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.

But, what do we do when models fail?
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

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Challenging</td>
<td>Control</td>
</tr>
<tr>
<td># Examples (# positive)</td>
<td>3,088 (1,987)</td>
<td>567 (547)</td>
</tr>
<tr>
<td># Revised (# positive)</td>
<td>1,610 (976)</td>
<td>46 (46)</td>
</tr>
<tr>
<td># Revised (% positive)</td>
<td><strong>52.1 (49.1)</strong></td>
<td><strong>8.1 (8.4)</strong></td>
</tr>
</tbody>
</table>

- 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

- 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Original</th>
<th>Revised</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN, masked</td>
<td>59.5</td>
<td>66.5</td>
</tr>
<tr>
<td>TRE</td>
<td>67.4</td>
<td>75.3</td>
</tr>
<tr>
<td>SpanBERT</td>
<td>70.8</td>
<td>78.0</td>
</tr>
<tr>
<td>KnowBERT</td>
<td>71.5</td>
<td>79.3</td>
</tr>
</tbody>
</table>
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.

Relation Def.: [Zhang Yinjun]_{tail:per}, spokesperson with one of China’s largest charity organization, the [China Charity Federation]_{head:org}, 109

Entity Type: [Christopher Bollyn]_{head:per} is an [independent]_{tail:religion} journalist per:religion 96

- 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)

<table>
<thead>
<tr>
<th>Groups</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface structure</td>
<td>argument distance, sentence length</td>
</tr>
<tr>
<td>Arguments</td>
<td>head and tail entity type</td>
</tr>
<tr>
<td>Context</td>
<td>distracting entities in context</td>
</tr>
<tr>
<td>Ground truth</td>
<td>positive examples, excluding “no relation”</td>
</tr>
</tbody>
</table>

- 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
- 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](https://github.com/DFKI-NLP/tacrev)
References


- Christoph Alt, Marc Hübner, and Leonhard Hennig. Improving relation extraction by pre-trained language representations. AKBC, 2019.