Towards a better understanding of neural relation extraction

Christoph Alt

October 2020
Relation extraction

The measures include Aerolineas’s domestic subsidiary, Austral.
The measures include Aerolineas’s domestic subsidiary, Austral.
The measures include Aerolineas’s domestic subsidiary, Austral.
The measures include Aerolineas’s domestic subsidiary, Austral.
The measures include Aerolineas’s domestic subsidiary, Austral.
Neural relation extraction

Example

Dataset
Neural relation extraction

Example

Dataset

The measures include Aerolineas’s domestic subsidiary, Austral.
Neural relation extraction

The measures include Aerolineas’s domestic subsidiary, Austral.
Neural relation extraction

The measures include Aerolineas’s domestic subsidiary, Austral.
Neural relation extraction

The measures include Aerolineas’s domestic subsidiary, Austral.

Example

Dataset

Prediction

org:subsidiaries
org:parents
org:members
...

?
Neural relation extraction

The measures include Aerolineas’s domestic subsidiary, Austral.

Example

Model → Prediction

Dataset

org:subsidiaries
org:parents
org:members
...

head

tail

Example

Dataset

Model → Prediction

org:subsidiaries
org:parents
org:members
...

head

tail

Example

Dataset

Model → Prediction

org:subsidiaries
org:parents
org:members
...

head

tail

Example

Dataset

Model → Prediction

org:subsidiaries
org:parents
org:members
...

head

tail
Neural relation extraction

The measures include Aerolineas’s domestic subsidiary, Austral.

org:subsidiaries
org:parents
org:members
...

head

The measures include Aerolineas’s domestic subsidiary, Austral.

tail
Neural relation extraction

The measures include Aerolineas’s domestic subsidiary, Austral.

org:subsidiaries
org:parents
org:members
...

head
 tête

final representation

Example

Dataset
How do we get a better understanding of neural relation extraction?
Understanding neural relation extraction

Model -> Prediction

Example

Dataset
Understanding neural relation extraction

Model -> Prediction

Example

Dataset

does it properly reflect the task?
Understanding neural relation extraction

- Model
- Example
- Dataset
- Prediction

- internal workings of the model
- does it properly reflect the task?
Understanding neural relation extraction

- Model
- Prediction
  - important features for prediction
  - internal workings of the model
- Example
- Dataset
  - does it properly reflect the task?
Understanding neural relation extraction

- Model
- Dataset
- Example
- Prediction

Does it properly reflect the task?

- Incorrect predictions
- Important features for prediction
- Internal workings of the model
Understanding neural relation extraction

Model → Prediction

Example → Prediction

Dataset → Prediction

Does it properly reflect the task?

Incorrect predictions

Important features for prediction

Internal workings of the model
In this talk

1. What linguistic aspects of the input do neural relation extraction models focus on?
In this talk

1. What linguistic aspects of the input do neural relation extraction models focus on?
2. Where do neural relation extraction models fail, and why?
1. What linguistic aspects of the input do neural relation extraction models focus on?
2. Where do neural relation extraction models fail, and why?

Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction. Christoph Alt, Aleksandra Gabryszak and Leonhard Hennig. ACL 2020
What properties are important to relation extraction?

The measures include Aerolineas’s domestic subsidiary, Austral.
What properties are important to relation extraction?

The measures include Aerolineas’s domestic subsidiary, Austral.
What properties are important to relation extraction?

The measures include Aerolineas's domestic subsidiary, Austral.

- argument order?
- entity type?
What properties are important to relation extraction?

The measures include Aerolineas's domestic subsidiary, Austral.

- Argument order?
- Distance between arguments?
- Entity type?
What properties are important to relation extraction?

The measures include:

- Aerolineas's domestic subsidiary, Austral.
- Sentence structure?
- Entity type?
- Distance between arguments?
- Argument order?
What properties are important to relation extraction?

The measures include:
- Aerolineas’s domestic subsidiary, Austral.
- Sentence structure?
- Entity type?
- Argument order?
- Distance between arguments?
Do representations contain any of these properties?

The measures include Aerolineas’s domestic subsidiary, Austral.

- head
- tail
- sentence structure?
- argument order?
- distance between arguments?
- entity type?
Do representations contain any of these properties?

