Neural Discourse Structure for Text Categorization

Yangfeng Ji and Noah Smith
Presented by Ji Li

Outline

• Motivation
• Background
• Model
• Experiment results
• Discussion and conclusion
Outline

• Motivation
• Background
• Model
• Experiment results
• Discussion and conclusion
Text Categorization

• Important for applications like sentiment analysis, inferring authorship and making predictions

• How to represent text?
  – Bag of words
  – Representation learning

• How does different part of text contribute to the representation?
  – Different part of text may influence the text category in different weight
Outline

• Motivation
• Background
• Model
• Experiment results
• Discussion and conclusion
Rhetorical Structure Theory

Document can be represented by a tree.

Elementary Discourse Unit (EDU), connected via discourse relations.

[Although the food was amazing]$^A$ [and I was in love with the spicy pork burrito,]$^B$ [the service was really awful.]$^C$ [We watched our waiter serve himself many drinks.]$^D$ [He kept running into the bathroom]$^E$ [instead of grabbing our bill.]$^F$
Discourse Parsing from Linear Projection

- Shift reduced discourse parsing
  - Maintain a stack and a queue
  - Consider features of 3 EDUs, \( v = [v1; v2; v3] \)

- Formulate the decision as a multi-class shift-reduce classifier

\[
\hat{m} = \arg\max_{m \in \{1, \ldots, C\}} w^T_m f(v; A)
\]

Outline

• Motivation
• Background
• Model
• Experiment results
• Discussion and conclusion
Model

- Recursive neural network build on discourse dependency tree
- Distributed representation of EDU achieved with bidirectional LSTM
- Let $v_i$ denote the vector representation of EDU $i$ and its descendants
  - If $v_i$ is leaf: $v_i = tanh(e_i)$
  - Other wise:

\[
    v_i = \tanh \left( e_i + \sum_{j \in \text{children}(i)} \alpha_{i,j} W_{r_{i,j}} v_j \right)
\]

\[
    \alpha_{i,j} = \sigma \left( e_i^\top W_\alpha v_j \right)
\]
Model ablations and baselines

• Unlabeled Model

\[ v_i = \tanh \left( e_i + \sum_{j \in \text{children}(i)} \alpha_{i,j} v_j \right) \]

• Root
  – Use discourse structure to select the root and use only the root EDU representation

• Additive
  – Use all EDUs with simple average
Outline

• Motivation
• Background
• Model
• Experiment results
• Discussion and conclusion
## Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Yelp</th>
<th>MFC</th>
<th>Debates</th>
<th>Movies</th>
<th>Bills</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prior work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Yang et al. (2016)</td>
<td>71.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. Card et al. (2016)</td>
<td>—</td>
<td>56.8</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. Yogatama and Smith (2014)</td>
<td>—</td>
<td>—</td>
<td>74.0</td>
<td>—</td>
<td>88.5</td>
</tr>
<tr>
<td>4. Bhatia et al. (2015)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>82.9</td>
<td>—</td>
</tr>
<tr>
<td>5. Hogenboom et al. (2015)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>71.9</td>
</tr>
<tr>
<td><strong>Variants of our model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. ADDITIVE</td>
<td>68.5</td>
<td>57.6</td>
<td>69.0</td>
<td>82.7</td>
<td>80.1</td>
</tr>
<tr>
<td>7. ROOT</td>
<td>54.3</td>
<td>51.2</td>
<td>60.3</td>
<td>68.7</td>
<td>70.5</td>
</tr>
<tr>
<td>8. UNLABELED</td>
<td><strong>71.3</strong></td>
<td><strong>58.4</strong></td>
<td><strong>75.7</strong></td>
<td><strong>83.1</strong></td>
<td><strong>78.4</strong></td>
</tr>
<tr>
<td>9. FULL</td>
<td><strong>71.8</strong></td>
<td>56.3</td>
<td><strong>74.2</strong></td>
<td><strong>79.5</strong></td>
<td><strong>77.0</strong></td>
</tr>
</tbody>
</table>

