same-role sub.			same-type sub.			diff.-type sub.			masking sub.		
				subj	obj	subj + obj	subj	obj	subj + obj	subj	obj	subj + obj	subj	obj	subj + obj
LUKE	72.0	54.2	-24.7%	69.2	65.5	64.9	67.8	60.7	57.3	60.9	35.0	31.7	66.7	43.1	27.7
SpanBERT	70.8	26.1	-63.0%	42.4	41.7	39.8	39.4	40.3	35.1	35.0	32.3	22.9	30.6	37.4	23.5
UniST	75.5	33.4	-39.4%	53.2	47.4	47.5	49.9	40.3	35.7	40.0	15.3	8.3	46.1	13.6	3.6
SuRE	74.8	22.4	-59.9%	37.7	32.5	28.9	36.5	29.2	23.5	29.8	12.5	8.1	24.8	5.7	0.1
TYP-Marker	72.0	50.6	-29.7%	68.7	64.5	63.3	64.0	57.1	48.4	52.4	26.4	15.5	66.0	44.6	36.0
NLI (w/)	68.6	30.4	-55.6%	63.4	58.4	54.8	53.6	52.1	38.2	24.4	14.9	4.8	0.0	0.0	0.0
NLI (w/o)	42.7	27.9	-34.4%	42.6	40.8	38.5	42.6	39.5	36.5	37.4	24.3	21.0	0.3	0.1	0.1
avg	68.0	35.0	-48.5%	53.8	50.1	48.1	50.5	45.6	39.2	40.1	22.9	16.0	33.5	20.6	14.5

Table 2: Fine-grained *F1* scores on *Tacred*. **std.** the evaluation on the standard test set, **adv.** the average evaluation over all the adversarial strategies, **diff.** the percentage of loss from standard to adversarial evaluation. The next columns contain *F1* score for each strategy.

object substitutions lead the models to predict the `no_relation` label with equal or greater frequency compared to subject substitutions.

In view of the above, we can assume that one of the main sources of inaccuracy is that models default to the most frequent label in the training set (in this case to `no_relation` by a large margin) when put under unexpected situations.

Furthermore, as depicted in Figure 3, a comparison is made between the predictions on the test set and those on the test set following *diff.-type* object substitutions, with LUKE serving as an exemplary model. The figure shows that the model, beyond predicting the `no_relation` label very frequently, also predicts a significant number of random relations.

This behavior was consistently observed across all models examined. Additionally, despite the high similarity of the relations (with distinctions primarily in the relation scope), the model's performance is notably superior for `org:city_of_headquarters` relation compared to `org:country_of_headquarters` and `org:state-orprovince_of_headquarters`. The cause of these phenomena remains uncertain.

Nevertheless, it is important to note that they are not attributable to a bias in the number of training samples.[11]

It is possible to imagine the situation that a model correctly identifies that r is the only relation that is expressed in a sentence d and then, after an entity in d was replaced with another entity, resulting in the sentence d', it is no mistake if the model identifies that the only relation that is expressed in d' is a relation r' which is different from r. For example, in a sentence like *"He lives in Italy"* that expresses the `per:country_of_residence` relation, it can be appropriate that after a country is replaced by a city, then the relation that is expressed is `per:cities_of_residence`. Here, the context

of the replaced entity remains the same, the same context might be equally indicative for each of the two relations.

These situations need to be taken into account. Otherwise, one would penalize a model for predicting the actual relation being expressed in the context of the sentence. We carry out a case study to investigate whether these effects take place in the context of the TACRED dataset. Note that the only substitution strategies where this phenomenon needs to be investigated is the *diff.-type* substitution.

As it turns out, there are certain relations in which the substition of an entity with an entity of another type might generate a different relation in the adversarial sentence compared to the original sentence. We refer to these sets as sets of **mutually confusable relations**. We list these sets of mutually confusable relations in Fig. 1. For example, if in a sentence that expresses the `per:countries_of_residence` relation we replace a country with a city, then in the resulting sentence it would make sense that the relation `per:cities_of_residence` is identified.

