Relation Extraction with Explanation

Hamed Shahbazi, Xiaoli Z. Fern, Reza Ghaeini, Prasad Tadepalli
Oregon State University

EMNLP 2020 (short)

*Some content from Hamed Shahbazi.*
Problem Setup

- **Given entity pair and its bag w/ labels**
  - a bag is a set of sentences that contain mentions of both entities
  - labeled with a set of relations between the two entities from a KB

- **Training objective**
  - Predict what relation holds for a given entity pair and its bag
  - Assign more importance to the sentences which are supporting a relation

---

1. They live in **Windhoek**, the **Namibia**’s capital and its largest city.
2. **Windhoek** in located in central **Namibia** in the Khomas Highland plateau area.
3. ...

- Entity pair = (Windhoek, Namibia)
- Bag annotation = [contain, capital]

\[
P(r = k | B_{i,j}) \quad k \in 1 \ldots K
\]
Two Baselines

- **DirectSup** (Beltagy et al., 2019)
  - assigns an importance weight for each sentence based on the output of a binary classifier learned from an additional direct supervision data
    - weights $\{\alpha_1, \ldots, \alpha_N\}$
    - bag representation
      - $\bar{x} = \text{Max-pool}(\{\alpha_1 x_1, \ldots, \alpha_N x_N\})$
      - $P(r = k|B) = \sigma(\bar{x} \cdot r_k + b_k)$

- **CNNs+ATT** (Lin et al., 2016)
  - same CNN sentence encoder, but use attention weights
    - $\alpha_{k,n} = \frac{\exp(x_n A q_k)}{\sum_{n=1}^N \exp(x_n A q_k)}$
  - bag representation
    - $\bar{x}_k = \sum_{n=1}^N \alpha_{k,n} x_n$
    - $P(r = k|B) = \sigma(\bar{x}_k \cdot r_k + b_k)$
Two Baselines

- Baselines provide sentence level aggregation weights
- Baselines’s explainability is not evaluated
  - This is what this paper does
- **Two approaches** to improve explainability (as well as performance)
Approach 1: Enhancing Sentence Representation for Explanation

- **Motivation**
  - The embeddings of entity mentions may contain sufficient evidence to support a relation regardless of the rest of the sentence.

- **Idea**
  - Substitute the entity mentions with their Fine-Grained Entity Types (FGET) to force the model to identify the relations through **textual evidence** of the sentences rather than word embeddings of mentions.

  ```plaintext
  They live in Windhoek, the Namibia’s capital and its largest city.
  ```

  ```plaintext
  They live in City, the Country’s capital and its largest city.
  ```
Approach 2: Learning from Distractors

- **Idea**
  - For any bag annotated with any relation $k$:
    - Generate a distractor sentence $x_d$ which does not support the relation $k$
    - Put $x_d$ into the bag (resulting in an augmented bag)
    - Mitigate the contribution of $x_d$ in predicting relation $k$

- **Distractor sentence: two requirements**
  - it belongs to a bag that has the same FGET pair
  - the bag is not annotated with relation label $k$
Approach 2: Learning from Distractors

- Mitigate the contribution of $x_d$ in predicting relation $k$

\[
\text{Loss} = \text{original loss} + \lambda \times \text{distractor loss}
\]

- Distractor loss
  - the first term ensures that the contribution of the distractor is lower than the maximum contribution of all the sentences in the original bag
  - the second term reduces the absolute contribution of the distractor

\[
l_d = \sum_k \left( \max(0, \gamma + c_{x_d,k} - \max_{x \in B_{i,j}} c_x,k) + \max(0, c_{x_d,k}) \right)
\]

- Contribution measure $c_{x,k}$
  - Saliency
  - Gradient $\times$ input (GI)
  - Leave One Out (loo)
Experiments

- Relation Extraction Performance (FB-NYT)
  - (+/-)E means whether to concatenate entity embeddings with bag representation

<table>
<thead>
<tr>
<th>model</th>
<th>AUC (-E)</th>
<th>AUC (+E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNNs+ATT</td>
<td>25.1</td>
<td>-</td>
</tr>
<tr>
<td>DirectSup</td>
<td>26.4</td>
<td>28.1</td>
</tr>
<tr>
<td>CNNs+ATT +F</td>
<td>26.1</td>
<td>31.5</td>
</tr>
<tr>
<td>DirectSup +F</td>
<td>26.9</td>
<td>33.3</td>
</tr>
<tr>
<td>CNNs+ATT +FE</td>
<td>27.4</td>
<td>33.1</td>
</tr>
<tr>
<td>DirectSup +FE</td>
<td>27.6</td>
<td>33.4</td>
</tr>
<tr>
<td>CNNs+ATT +LD</td>
<td>27.1</td>
<td>33.6</td>
</tr>
<tr>
<td>CNNs+ATT +F +LD</td>
<td>27.7</td>
<td>33.9</td>
</tr>
<tr>
<td>DirectSup +F +LD</td>
<td><strong>27.8</strong></td>
<td><strong>34.1</strong></td>
</tr>
</tbody>
</table>

