TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task

ACL 2020

Christoph Alt, Aleksandra Gabryszak, Leonhard Hennig

German Research Center for AI (DFKI)
Speech and Language Technology Lab

Relation Extraction

Relation extraction (RE) is concerned with extracting semantic relations from text

?

The measures included Aerolineas's domestic subsidiary, Austral.

head

tail

org:subsidiaries ✓

org:parents ✗

org:members ✗

...

But, what do we do when models fail?

Motivation

- **Goal:** Understand *where* and *why* relation extraction (RE) models fail
- Current state of the art in RE:
  - TACRED [Zhang et al., 2017]: 71.5 F1 [Baldini Soares et al., 2019; Peters et al., 2019]
  - SemEval 2010 Task 8 [Hendrickx et al., 2010]: 91.0 F1 [Li et al., 2020]
  - ACE 2005 [Walker et al., 2006]: 63.2 F1 [Luan et al., 2019]
- **Problem:**
  - A single metric, e.g., F1, precision, recall, is insufficient to understand model capabilities
    - → Instead, we should focus on errors
  - However, difficult to determine the cause
    - → model errors
    - → dataset bias
    - → annotation errors

Research Questions

- TACRED is one of the largest, most widely used RE benchmarks
- **Observation**: Error rate of almost 30% is still high

Questions

- Is there still room for improvement, and can we identify the underlying factors that contribute to this error rate?
  - To what extent does the quality of crowd based annotations contribute to the error rate?
  - What can be attributed to dataset and models?

Re-Annotation of Challenging Examples

- **Goal:**
  - Identify most challenging examples (development and test split)
- **Approach:**
  - Rank each example according to evidence from 49 different RE model predictions
  - Select examples for manual evaluation
    - → *Challenging* → misclassified by at least half of the models
    - → *Control* → Correctly classified by at least 39 models
  - Manual re-annotation of selected examples
    - → According to TAC KBP guidelines
    - → With label suggestions, similar to original crowdsourcing

Results (1): Label Error Analysis

	Dev		Test	
	Challenging	*Control*	*Challenging*	*Control*
# Examples (# positive)	3,088 (1,987)	567 (547)	1,923 (1,333)	427 (407)
# Revised (# positive)	1,610 (976)	46 (46)	960 (630)	38 (38)
# Revised (% positive)	**52.1 (49.1)**	**8.1 (8.4)**	**49.9 (47.3)**	**8.9 (9.3)**

- Overall approx. 5k challenging examples re-annotated
- Approx. 50% of challenging examples were revised (relabeled)
- Only 8% of examples in control were revised

Results (1): Label Error Analysis

Model	Original	Revised
CNN, masked	59.5	66.5
TRE	67.4	75.3
SpanBERT	70.8	78.0
KnowBERT	71.5	79.3

- Approx. 8% absolute improvement in F1 score across all models
- Average score across 49 models from 62.1 to 70.1 F1
- State-of-the-art improved to 79.3 F1

Error Categories for Model Misclassifications

- **Goal:**
  - Manual explorative analysis of model misclassifications
  - Categorization into linguistically motivated error categories
- **Approach:**
  - Explore possible linguistic aspects causing incorrect predictions
    - → e.g., entity type errors or distracting phrases
  - Iteratively develop error categories
  - Annotate each misclassification with category

Results(2): Model Error Categories

- 1017 examples, categorized into 9 error categories
- 7 categories related to context, 2 categories related to arguments

Wrong Args	Authorities said they ordered the detention of Bruno 's wife , [Dayana Rodrigues]$_{tail:per}$, who was found with [Samudio]$_{head:per}$'s baby .	*per:spouse*	109
Relation Def.	[Zhang Yinjun]$_{tail:per}$, spokesperson with one of China 's largest charity organization , the [China Charity Federation]$_{head:org}$	*org:top_mem.*	96
Entity Type	[Christopher Bollyn]$_{head:per}$ is an [independent]$_{tail:religion}$ journalist	*per:religion*	31

- Context misinterpretations account for ~96% of errors
- Argument errors account for ~4% of errors
- Incorrect assignment of "no relation" is the most common error

Automated Analysis

- **Goal:**
  - Attribute errors to dataset or models
  - Prevent focusing on hypotheses well handled on average
- **Approach:**
  - Extend misclassification categories to testable hypotheses (groups)
    - → Group examples according to attribute, e.g., “has distracting entity in context”
    - → Automatically verifiable on whole dataset split
  - Validate whether the hypothesis holds
    - → I.e., group of instances shows an above average error rate
    - → Based on the approach of [Wu et al., 2019]