Figure 3 shows the predictions made for 162 sentences for the set of `headquarter` relations for the LUKE model before (on the left) and after the *diff.-type* object entity substitution (on the right). One can observe that the change in prediction to other relations within the same set of mutually confusable relations is not severe. It is rarely the case that a sentence where the `org:country_of_headquaters` relation is identified is transformed into a sentence where the `org:city_of_headquaters` relation is identified (2 times), it is rarely the case that a sentence where the `org:state-orprovince_of_headquaters` relation is identified is transformed into a sentence where the `org:country_of_headquaters` relation is identified (2 times), and so on.

We studied this phenomenon for all models and all sets of relations. We computed the percentage of relations that are changed into other relations within the same set. An example for LUKE

---

[11]We generated plots for all sets of mutually confusable relations for all substitution strategies and models. The scripts to generate these plots along with the generated plots can be found in our GitHub repository. See `https://github.com/Nolanogenn/adv_re`.

Model	std.	same-role sub.			same-type sub.			diff.-type sub.			masking sub.		
		subj	obj	subj + obj	subj	obj	subj + obj	subj	obj	subj + obj	subj	obj	subj + obj
LUKE	-1.7%	+4.0%	+6.4%	+8.0%	+2.0%	+6.5%	+8.4%	+6.4%	+17.8%	+18.7%	+2.8%	+10.9%	+24.7%
SpanBERT	+7.0%	+12.4%	+14.2%	+17.7%	+13.3%	+12.3%	+17.3%	+15.7%	+18.7%	+25.5%	+19.2%	+8.7%	+24.1%
UniST	+1.0%	+5.7%	+8.9%	+11.2%	+6.3%	+9.7%	+23.2%	+28.4%	+28.5%	+35.3%	+39.2%	+39.2%	+39.2%
SuRE	-1.1%	+26.2%	+28.4%	+30.3%	+27.6%	+28.9%	+31.3%	+29.0%	+34.8%	+36.0%	+31.6%	+37.8%	+38.8%
TYP-Marker	+1.3%	+6.9%	+9.4%	+11.6%	+6.8%	+10.4%	+13.3%	+13.0%	+21.8%	+28.1%	+6.6%	+20.6%	+23.8%
NLI (w/)	+3.3%	+13.5%	+16.0%	+19.6%	+16.6%	+15.7%	+23.2%	+28.4%	+28.5%	+35.3%	+39.2%	+39.2%	+39.2%
NLI (w/o)	+3.2%	+11.8%	+14.3%	+18.4%	+10.6%	+12.9%	+16.8%	+10.6%	+22.6%	+27.2%	+39.1%	+38.7%	+8.9%

Table 3: Changes in predicted *no_relation* labels compared to actual ones, for every model across all adversarial datasets and testing set. The **std.** column refers to the standard test set.

Figure 1: Sets of mutually confusable relations.

Subject sensitive
residence: *per:countries_of_residence, per:cities_of_residence, per:stateorprovinces_of_residence*
headquarter: *org:country_of_headquarters, org:city_of_headquarters, org:stateorprovince_of_headquarters*
death: *per:city_of_death, per:stateorprovince_of_death, per:country_of_death*
birth: *per:city_of_birth, per:stateorprovince_of_birth, per:country_of_birth, per:origin*
Object sensitive
name: *org:alternate_names, per:alternate_names*
religion: *per:religion, org:political/religious_affiliation*
member: *org:member_of, org:top_members/employees, per:employee_of*

Figure 2: Percentage of predictions within the same set of relations for LUKE.

is shown in Figure 2. The only critical values are for the *diff.-type* subject substitution on the set of *religion* relations and for the *diff.-type* object substitution on the *headquarter* set. However, these values are not representative with 13 vs. 35 samples with an average of 98 samples per cell across the plot.

Thus, although in principle by following the way we evaluate the models one could happen to underestimate the *F1* score of a model due to this phenomenon, we found that this phenomenon rarely occurs for the TACRED datasets, our substitution strategies, and the models we analyzed.

