UNIVERSITÄT BIELEFELD

Scan to view on GitHub

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
nolanogenn@gmail.com {mblum, bell, cimiano}@techfak.uni-bielefeld.de

Introduction

In relation extraction (RE), we would expect a model to identify the same relation independently of the entities involved in it.

Our analyses show that state-of-the-art RE 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.

Methodology

We developed 12 semantically-motivated strategies to generate adversarials by replacing entity mentions. In doing so, the contextual information between and around the entities, and the relation expressed by the sentence, is left untouched.

	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

Experiments

Data The starting point is the human-labeled TACRED dataset. After some filtering, this left us with 6,277 sentences in the test set for generating adversarial examples.

Models We test different state-of-the-art models in RE: **LUKE**, **SpanBERT**, **UniST**, **SuRE**, **TYP-Marker**, and **NLI**.

Results

- The models are affected by the adversarials, with an average loss of 48.5% in F_1 score. The most robust model is LUKE, which loses on average 24.7% of F_1.
- The *same-role* & the *same-type* sub. are the least impactful ones, while the *masking* sub. has the strongest impact on the results.
- The models generally fare better when the subject entity, rather than the object entity, is substituted. Substituting both entities has the strongest negative impact.
- The models, when put under pressure, tend to default to the `no_relation` label, which turns out to be the most frequently predicted relation.

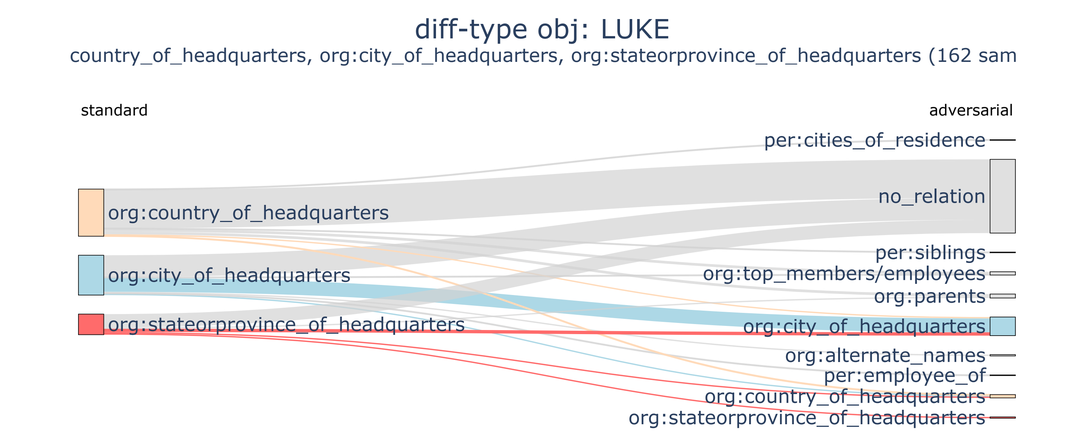

Figure 1: Comparison: Model predictions on the standard test set vs. the predictions on the test set following the *diff.-type* object substitution strategy for selected relations.

- The hypothesis that the investigated models use information about entity types is proved wrong.

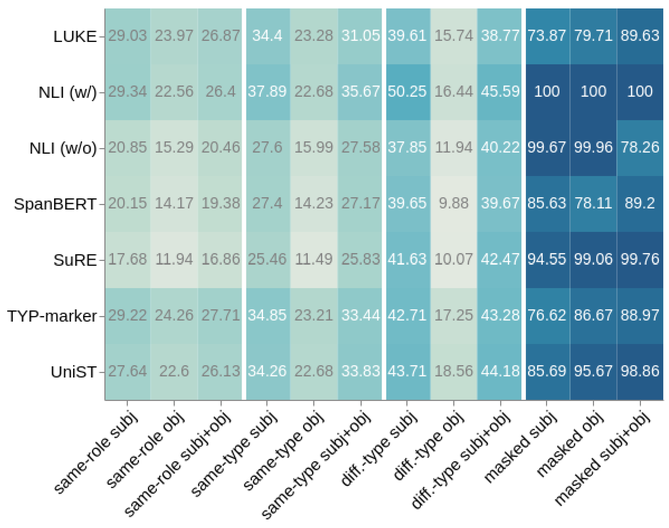

Figure 2: Percentages of predicted relations adhering to the type constraints posed by the adversarial examples.

Conclusions

Our analyses show that the performance of state-of-the-art RE models significantly deteriorates on the adversarials (avg. of -48.5% in F_1), indicating an overreliance on the surface form of entities to predict the correct relation label. Instead, models should rely on the actual linguistic structure of the relation expressed in a sentence or on the entities' types.

Table 2: Fine-grained F_1 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 F_1 score for each strategy.

Model	**std.**	**adv.**	**diff.**	**Same-role**			**Same-type**			**Diff.-type**			**Masking**		
				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