Carnegie Mellon University

Re-TACRED: Addressing Shortcomings of the TACRED Dataset

George Stoica, Emmanouil Antonios Platanios, Barnabás Póczos

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

ML 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 evaluations
- .54 Fleiss' Kappa from 761 randomly sampled annotations
- 106,264 instances spread among 42 relations

TACREV [2]

- 5K most challenging examples yield *>50% error* and reduce performance by *8.1% average F1-score*
- However,
  1. Selection bias constrained error estimation
  2. Analysis limited by largely uncorrected dataset

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

Relation Refinements

- **Addressing Ambiguous Relations**

  "*Holly* showed off *her* jewelry."

  per:identity

- **Merging Very Similar Relations**

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

  org:member_of *or* org:parents?

- **Relaxing Challenging Criteria**

  "*ORGANIZATION* from *CITY*."

  org:city_of_headquarters?

- **Enforcing Mutual-Exclusivity**

  "*He* is a native of *Pittsburgh*, PA."

  per:city_of_birth or per:city_of_residence?

Edge-Case Handling

Previous work provides workers with *all type-compatible* relations between sentence entities

- However, what if assigned entity types are incorrect?

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

We address this issue in two ways:

1. Add a "wrong_type" relation to each relation set
2. Extend label sets to include relations from frequently confused types

(*PERSON*:*CITY*, *STATE/PROVINCE*, *COUNTRY*, *LOCATION*) ⟶ (*PERSON*:*LOCMULTI*)

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 labeling task
- **Careful**: Have correctly answered at least 80% of observed gold-sentences

Results

Model	Dataset	All Labels Metric F1	Precision	Recall	Train Split	Test Split	Non-Refined Labels Metrics F1	Precision	Recall	Dataset	Refined Labels Category F1 Ambiguous	Similar	Challenging	Exclusive
PA-LSTM*	TACRED	66.2	68.1	64.5	$\text{TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	72.3	71.3	73.3	TACRED	46.7	21.2	55.9	51.9
	Re-TACRED	77.9	78.3	77.6	$\text{TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	73.3	**76.7**	70.2	Re-TACRED	87.6	48.8	68.8	53.4
	Change %	**+11.7**	**+10.2**	**+13.1**	$\text{Re-TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	**75.9**	75.8	**76.1**	Change %	**+30.9**	**+27.6**	**+12.9**	**+1.5**
C-GCN*	TACRED	66.3	68.5	64.4	$\text{TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	72.6	71.1	74.3	TACRED	14.6	22.7	56.7	51.5
	Re-TACRED	79.1	79.7	78.5	$\text{TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	73.2	76.0	70.6	Re-TACRED	88.1	51.9	73.7	54.2
	Change %	**+12.8**	**+11.2**	**+14.1**	$\text{Re-TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	**77.3**	**78.2**	**76.5**	Change %	**+73.5**	**+29.2**	**+17.0**	**+2.7**
SpanBERT*	TACRED	69.7	70.1	69.2	$\text{TACRED}_{\text{train}}$	$\text{TACRED}_{\text{test}}$	75.0	74.7	75.3	TACRED	44.1	51.9	66.8	55.9
	Re-TACRED	84.2	84.6	83.9	$\text{TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	76.8	81.2	72.8	Re-TACRED	91.7	65.1	74.0	69.8
	Change %	**+14.5**	**+14.5**	**+14.7**	$\text{Re-TACRED}_{\text{train}}$	$\text{Re-TACRED}_{\text{test}}$	**84.1**	**85.0**	**83.1**	Change %	**+47.6**	**+13.2**	**+7.2**	**+13.9**

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

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