5.3. Type-aware analysis

Since the work by Petroni et al. (2019), it is a common assumption that large language models show

a strong ability to recall factual knowledge without any explicit access to a knowledge base (Jiang et al., 2020; Liu et al., 2021), e. g., by retaining information about entity types and domains/ranges of relations.

We hypothesize that such an ability can have an impact on the results in adversarial settings. Thus, we investigate whether this sort of information implicitly learned by the models at training time has any impact in the adversarial setting.

While learning this information can be useful for specific tasks (e. g., ontology induction), models could find the shortcut to reply to the types of entities and pay less or no attention to the context of the entities in the text.

For instance, consider the following adversarial sample, generated by the *same-role* strategy on subject and object entities:

The first $ 9 billion in proceeds from the sale will go toward redeeming preferred shares in Fidelity Investments$_{SUBJ:ORGANIZATION}$, *held by the NASA*$_{OBJ:ORGANIZATION}$ *[…]*

The correct relation is `org:parents`, but the combination of entity types *organization* and *organization* also adheres to the domain/range constraints of the following relations: `org:shareholders`, `org:alternate_names`, `org:member_of`, `org:members`, and `org:subsidiaries`.

If the hypothesis that models inherently learn information about domain and range is accepted, it would follow that, by leveraging such informtion, the models would tend to predict a relation label from this list.

In this case, even if the relation label is incorrect, the models would still show some sort of semantic
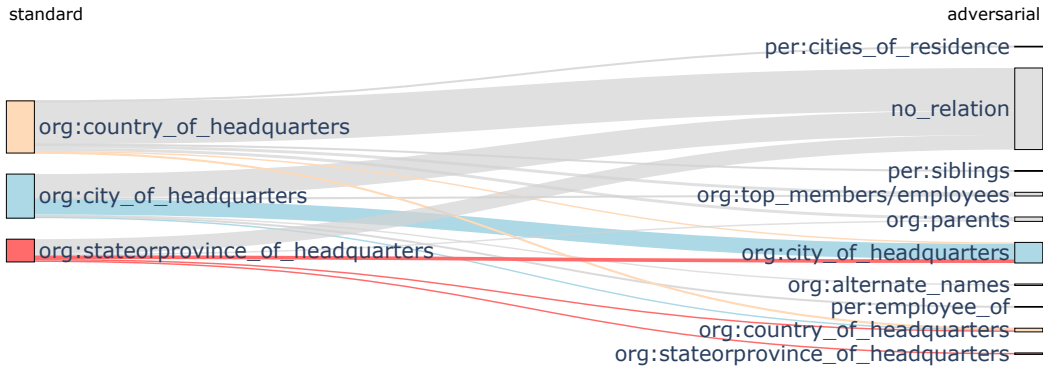
Figure 3: Comparison: LUKE's predictions on the standard test set vs. the predictions on the test set following the *diff.-type* object substitution strategy for selected relations.

knowledge on entity tpes.

To test this hypothesis, we first create a list of all the possible subject type and object type combinations that occur in the training split, and the relation they are involved in.

Then, for every example in the adversarial datasets, we check whether the predicted relation falls into one of the possible labels according to the domain and range of the generated sentence. If a combination of entity types is never found in the training set (for instance when at least one entity is masked), the only available relation is `no_relation`.

