Derivational Smoothing for Syntactic Distributional Semantics

Sebastian Padó*, Jan Šnajder†, and Britta Zeller*

*Institute for Computational Linguistics, Heidelberg University
†Faculty of Electrical Engineering and Computing, Zagreb University

The 51st Annual Meeting of the Association for Computational Linguistics
August 6, 2013
Distributional Semantics

- Representation of word meaning as vectors
  - Vector components: co-occurrences with context features
  - Firth (1957): *You shall know a word by the company it keeps*

  \[
  \begin{array}{l}
  \text{Peter convinced himself to write reports} \\
  \Rightarrow \\
  \hline
  \text{report} \\
  \text{Peter} & 1 \\
  \text{convince} & 1 \\
  \text{write} & 1 \\
  \end{array}
  \]

- Vector similarity approximates semantic similarity
  - Simple, unsupervised induction of word meaning
  - Used in variety of tasks (Turney and Pantel, 2010)
Lexical (word) context captures **topical** similarity

Syntactic (word-relation) context captures **relational** similarity

- Can model fine-grained information (Baroni and Lenci, 2010)
- More appropriate for free word order languages
A problem for syntactic vector spaces: Sparsity

- Syntactic vector spaces are very **sparse**
  - Even if constructed from very large corpora
- Reason: Less cooccurrences

Peter convinced himself to write reports

\[ \Rightarrow \quad \text{report} \quad \text{write} \quad 1 \]

- Many word pairs receive semantic similarities of zero
  - Real dissimilarity or missing data?
The question
Where can we get semantic relatedness information to smooth distributional similarity?

The answer: Derivational morphology

- Consider derivational families:
  - Words that are derived from one another have similar meaning
  - Available from resources like CatVar (Habash and Dorr, 2003)
If vectors are sparse, do not compute semantic similarity directly.
Instead, back off to less sparse members of derivational families.

\[ \text{sim}(\text{arguably, debatably}) = 0 \]
\[ \text{sim}(\text{argue, debate}) > 0 \]

\[ \text{smoothed-sim}(\text{arguably, debatably}) = f(\text{arguably, debatably}) \]

(Similar to back-off to less sparse \( n - 1 \) grams in LMs)
Derivational parameters: Two parameters

1. **Smoothing trigger:** When is a vector considered too sparse?
   - Smooth always
   - Smooth only if $\text{sim}(l_1, l_2) = 0$ (or undefined)

2. **Smoothing scheme:** How to bring in derivational family
   - **maxSim:** Consider most similar pair between families
   - **avgSim:** Consider average similarity of all pairs
   - **centSim:** Consider similarity of family centroids
Experiments

Language choice: German

- Resource situation comparable to English, but not quite as good
- Derivation important process of word formation

Distributional models

- Base Model: German Distributional Memory $\text{DM.DE}$ (Padó and Utt, 2012)
  - 900M-token SDEWAC web corpus (Faaß et al., 2010)
- $\text{DERIVBASE}$ derivational families (Zeller et al., 2013)
  - Rule-based resource for German, focus on precision
  - 18,000 non-singleton families covering 60,000 lemmas
- Baseline: Bag-of-words models (same corpus)
Task 1: Synonym choice

- 980 targets with four candidates each (Reader’s Digest)
  “Which term is antiquated most similar to? 
  (a) venerable, (b) old, (c) unusable, (d) outdated?”
- Prediction: candidate with max cosine similarity to target
- Evaluation: Accuracy (%) + Coverage (%)

Task 2: Word similarity prediction

- 350 pairwise judgments on 5-point scale (Zesch et al., 2007)
  (monkey, macaque) ⇒ 4
  (office, tiger) ⇒ 1
- Prediction: Cosine similarity
- Evaluation: Correlation (Pearson’s r) + Coverage (%)
### Results: Synonym choice

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc. %</th>
<th>Cov. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM.DE, unsmoothed</td>
<td>53.7</td>
<td>80.8</td>
</tr>
<tr>
<td>DM.DE, smooth always</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AvgSim</td>
<td>46.0</td>
<td>86.6</td>
</tr>
<tr>
<td>MaxSim</td>
<td>50.3</td>
<td>86.6</td>
</tr>
<tr>
<td>CentSim</td>
<td>49.1</td>
<td>86.6</td>
</tr>
<tr>
<td>DM.DE, smooth if sim = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AvgSim</td>
<td>52.6</td>
<td>86.6</td>
</tr>
<tr>
<td>MaxSim</td>
<td>51.2</td>
<td>86.6</td>
</tr>
<tr>
<td>CentSim</td>
<td>51.3</td>
<td>86.6</td>
</tr>
<tr>
<td>BoW “baseline”</td>
<td>56.9</td>
<td>98.5</td>
</tr>
</tbody>
</table>

- Gain in coverage (+6%), but small loss in accuracy (-1%)
  - BoW “baseline” performs best
- Conservative trigger (smooth if necessary) works best
## Results: Semantic similarity

<table>
<thead>
<tr>
<th>Model</th>
<th>r</th>
<th>Cov. %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DM.DE, unsmoothed</strong></td>
<td>.44</td>
<td>58.9</td>
</tr>
<tr>
<td>DM.DE, smooth always</td>
<td></td>
<td></td>
</tr>
<tr>
<td>avgSim</td>
<td>.30</td>
<td>88.0</td>
</tr>
<tr>
<td>maxSim</td>
<td>.43</td>
<td>88.0</td>
</tr>
<tr>
<td>centSim</td>
<td>.44</td>
<td>88.0</td>
</tr>
<tr>
<td><strong>DM.DE, smooth if sim = 0</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avgSim</td>
<td>.43</td>
<td>88.0</td>
</tr>
<tr>
<td>maxSim</td>
<td>.42</td>
<td>88.0</td>
</tr>
<tr>
<td>centSim</td>
<td><strong>.47</strong></td>
<td><strong>88.0</strong></td>
</tr>
<tr>
<td><strong>BoW baseline</strong></td>
<td>.36</td>
<td><strong>94.9</strong></td>
</tr>
</tbody>
</table>

- Again, conservative trigger works best
- Big increase in coverage (+30%), small increase in correlation
Semantic similarity benefits more from derivational smoothing than synonym choice

- Derivational families contain related words, not synonyms
Summary

- Sparsity is a problem for syntax-based distributional models
  - “Derivational smoothing”: Back off from rare word to derivational family
- Initial experiments
  - Conservative trigger (smooth only when sim=0) works best
  - Jury still out on smoothing scheme (combination method)
- Future work
  - More experiments on smoothing schemes
  - Use richer information about derivational families