F: Replace entity mention with FGET
FE: Replace entity mention with concatenation of FGET and entity mention
LD: Learning from distractor

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>472,963</td>
<td>172,448</td>
</tr>
<tr>
<td>Positive bags</td>
<td>16,625</td>
<td>1,950</td>
</tr>
<tr>
<td>Negative bags</td>
<td>236,811</td>
<td>94,917</td>
</tr>
</tbody>
</table>

Table 1: FB-NYT modified dataset.
Experiments

- Relation Extraction Performance (FB-NYT)
Experiments

- Evaluation of Explanations
  - manually annotate the positive bags of the test split of FB-NYT with ground-truth explanations
  - for each pair of (sentence, relation) in a bag, the sentence is either a **rationale** (supportive) to the relation or it is **irrelevant**
  - expl-eval: (bag-id, k, rationale sentence, irrelevant sentence)
Experiments

- **Evaluation of Explanations**
  - apply the contribution measures (S, GI, etc.) to produce sentence importance scores for the test set and compute the Kendall Tau correlations for the importance scores using `expl-eval`.

<table>
<thead>
<tr>
<th>model</th>
<th>loo (H)</th>
<th>loo (L)</th>
<th>$S_{x_{n,k}}$ (H)</th>
<th>$S_{x_{n,k}}$ (L)</th>
<th>$GI_{x_{n,k}}$ (H)</th>
<th>$GI_{x_{n,k}}$ (L)</th>
<th>$\alpha_{x_{n}}$ (H)</th>
<th>$\alpha_{x_{n}}$ (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNNs+ATT</td>
<td>0.16</td>
<td>-0.08</td>
<td>0.19</td>
<td>-0.02</td>
<td>0.20</td>
<td>0.04</td>
<td>0.69</td>
<td>0.21</td>
</tr>
<tr>
<td>DirectSup</td>
<td>0.19</td>
<td>0.12</td>
<td>0.08</td>
<td>0.15</td>
<td>0.29</td>
<td>0.19</td>
<td>0.26</td>
<td>-0.12</td>
</tr>
<tr>
<td>CNNs+ATT +F</td>
<td>0.21</td>
<td>0.10</td>
<td>0.36</td>
<td>0.03</td>
<td>0.23</td>
<td>0.00</td>
<td>0.73</td>
<td>0.11</td>
</tr>
<tr>
<td>DirectSup +F</td>
<td>0.24</td>
<td>0.15</td>
<td>0.31</td>
<td>-0.19</td>
<td>0.40</td>
<td>-0.17</td>
<td>0.28</td>
<td>0.15</td>
</tr>
<tr>
<td>CNNs+ATT +FE</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.21</td>
<td>-0.14</td>
<td>0.20</td>
<td>-0.20</td>
<td>0.24</td>
<td>0.01</td>
</tr>
<tr>
<td>DirectSup +FE</td>
<td>0.14</td>
<td>-0.12</td>
<td>0.19</td>
<td>-0.10</td>
<td>0.29</td>
<td>0.06</td>
<td>0.17</td>
<td>-0.11</td>
</tr>
<tr>
<td>CNNs+ATT +LD</td>
<td>0.18</td>
<td>-0.01</td>
<td>0.22</td>
<td>0.10</td>
<td>0.21</td>
<td>0</td>
<td>0.67</td>
<td>0.11</td>
</tr>
<tr>
<td>CNNs+ATT +LD +F</td>
<td>0.22</td>
<td>-0.11</td>
<td>0.43</td>
<td>0.09</td>
<td>0.28</td>
<td>0.07</td>
<td>0.70</td>
<td>0.12</td>
</tr>
<tr>
<td>DirectSup +LD +F</td>
<td>0.23</td>
<td>0.14</td>
<td>0.38</td>
<td>0.01</td>
<td>0.49</td>
<td>0.20</td>
<td>0.45</td>
<td>0.02</td>
</tr>
</tbody>
</table>

H: High confidence $P(\tau) \in [0.76, 1.0]$
L: Low confidence $P(\tau) \in [0, 0.25]$

**Observations:**
- Higher corr on H
- F helps while FE hurts
- LD helps explanation

**Table 3:** Kendall correlations for top confidence and least confidence range.
Summary

- This paper presents a test set with ground-truth sentence-level explanations to evaluate the quality of explanations afforded by relation extraction models.
  - expl-eval

- They introduce two methods for improving relation extraction with explanation.
  - replacing the entity mentions with their fine-grained entity types for sentence representation
  - augment the bags with automatically generated distractor sentences
Thanks