The percentages of predicted relations that adhere to the constraints posed by the newly generated sentences are shown in Figure 4. According to the results, the initial hypothesis that
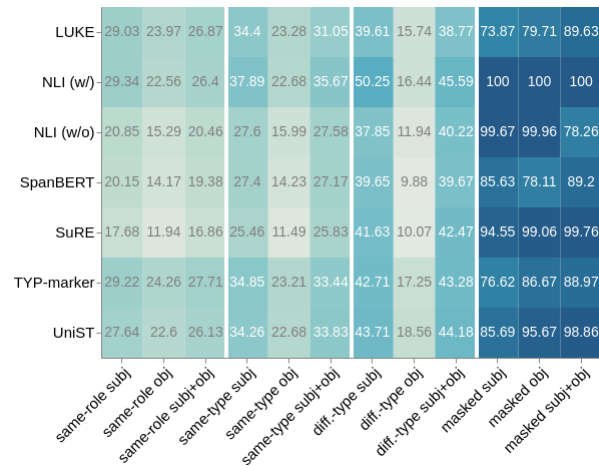

Figure 4: Percentages of predicted relations adhering to the type constraints posed by the adversarial examples.

the investigated models use information about entity types is proved wrong. For instance, for the example shown above, all the models, except for `TYP_marker`, predicted the relation shown in the sentence as being `no_relation`.

As shown in Figure 4, this is not an anomaly. In fact, models only adhere to constraints posed by domain/range by 50% at most, but in most cases much lower, with the only exception being in the case of *masking* adversarials. However, this is expected since in this case the only available relation is *no_relation*.

For the other strategies of substitution, the models weakly adhere to entity types combination, even when entity types are not changed by the adversarial strategy (i. e., for *same-role* and *same-type* adversarials).

Similarly, since the substitutions are randomly generated, it might happen for the *diff.-type* strategy that the generated adversarials adhere to a different relation than the one holding in the original sentence (see Section 7). Even in those cases, the models adhere to the newly posed domain/range constraints with a low frequency.

At best, NLI (w/) adheres to such constraints 50.25% of the time, for *diff.-type* substitution on subject entities, with all the other models faring worse than this.

We can thus assume that models take into account entities' surface forms rather than implicit information about entity types and domain/range.

As mentioned already in Section 5.1, in this case as well, the models are less likely to adhere to known entity types constraints when the object entity is the one being substituted. This is clearly seen in the *diff.-type* substitution strategy, where the examples in which the objects are substituted are, on average, 66.16% less likely to adhere to constraints posed by entity types.

6. Conclusion and Future Work

In this work, we test a set of SOTA Relation Extraction models on multiple adversarial datasets created through semantically motivated substitution strategies. As shown by the results, the models are heavily affected by this kind of adversarials, and in particular, they show an overreliance

on the surface form of entities to predict the correct relation label, rather than relying on the actual linguistic structure of the relation expressed in a sentence or on the entities' types.

The models lose an average of 48.5% of *F1* score when facing adversarials, with certain adversarial strategies being more impactful than others.

The possibility that models are led astray by entity types is proven false by a deeper analysis of the results. In fact, models generally default to the `no_relation` class even in the face of adversarials where a relation between the two entities might exist, according to their types. While, given the unbalanced nature of the TACRED dataset, this is to be expected, it also means that models are easily led astray by changes in the input data, even in cases that leave the relation unchanged.

In future work, applying the adversarial methodology to another dataset might prove useful in testing whether models trained on another set of data might be more resilient to adversarial attacks.

Furthermore, while semantically motivated adversarials have proven to be effective when evaluating RE models, it is possible to build *syntactically* motivated adversarials that substitute or mask specific parts-of-speech in sentences, e. g., verbs, to test models' reliance on such features.

Such adversarials can be included in future work, to prove whether models actually tend to rely on entity's surface forms, or are just led astray by any kind of unexpected change in the testing set.

While this work focuses on the use of adversarials as a means to explore the workings of language models, adversarials have also been proposed as a way to train more robust models (Szegedy et al., 2014). Despite being out of scope for the present paper, an investigation of how model performance improves / robustness increases when trained on the generated adversarials can help to gain a deeper understanding of the models.

7. Limitations

We outlined that the adversarial strategy *diff.-type* might generate samples expressing another relation where two relations are expressed by similar structures. An example and a case study analyzing the models' behavior is shown in Section 5.2.

To investigate the amount of such possible false negatives independently of the substitution strategy for TACRED, we compute the Jaccard similarity on the token basis for sentences within groups of mutually confusable relations. To compute this similarity, entity mentions were removed from every sentence. The results are shown in Figure 5.

