The Effects of Lexical Resource Quality on Preference Violation Detection

Jesse Dunietz, Lori Levin and Jaime Carbonell
ACL 2013
August 6, 2013
Resources often play second fiddle to algorithm development.

The usual model:
Curating resources interacts synergistically with algorithmic improvements.

A better model:

- Performance
- Algorithms
- Lexical Resources

Linguistic resource improvements needed for better POS tagging [Manning 2011]
The politician pruned laws regulating plastic bags, and created new fees for inspecting dairy farms.

Preference violation $\rightarrow$ likely metaphor [Wilks 1978]

Corpus:
- 715 sentences
- 2 annotators following manual
- Each clause marked for violations under most concrete meaning
The politician pruned laws regulating plastic bags

The politician pruned laws regulating plastic bags

AGENT: CARVE-21.2-2
+ INT_CONTROL

PATIENT: +CONCRETE

The politician pruned laws regulating plastic bags

+ animate_being.n.01 OR + physical_object.n.01 OR
+ person.n.01 OR + matter.n.01 OR
+ machine.n.01 OR + substance.n.01
...

Violation!
DAVID unifies 6 tools & resources to detect preference violations.

Diagram:
- Text → Senna → Proposition Bank → SemLink → VerbNet
- WordNet
- WordNet: A lexical database for English
- Violation?
Initial results suggested the technique was unusable.

End-to-end $F_1$ score: 28%
Error analysis revealed that <24% of errors were endemic to the technique.

Sources of DAVID errors on 90 randomly selected sentences (total errors: 32)

- Endemic errors: 23.40%
- Resource errors: 39.10%
- Tool errors: 37.50%
To explore the algorithm’s viability, we hand-corrected some resources/parses.

**VerbNet**

**Seek:**
\[\text{Agent} [+\text{Animate}]\]

**SemLink**

**KEEP.04 \rightarrow SUSTAIN-55.6**

**KEEP.04 \rightarrow KEEP-15.2**

20 VerbNet classes updated

20 SemLink entries updated

**Senna**

It’s named after Mr. Scott’s father.

ARG1 V ARGM-TMP

ARG1 V ARG2
Results: resource improvements > tool improvements.

DAVID performance with various degrees of correction

- 30 sentences:
  - Uncorrected: 40%
  - Tools corrected: 44%
  - Resources corrected: 50%

- 60 sentences:
  - Uncorrected: 31%
  - Tools corrected: 31%
  - Resources corrected: 45%

- 90 sentences:
  - Uncorrected: 30%
  - Tools corrected: 38%
  - Resources corrected: 42%
Results: resource improvements & tool improvements add superlinearly.

DAVID performance with various degrees of correction

<table>
<thead>
<tr>
<th>F1 score</th>
<th>30 sentences</th>
<th>60 sentences</th>
<th>90 sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrected</td>
<td>40%</td>
<td>31%</td>
<td>30%</td>
</tr>
<tr>
<td>Tools corrected</td>
<td>44%</td>
<td>31%</td>
<td>38%</td>
</tr>
<tr>
<td>Resources corrected</td>
<td>50%</td>
<td>45%</td>
<td>42%</td>
</tr>
<tr>
<td>Both corrected</td>
<td>70%</td>
<td>69%</td>
<td>62%</td>
</tr>
</tbody>
</table>
Our resource improvements generalize to novel sentences.

**DAVID performance on 625 uncorrected sentences with various degrees of correction**

- None: 27.98%
- Out-of-date resources: 28.15%
- 30 sentences: 28.82%
- 60 sentences: 28.74%
- 90 sentences: 28.99%
NLP tools demand heavy investment in resource quality.

We have demonstrated that:

• Preference violations can be detected with lexical resources
• Resource quality can matter more than tool performance
• Resource and tool improvements add synergistically