Table 2: Test-set accuracy across five datasets. Results from prior work are reprinted from the corresponding publications. Boldface marks performance stronger than the previous state of the art.
## Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Yelp</th>
<th>MFC</th>
<th>Debates</th>
<th>Movies</th>
<th>Bills</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prior work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Yang et al. (2016)</td>
<td>71.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Card et al. (2016)</td>
<td></td>
<td>56.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Yogatama and Smith (2014)</td>
<td></td>
<td></td>
<td>74.0</td>
<td></td>
<td>88.5</td>
</tr>
<tr>
<td>4. Bhatia et al. (2015)</td>
<td></td>
<td></td>
<td></td>
<td>82.9</td>
<td></td>
</tr>
<tr>
<td>5. Hogenboom et al. (2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>71.9</td>
</tr>
<tr>
<td><strong>Variants of our model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. ADDITIVE</td>
<td>68.5</td>
<td>57.6</td>
<td>69.0</td>
<td>82.7</td>
<td>80.1</td>
</tr>
<tr>
<td>7. ROOT</td>
<td>54.3</td>
<td>51.2</td>
<td>60.3</td>
<td>68.7</td>
<td>70.5</td>
</tr>
<tr>
<td><strong>8. UNLABELED</strong></td>
<td>71.3</td>
<td>58.4</td>
<td>75.7</td>
<td>83.1</td>
<td>78.4</td>
</tr>
<tr>
<td><strong>9. FULL</strong></td>
<td>71.8</td>
<td>56.3</td>
<td>74.2</td>
<td>79.5</td>
<td>77.0</td>
</tr>
</tbody>
</table>

Table 2: Test-set accuracy across five datasets. Results from prior work are reprinted from the corresponding publications. Boldface marks performance stronger than the previous state of the art.
## Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Yelp</th>
<th>MFC</th>
<th>Debates</th>
<th>Movies</th>
<th>Bills</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prior work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Yang et al. (2016)</td>
<td>71.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. Card et al. (2016)</td>
<td>—</td>
<td>56.8</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. Yogatama and Smith (2014)</td>
<td>—</td>
<td>—</td>
<td>74.0</td>
<td>—</td>
<td>88.5</td>
</tr>
<tr>
<td>4. Bhatia et al. (2015)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>82.9</td>
<td>—</td>
</tr>
<tr>
<td>5. Hogenboom et al. (2015)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>71.9</td>
</tr>
<tr>
<td><strong>Variants of our model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. ADDITIVE</td>
<td>68.5</td>
<td>57.6</td>
<td>69.0</td>
<td>82.7</td>
<td>80.1</td>
</tr>
<tr>
<td>7. ROOT</td>
<td>54.3</td>
<td>51.2</td>
<td>60.3</td>
<td>68.7</td>
<td>70.5</td>
</tr>
<tr>
<td>8. UNLABELED</td>
<td>71.3</td>
<td>58.4</td>
<td>75.7</td>
<td>83.1</td>
<td>78.4</td>
</tr>
<tr>
<td>9. FULL</td>
<td>71.8</td>
<td>56.3</td>
<td>74.2</td>
<td>79.5</td>
<td>77.0</td>
</tr>
</tbody>
</table>

Table 2: Test-set accuracy across five datasets. Results from prior work are reprinted from the corresponding publications. Boldface marks performance stronger than the previous state of the art.
Case Study

From DPLP:

From DPLP:  

\[ 2C \xrightarrow{0.70} 2B \xrightarrow{0.87} 2A \xrightarrow{0.61} 2D \]

Evaluation Evaluation Elaboration

[I love these people.]\(^{2A}\) [They are very friendly and always ask about my life.]\(^{2B}\) [They remember things I tell them then ask about it the next time I’m in.]\(^{2C}\) [Patrick and Lily are the best but everyone there is wonderful in their own ways.]\(^{2D}\)

(b) true label: 5, predicted label: 5
From DPLP:

```
Attribution 0.32  0.16  Elaboration
  3A    3C 0.62    3D 0.32    3E 0.47    3F
Elaboration     Elaboration      Elaboration
```

Manually constructed:

```
Cause
  3F

Explanation
  3A Background  3B Explanation  3E

Explanation
  3C                3D
```

[We use to visit this pub 10 years ago because they had a nice english waitress and excellent fish and chips for the price.] \(^{3A}\) [However we went back a few weeks ago and were disappointed.] \(^{3B}\) [The price of the fish and chip dinner went up and they cut the portion in half.] \(^{3C}\) [No one assisted us in putting two tables together we had to do it ourselves.] \(^{3D}\) [Two guests wanted a good English hot tea and they didn't brew it in advance.] \(^{3E}\) [So we've decided there are newer and better places to eat fish and chips especially up in north phoenix.] \(^{3F}\)
Other experiment

Further improvement on Discourse structure can potentially enhance performance.
Outline

• Motivation
• Background
• Model
• Experiment results
• Discussion and conclusion
Discussion and Conclusion

• Automatically-derived discourse structure can be helpful to text categorization.

• Two major limitations:
  – Performance of discourse parser.
  – Domain mismatch between training corpus and future test data.

• Motivates further improvement of discourse parser and exploration of domain adaptation methods.
Thank you