Since sentences referring to different relations have, in general, an average of one third of their tokens in common (this also takes into account function words), it can be assumed that the actual over-
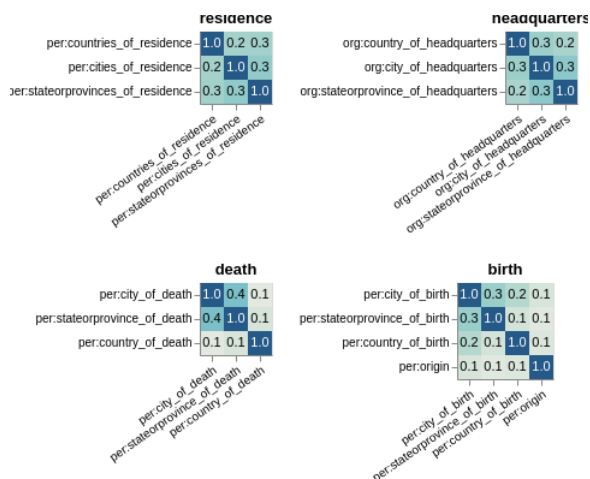

Figure 5: Jaccard similarity between groups of similar relations

lap between sentences expressing different relations is low enough for false negatives not to be considered an issue.

We further checked how many times the *diff.-type* strategy generates sentences representing similar relations. To do so, we compute how many sentences generated from one relation generate sentences representing another, similar one. These percentages are shown in Table 4.

Relation	%
`per:cities_of_residence`	4.0
`per:countries_of_residence`	4.51
`per:stateorprovinces_of_residence`	2.38
`per:origin`	3.47
`per:cities_of_death`	4.17
`per:country_of_death`	4.17
`per:stateorprovince_of_death`	1.19
`per:stateorprovince_of_birth`	2.78
`per:country_of_birth`	6.67
`per:city_of_birth`	1.67
`org:stateorprovince_of_headquarters`	5.36
`org:country_of_headquarters`	5.66
`org:city_of_headquarters`	3.65

Table 4: Percentages of generated sentences that represent similar relations

Even though the substitutions might generate a correct relation in some cases, we have already shown that the models tend not to agree to domain/range information in the newly generated sentences, and this is the case for all models across all types of generated adversarials.

8. Ethics

Our work primarily focuses on the investigation of existing language models, which in itself does not introduce new ethical biases. Instead, our work contributes to the broader conversation about the

ethical development of language models, emphasizing the need for transparency and fairness. Nevertheless, when considering the use of adversarials for training, it is crucial to ensure that such manipulations do not introduce new harmful biases. By default, our methodology might generate false facts, which, if used to train a model, would then be learned by the model.

Finally, since the TACRED dataset is based on the English language, this investigation might show different results with different languages, and as such should not be taken for granted in a multilingual setting.

Acknowledgements

9. 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*.

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

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

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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

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)*.

Brian Formento, Chuan Sheng Foo, Luu Anh Tuan, and See Kiong Ng. 2023. Using punctuation as an adversarial attack on deep learning-based NLP systems: An empirical study. In *Findings of the Association for Computational Linguistics: EACL 2023*.

Piotr Gaiński and Klaudia Bałazy. 2023. Step by step loss goes very far: Multi-step quantization for adversarial text attacks. In *Proceedings of the 17th Conference of the European Chapter of the 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*.

Shreyansh Goyal, Sumanth Doddapaneni, Mitesh M.Khapra, and Balaraman Ravindran. 2022. A survey of adversarial defences and robustness in nlp.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on 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*.

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

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2).

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*.

Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. 2020. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, Muhao Chen, et al. 2022. Summarization as indirect supervision for relation extraction. In *EMNLP - Findings*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

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*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *ArXiv*, abs/1909.01066.

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)*.

Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. 2020. Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Oscar Sainz and German Rigau. 2021. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In *Proceedings of the 11th Global Wordnet Conference*.

