Combining Deterministic Dependency Parsing and Linear Classification for Robust RTE

Alexander Volokh, Günter Neumann, Bogdan Sacaleanu
DFKI, Saarbrücken, Germany
Outline

• RTE6 Main Task:
  - Former DFKI approaches for RTE
  - Feature-based approach of this year

• Parser comparison via RTE
General RTE Architecture
Architecture with Voting

- Expert #1
- Expert #2
- Expert #3

Voting

Context
- Text

RTE System

Hypothesis

entailed?

Yes/No
Architecture with Voting

- Syntactic Information
- Semantic Roles
- Temporal Information

Voting

Context
- Text

RTE System

Hypothesis

entailed?

Yes/No
Architecture with Voting

- Syntactic Information
- Semantic Roles
- Temporal Information

Voting
Majority decision

RTE System

Context
Text
Hypothesis

entailed?

Yes/No

Expert #1
Expert #2
Expert #3

Voting
Fallback strategy
Architecture with Voting

- Syntactic Information
- Semantic Roles
- Temporal Information
- Voting
  - Majority decision
- Bag of Words
- RTE System
- Context
- Text
- Hypothesis
- entailed?
- Yes/No

Expert #1
Expert #2
Expert #3

Majority decision

Voting
RTE-6 Changes

- This year's RTE task:
  - Explore the influence of the different knowledge sources
  - Each of the "arrows" is a factor when computing the contribution
  - Main / Fallback strategies make the computation even more difficult

→ Principal change of architecture
Feature-Based Architecture

Common Feature Representation

- Named Entity Features
- Structural Features
- Word-level Features

entailed?

Yes/No
Ablation Tests

• **Official F-Score:** 38.26

• **Features:**
  - Dependency structure comparison(1,2,3,4)
  - Overlap and similarity on word level(4,5,7)
  - Named entity overlap and type similarity(6)
  - Coreference resolution(8)

• **Classifier Parameters:**
  - Threshold for YES/NO classification

<table>
<thead>
<tr>
<th>Test #</th>
<th>F Measure</th>
<th>Impact</th>
<th>Left out Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.31</td>
<td>2.95</td>
<td>1 (root features)</td>
</tr>
<tr>
<td>2</td>
<td>39.11</td>
<td>-0.85</td>
<td>2 (depth 1 features)</td>
</tr>
<tr>
<td>3</td>
<td>38.54</td>
<td>-0.28</td>
<td>3 (depth 2 features)</td>
</tr>
<tr>
<td>4</td>
<td>33.27</td>
<td>4.99</td>
<td>4 (word form and pos features)</td>
</tr>
<tr>
<td>5</td>
<td>19.22</td>
<td>19.04</td>
<td>6 (content word features)</td>
</tr>
<tr>
<td>6</td>
<td>36.04</td>
<td>2.22</td>
<td>7 (named entities features)</td>
</tr>
<tr>
<td>7</td>
<td>38.49</td>
<td>-0.23</td>
<td>5 (WordNet similarity features)</td>
</tr>
<tr>
<td>8</td>
<td>36.72</td>
<td>1.54</td>
<td>Coreference resolution features. No additional features were introduced or left out, but the content of all T-H-pairs was first processed with the LingPipe coreference resolution tool.</td>
</tr>
<tr>
<td>9</td>
<td>39.10</td>
<td>-0.84</td>
<td>Threshold 0.15</td>
</tr>
<tr>
<td>10</td>
<td>39.11</td>
<td>-0.85</td>
<td>Threshold 0.13</td>
</tr>
</tbody>
</table>
PETE

- Parser Evaluation using Textual Entailments (PETE), task #12 at SemEval-2, 2010 (Yuret et al.)
- **subject-verb** dependency
  "John kissed Mary." --> "John kissed somebody."
- **verb-object** dependency
  "John kissed Mary." --> "Mary was kissed."
- **noun-modifier** dependency
  "The big red boat sank." --> "The boat was big."

- Natural evaluation instead of LAS/UAS
- Formalism independent comparison
Parser Comparison

- Two parsers: MaltParser (Nivre et al.), MDPARSER

<table>
<thead>
<tr>
<th></th>
<th>Parsing Time</th>
<th>Sentences per Second</th>
<th>Tokens per Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDPARSER</td>
<td>73.188s</td>
<td>46.128</td>
<td>1015.55</td>
</tr>
<tr>
<td>MaltParser</td>
<td>1954.684s</td>
<td>1.73</td>
<td>38.02</td>
</tr>
</tbody>
</table>

- RTE results:
  - MDPARSER: f-score = 38.26
  - MaltParser: f-score = 39.81
Summary

• We thank RTE6 for the opportunities:
  - implementation of a robust RTE approach
  - parser comparison in a real-world application

• RTE6 shows that no specific data is necessary to evaluate parsers
  - on the contrary to PETE data more than syntax was necessary, but the comparison worked