**Probing tasks**

- Probing task, diagnostic classifier or auxiliary prediction task [Adi et al., 2017, Conneau et al., 2018]
  - Simple classification task, classifier trained on representations
  - Performance measures how well the information is encoded
    - Assumption: Information is used for model prediction
Probing tasks

Final representation → Model → Prediction
Probing tasks

Model architectures

Final representation → Model → Prediction

Bag of embeddings
CNN
GCN (graph conv.)
(Bi-) LSTM
Self-attention
Probing tasks

Supporting linguistic features

Final representation ➔ Model ➔ Prediction

Bag of embeddings
CNN
GCN (graph conv.)
(Bi-) LSTM
Self-attention

Entity masking
Contextual word represent.
Probing tasks

Tasks

- Bag of embeddings
- CNN
- GCN (graph conv.)
- (Bi-) LSTM
- Self-attention
- Entity masking
- Contextual word represent.

Model

Final representation

Prediction

Surface properties
Syntactic properties
Semantic properties
Tasks

Surface properties

- Sentence length
- Argument distance
- Named entity between arguments
Tasks

Surface properties
- Sentence length
- Argument distance
- Named entity between arguments

Syntactic properties
- Dependency tree depth
- Shortest dependency path tree depth
- Argument order
- POS of tokens to the left and right of {head, tail}
Tasks

Surface properties
- Sentence length
- Argument distance
- Named entity between arguments

Syntactic properties
- Dependency tree depth
- Shortest dependency path tree depth
- Argument order
- POS of tokens to the left and right of \{head, tail\}

Semantic properties
- Named entity type of \{head, tail\}
- Grammatical role of \{head, tail\}
Experiment Setup

**Probing task dataset:**
- Collect sentences from TACRED [Zhang et al., 2017] and SemEval 2010 Task 8 [Hendrickx et al., 2010]
- Assign probing task label
  - syntactic and semantic probing tasks labels via Stanford CoreNLP [Manning et al., 2014]

**Evaluation approach:**
- Train relation extraction model, e.g., on TACRED
- Evaluate accuracy of probing task model trained on final representation
Results

Overall relation extraction performance

![Graph showing overall relation extraction performance for different models: BoE, CNN, CNN, mask, Bi-LSTM, Bi-LSTM, mask, GCN, GCN, mask, S-Att., S-Att., mask. The graph displays precision (P), recall (R), and F1 scores for each model.](image-url)
Results

General probing task performance
Results

Neural network architecture
Results

Entity masking

![Graph showing probing task accuracy for different models and scenarios.](image-url)
Results

Contextual word representations

Probing task accuracy
Summary

- Extensive evaluation showed that
  - self-attentive encoders are well suited for RE
  - but perform lower on probing tasks
  - bias induced by different architectures is reflected in probing task performance
    - e.g., distance and dependency related tasks
- However, probing task performance *not correlated with RE performance*

Software libraries:

- REval: framework to develop and evaluate probing tasks for neural RE, based on SentEval [Conneau and Kiela, 2018]
- RelEx: binary RE framework based on AllenNLP [Gardner et al., 2017]
1. What linguistic aspects do neural relation extraction models focus on?
2. Where do neural relation extraction models fail, and why?

**TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task.**
Christoph Alt, Aleksandra Gabryszak and Leonhard Hennig. ACL 2020
In May, he secured $96,972 in working capital from GE Healthcare Financial Services.
In May, he secured $96,972 in working capital from GE Healthcare Financial Services.
Model errors

In May, he secured $96,972 in working capital from GE Healthcare Financial Services.
What is causing model errors?

Model → Prediction → Example → Dataset
What is causing model errors?

Model \[\rightarrow\] Prediction

Example

Dataset
What is causing model errors?

- Dataset
  - dataset bias?
  - annotation errors?
- Example
- Model
- Prediction
What is causing model errors?

Dataset -> Example -> Model -> Prediction

- Dataset bias?
- Annotation errors?
- The model?
What is causing model errors?

- Model
- Prediction
- Example
- Dataset

- dataset bias?
- annotation errors?
- properties of examples?
- the model?
What is causing model errors?

