Re-TACRED: A New Relation Extraction Dataset

Carnegie Mellon University

George Stoica, Emmanouil Antonios Platanios, Barnabás Póczos

gis@cs.cmu.edu, emplata@microsoft.com, bapoczos@cs.cmu.edu

MACHINE LEARNING DEPARTMENT

Relation Extraction

Relation Extraction (RE): Determine Relationships between subjects and objects in text

"Farland joined the FBI in 1942."

per:employee_of

TACRED [1]

- One of the *largest* and *most-popular* crowd-sourced RE datasets
- Collected from 2009-2014 TAC KBP [2] evaluations
- *Quality*: .54 Fleiss' Kappa from 761 randomly sampled annotations
- 106,264 instances spread among 42 relations

TACREV [3]

- Manually verified 5K most misclassified TACRED evaluation instances from 49 RE models
- Observed *>50% label error* amongst these examples
- Achieved Fleiss' Kappas of .80 on dev & .87 on test
- Revised dataset improved average F1-score by 8.1%
- However, generalization limited by bias

Re-TACRED

- **Comprehensive**: We verify *entire* TACRED dataset
- **Improved Annotation:** We deploy an improved crowd-sourced annotation strategy
- **High Quality**: We achieve .77 Fleiss' Kappa over *entire* dataset
- **Performance:** We improve model F1-score performance on average by *13%*

Improved Annotation Strategy

Quality Assurance

All workers must be:

- **Experienced**: Have previously completed at least 500 tasks
- **Reliable**: Have at least 95% task approval rate
- **Specialized**: Have passed custom qualification exams for each super-cluster
- **Careful**: Have correctly answered at least 80% of observed gold-sentences

Wrong Type Handling

- [1] and [3] label sentences by providing workers *all type-compatible* relations with sentence entities
- However, all candidate relations are incorrect when assigned entity type is wrong (occurred in 5% of a random sample of size 1K)

"*Thomas More Law Center*" ⟶ *PERSON* or *ORGANIZATION*?

We address this issue in two ways:

1. We extend each label set to include a "wrong_type" relation
2. We preempt "wrong_type" assignments by merging frequently miss-typed entity pairs into "super-clusters"

(*PERSON* | *CITY*, *STATE/PROVINCE*, *COUNTRY*, *LOCATION*) ⟶ *per:locmulti*

Relation Refinements

(i) Explicitly Enabling Identity Relations

"Holly showed off her jewelry."

per:identity

(ii) Merging Very Similar Relations

"... Badr is the armed wing of the ISCI."

org:member_of *or* org:parents?

(iii) Relaxing Challenging Criteria

"Facebook in NYC."

org:city_of_headquarters?

(iv) Enforcing Mutual-Exclusivity

"He is a native of Pittsburgh, PA."

per:city_of_birth or per:city_of_residence?

Results

Dataset	Metric	PALSTM*	C-GCN*	SpanBERT*
TACRED	Precision	68.1	68.5	70.1
	Recall	64.5	64.4	69.2
	F1	66.2	66.3	69.7
Re-TACRED	Precision	78.3	79.7	84.6
	Recall	77.6	78.5	83.9
	F1	77.9	79.1	84.2
Change %	Precision	+12.2	+11.2	+14.5
	Recall	+13.1	+14.1	+14.7
	F1	+11.7	+12.8	+14.5

Model	Dataset	Refined Labels (i)	(ii)	(iii)	(iv)
PALSTM*	TACRED	46.7	21.2	55.9	51.9
	Re-TACRED	87.6	48.8	68.8	53.4
	Change %	+30.9	+27.6	+12.9	+1.5
C-GCN*	TACRED	14.6	22.7	56.7	51.5
	Re-TACRED	88.1	51.9	73.7	54.2
	Change %	+73.5	+29.2	+17.0	+2.7
SpanBERT*	TACRED	44.1	51.9	66.8	55.9
	Re-TACRED	91.7	65.1	74.0	69.8
	Change %	+56.6	+13.2	+7.2	+13.9

Model	Train Split	Test Split	Metrics F1	Precision	Recall
PALSTM*	$\text{TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	72.3	71.3	73.3
	$\text{TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	73.3	76.7	70.2
	$\text{Re-TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	68.3	65.9	70.9
	$\text{Re-TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	75.9	75.8	76.1
C-GCN*	$\text{TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	72.6	71.1	74.3
	$\text{TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	73.2	76.0	70.6
	$\text{Re-TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	69.2	68.5	69.8
	$\text{Re-TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	77.3	78.2	76.5
SpanBERT*	$\text{TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	75.0	74.7	75.3
	$\text{TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	76.8	81.2	72.8
	$\text{Re-TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	74.1	70.9	77.7
	$\text{Re-TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	84.1	85.0	83.1

Dataset Link: *https://github.com/gstoica27/Re-TACRED.git*

Citations: [1] Y. Zhang et. al., ACL 2017; [2] C. Zhang et. al., ACL 2012; [3] C. Alt et. al., ACL 2020