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

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

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*.

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)*.

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*.

Jincheng Xu and Qingfeng Du. 2020. Texttricker: Loss-based and gradient-based adversarial attacks on text classification models. *Eng. Appl. Artif. Intell.*, 92:103641.

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)*.

Jin Yong Yoo, John Morris, Eli Lifland, and Yanjun Qi. 2020. Searching for a search method: Benchmarking search algorithms for generating NLP adversarial examples. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Lifan Yuan, Yichi Zhang, Yangyi Chen, and Wei Wei. 2023. Bridge the gap between cv and nlp! a gradient-based textual adversarial attack framework.

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*.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

A. Entity Masking

Certain RE models replace the entity strings with masks during training, while others do so during training and testing. The implemetation of entity masking has been proven to counteract overfitting that might be caused by label noise, while also avoiding the over-reliance on entity-level heuristics during fine-tuning.

Often, the masking at training time is limited to a certain percentage of randomly chosen tokens. Both LUKE and SpanBERT follow the BERT methodology for masking, which applies masks to 15% of the tokens. As such, most of the entities will not be replaced, and this might lead the models to learn undesired shortcuts despite the efforts of implementing masking. An additional investigation might be needed to better quantify the impact of such masking strategies, but it is beyond the scope of the present work.

We used the pre-trained models, as well as the fine-tuned RE models (when available, otherwise trained with the original implementation and hyperparameters), as described in the respective papers. As such, in our experiments, masking is implemented as per the original papers. For our evaluation, we used the official codes with their default configurations.

In the following, we discuss the process of entity masking for the specific processes implemented in the paper.

LUKE In the original publication of LUKE it is stated that two special entities, *[HEAD]* and *[TAIL]*, are representing the head and tail entities. These entities, along with the input words, are fed into the model for each sentence. Subsequently, the task is addressed using a linear classifier that relies on a concatenated representation of the head and tail entities. We evaluate LUKE using the implementation from Huggingface.[12] As per the documentation, the use of *[MASK]* tokens at testing time is completely optional for LUKE, and its implementation can be used to make the model focus on the context more. The use of these tokens would make the main premise of our work meaningless, as entity replacement would have just resulted in masked tokens, without any impact on the results.

SpanBERT The default test configuration provided in the official repository of SpanBERT[13] does not mask any of the tokens in the input sentence during test time. This default configuration is also implemented in this work, and as such no masking is applied.

UniST This model does not apply any entity masking. Since entities in TACRED are also annotated with entity types, the entity type labels are placed in front of their corresponding entity mentions in the task description to provide additional information for relation classification.

SuRE Does not apply any entity masking.

TYP-Marker Does not apply any entity masking. Instead, entity markers are applied that enclose the subject and object entities.

NLI (RoBERTa) Does not apply any entity masking.

B. Linguistic Patterns vs. Entity Mentions in RE

It is well-established in linguistics and cognitive science that humans rely on linguistic patterns for understanding relations in text. However, we are aware of the crucial role entities play for RE. Our claim is not to entirely disregard entities, but rather to highlight the importance of the sentence context in the RE process.

In certain tasks, looking at entities is indeed crucial. However, the sentence context is always important. Our experiments with various substitution operations have shown that models can be significantly misled. The experiments aim not to develop an approach disregarding entities but rather to empirically demonstrate the extent to which current models depend on entities. This dependence can be viewed as a limitation to their generalization capabilities.

Furthermore, we argue that a model's ability to generalize and adapt to unforeseen circumstances is important. LMs are trained on data that may become outdated, and they should exhibit robustness to unexpected situations. By encouraging models to pay more attention to the sentence context, they would in turn show a better robustness to new entities.

In summary, we suggest a balanced consideration of linguistic patterns and entity mentions in RE models. The sentence context enhances the model's ability to adapt to novel situations and align more closely with human-like language understanding.

---

[12]https://huggingface.co/docs/transformers/model_doc/luke

[13]https://github.com/facebookresearch/SpanBERT/