- Example properties of examples?
- Dataset bias? annotation errors?
- Prediction the model?
General approach

1. Data selection
2. Human evaluation
3. Misclassification annotation
4. Automated analysis

Examples
Evaluation

TACRED
(106k examples, 41 relations)

1. Data selection
2. Human evaluation
3. Misclassification annotation
4. Automated analysis

Examples

58
Evaluation

Data selection and human evaluation

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
Evaluation

1. Data selection
2. Human evaluation
3. Misclassification annotation
4. Automated analysis

Examples

TACRED
(106k examples, 41 relations)
Evaluation

1. Examples
2. Data selection
3. Human evaluation
4. Misclassification annotation

TACRED (106k examples, 41 relations)

Based on errors of 49 different RE models

Automated analysis
Evaluation

1. Data selection
   - TACRED (106k examples, 41 relations)
   - Based on errors of 49 different RE models

2. Human evaluation
   - Re-annotation of 5k examples (dev and test)

3. Misclassification annotation

4. Automated analysis
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>52.1 (49.1)</td>
<td>8.1 (8.4)</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>IAA</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H1,H2</td>
<td>H,C</td>
</tr>
<tr>
<td>Challenging</td>
<td>0.78</td>
<td>0.43</td>
</tr>
<tr>
<td>Control</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>All</td>
<td>0.80</td>
<td>0.53</td>
</tr>
</tbody>
</table>

- **High inter-annotator agreement (IAA)**
- Challenging set more difficult for re-annotators (H1, H2), too
- Moderate agreement between re-annotators and crowd (H, C)
- Typical crowd errors
  - incorrect positive (49%) → revised to “no relation”
  - incorrect negative (36%)

\[ \text{H1, H2} \rightarrow \text{agreement between human re-annotators} \]
\[ \text{H, C} \rightarrow \text{Average agreement between re-annotators and crowd} \]
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 from 71.5 to 79.3 F1
Evaluation

Misclassification annotation

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
Evaluation

Examples

Data selection

TACRED
(106k examples, 41 relations)

based on errors of 49 different RE models

Human evaluation

re-annotation of 5k examples (dev and test)

Misclassification annotation

Automated analysis
Evaluation

1. Data selection
   - TACRED (106k examples, 41 relations)
   - based on errors of 49 different RE models

2. Human evaluation
   - re-annotation of 5k examples (dev and test)

3. Misclassification annotation
   - 9 error categories

4. Automated analysis
### Results (2): model error categories

<table>
<thead>
<tr>
<th>Context</th>
<th>Example</th>
<th>Category</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverted Args</td>
<td>[Ruben van Assouw]_head:per, who had been on safari with his 40-year-old father <em>per:children</em> 25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Patrick]_tail:per, mother Trudy, 41, and brother Enzo, 11.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong Args</td>
<td>Authorities said they ordered the detention of Bruno’s wife, [Dayana Rodrigues]_tail:per, who was found with [Samudio]_head:per’s baby. <em>per:spouse</em> 109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ling. Distractor</td>
<td>In May, [he]_head:per secured $96,972 in working capital from [GE Healthcare Financial Services]_tail:org. <em>per:employ.of</em> 35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factuality</td>
<td>[Ramon]_head:per said he hoped to one day become an [astronaut]_head:title <em>per:title</em> 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neither he nor [Aquash]_head:per were [American]_tail:nationality citizens. <em>per:origin</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relation Def.</td>
<td>[Zhang Yinjun]_tail:per, spokesperson with one of China’s largest charity organization, the [China Charity Federation]_head:org <em>org:top.mem.</em> 96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context Ignored</td>
<td>[Bibi]_head:per, a mother of [five]_tail:number, was sentenced this month to death. <em>per:age</em> 52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Relation</td>
<td>[He]_head:per turned a gun on himself committing [suicide]_tail:causeofdeath. <em>no_relation</em> 646</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1017 examples, categorized into 9 error categories

7 categories related to context
Results (2): further error categories

<table>
<thead>
<tr>
<th>Arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Span</td>
</tr>
<tr>
<td>Entity Type</td>
</tr>
</tbody>
</table>

This is a tragic day for the **Australian** [Defence Force][head:org] ([ADF][tail:org]) [Christopher Bollyn][head:per] is an **independent** [tail:religion] journalist

The company, which [Baldino][head:org] founded in [1987][tail:date] sells a variety of drugs

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

Automated analysis

Approach:

■ Extend misclassification categories to testable hypotheses (error 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., whether a group of instances shows an above average error rate
  ■ Based on the approach of [Wu et al., 2019]
Evaluation

1. Data selection
2. Human evaluation
3. Misclassification annotation
4. Automated analysis

TACRED (106k examples, 41 relations) based on errors of 49 different RE models
re-annotation of 5k examples (dev and test)
9 error categories
Evaluation