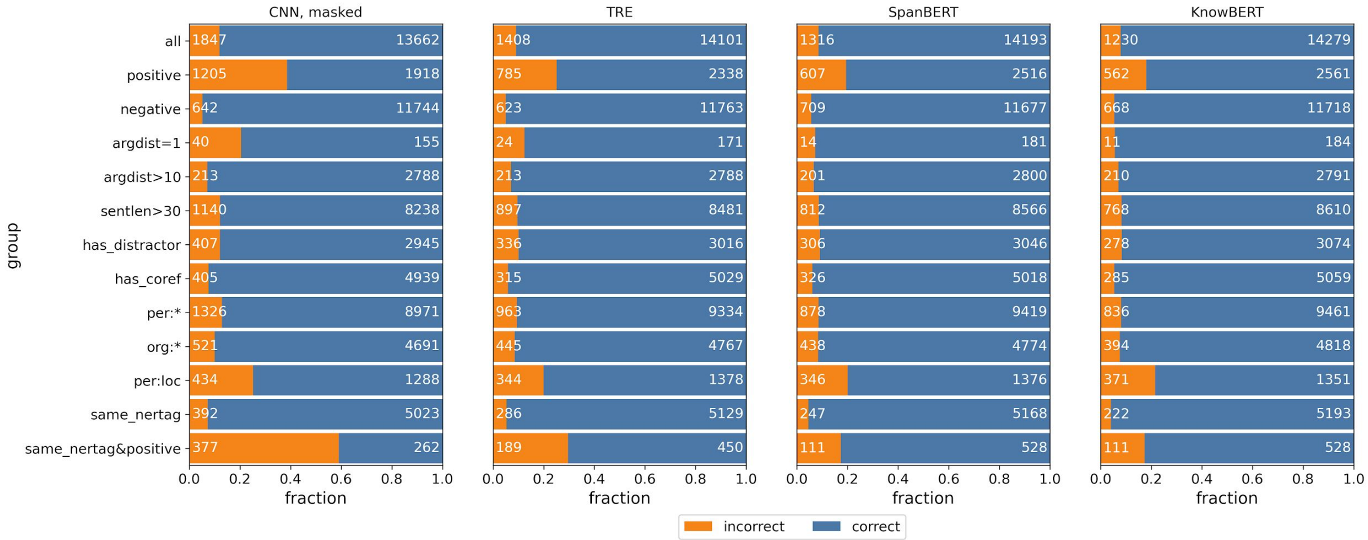
Experimental Setup

- Formulate hypotheses (error groups)

Groups	Attributes
Surface structure	argument distance, sentence length
Arguments	head and tail entity type
Context	distracting entities in context
Ground truth	positive examples, excluding “no relation”

- Compare state-of-the-art model error rates per group
  - TRE [Alt et al., 2019] → OpenAI GPT
  - SpanBERT [Joshi et al., 2019] → BERT, pre-trained on span level
  - KnowBERT [Peters et al., 2019] → BERT, pre-trained jointly with entity linking

Results (3): Per Group Error Rates



- Large fraction of errors caused by two ambiguous groups of relations
  - per:loc
    - → expressed in similar context, e.g., *per:cities_of_resid.* vs. *per:countries_of_resid.*
  - same_nertag&positive
    - → same argument types, e.g., *per:parents*, *per:children* and *per:other_family*

Conclusion

- Manual re-annotation of 5k most challenging TACRED examples (development and test split)
  - → Release of revised dataset, as patch
- Careful evaluation of development and test splits necessary if dataset is crowdsourced
  - → to ensure progress can be measured accurately
- Models often unable to predict a relation even if clearly expressed
- Models frequently ignore argument roles or ignore sentential context
- Two groups of ambiguous relations mainly responsible for remaining errors

Thank you

TACRED patch and code: https://github.com/DFKI-NLP/tacrev

References

- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. EMNLP, 2017.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. SemEval, 2010.
- Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia 57, 2006.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. ACL, 2019.
- Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. EMNLP, 2019.
- Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, Hannaneh Hajishirzi. A General Framework for Information Extraction using Dynamic Span Graphs. NAACL, 2019.
- Cheng Li, Ye Tian. Downstream Model Design of Pre-trained Language Model for Relation Extraction Task. arxiv:2004.03786, 2020.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Errudite: Scalable, reproducible, and testable error analysis. ACL, 2019.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke S. Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. TACL, 2019.
- Christoph Alt, Marc Hübner, and Leonhard Hennig. Improving relation extraction by pre-trained language representations. AKBC, 2019.