1. Data selection
   - TACRED (106k examples, 41 relations)
   - Based on errors of 49 different RE models

2. Human evaluation
   - Re-annotation of 5k examples (dev and test)
   - 9 error categories

3. Misclassification annotation
   - 10 error groups

4. Automated analysis
Evaluation

Error groups

Surface structure

  e.g., argument
distance or
sentence length
Evaluation

Error groups

Surface structure
- e.g., argument distance or sentence length

Arguments
- e.g., head and tail entity type
Evaluation

Error groups

Surface structure
- e.g., argument distance or sentence length

Arguments
- e.g., head and tail entity type

Context
- e.g., distracting entities in context
Evaluation

Error groups

Surface structure
- e.g., argument distance or sentence length

Arguments
- e.g., head and tail entity type

Context
- e.g., distracting entities in context

Ground truth
- e.g., positive examples (excluding “no relation”)
Evaluation

Error groups

- Surface structure: e.g., argument distance or sentence length
- Arguments: e.g., head and tail entity type
- Context: e.g., distracting entities in context
- Ground truth: e.g., positive examples (excluding “no relation”)

- Compare state-of-the-art model error rates per group
  - TRE [Alt et al., 2019] \(\rightarrow\) OpenAI GPT
  - SpanBERT [Joshi et al., 2019] \(\rightarrow\) BERT, pre-trained on span level
  - KnowBERT [Peters et al., 2019] \(\rightarrow\) BERT, pre-trained jointly with entity linking
Results (3): per group error rates

<table>
<thead>
<tr>
<th>Group</th>
<th>CNN, masked</th>
<th>TRE</th>
<th>SpanBERT</th>
<th>KnowBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>13662</td>
<td>14101</td>
<td>14193</td>
<td>14279</td>
</tr>
<tr>
<td>positive</td>
<td>1918</td>
<td>2338</td>
<td>2516</td>
<td>2561</td>
</tr>
<tr>
<td>negative</td>
<td>11744</td>
<td>11763</td>
<td>11677</td>
<td>11718</td>
</tr>
<tr>
<td>argdist=1</td>
<td>155</td>
<td>171</td>
<td>181</td>
<td>184</td>
</tr>
<tr>
<td>argdist&gt;10</td>
<td>2788</td>
<td>2788</td>
<td>2800</td>
<td>2791</td>
</tr>
<tr>
<td>sentlen&gt;30</td>
<td>8238</td>
<td>8566</td>
<td>8610</td>
<td>9461</td>
</tr>
<tr>
<td>has_distractor</td>
<td>2945</td>
<td>3016</td>
<td>3046</td>
<td>3074</td>
</tr>
<tr>
<td>has_coref</td>
<td>4939</td>
<td>5029</td>
<td>5018</td>
<td>5059</td>
</tr>
<tr>
<td>per:*</td>
<td>8971</td>
<td>9334</td>
<td>9419</td>
<td>9461</td>
</tr>
<tr>
<td>org:*</td>
<td>4691</td>
<td>4767</td>
<td>4774</td>
<td>4818</td>
</tr>
<tr>
<td>per:loc</td>
<td>1288</td>
<td>1378</td>
<td>1376</td>
<td>1351</td>
</tr>
<tr>
<td>same_nertag</td>
<td>5023</td>
<td>5129</td>
<td>5168</td>
<td>5193</td>
</tr>
<tr>
<td>same_nertag&amp;positive</td>
<td>262</td>
<td>450</td>
<td>528</td>
<td>528</td>
</tr>
</tbody>
</table>
Results (3): per group error rates

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

- Manual re-annotation of 5k most challenging TACRED examples (development and test split)
  - Results: Release of revised dataset

Lessons learned:

- 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
A clear definition of the (practical) purpose of the task
  e.g., IE for knowledge base construction vs. question answering
Probing → causation, i.e., encoded information actually impacts prediction [Elazar et al. 2020]
More detailed investigation of datasets and linguistic phenomena
  e.g., context vs. entity mentions [Peng et al., 2020] or via challenge sets [Rosenman et al., 2020]
Pre-training focused on semantic relations
Thank you!
Questions?

Github: github.com/DFKI-NLP

Website: christophalt.github.io

Credits: Some of the icons and graphics were created by Slidesgo, including icons from Flaticon.
References


- [Manning et al., 2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. ACL 2014 (System Demonstrations).


