Aims

Current Aims
- Predict prepositions in native English text.
- For predictions, use a purely surface-based, “web-as-corpus” approach (e.g., Lapata & Keller, 2005).
- Determine the tradeoff between
  - Informativeness of context (→ the size of the context n-gram)
  - Data-spariness (→ can we find the n-gram on the web?)

Mid-Term Aims
- Correct prepositions in text written by learners of English.

Related Work

Correct use of prepositions is a major problem for learners of English.

→ The computational analysis of preposition usage has attracted significant attention in recent years in the field of Intelligent Computer Assisted Language Learning.

(De Felice & Pulman, 2009; De Felice, 2008; Lee & Knutsson, 2008; Gamon et al., 2008; Chodorow et al., 2007; Tetreault & Chodorow, 2008a,b)

Most of the approaches first tackle the task of predicting prepositions in native language as a point of reference.

- The task is usually approached as a classification problem:
  - The classes are the prepositions.
  - The instances to be classified are the contexts.
  - Contexts are represented by a rich set of linguistic features (POS tags, PP attachment sites, WordNet classes of PP object and modified item, etc.).

Our Approach

Using purely surface-based predictions, how much information can we find in the immediate distributional context of the preposition?

Data

- Random sample of section J of the BNC XML Edition (8060 sentences)
- We extract one prediction task for each occurrence of one of the top nine prepositions in the BNC of, to, in, for, on, with, at, by, from
  → Same corpus section and predictions as used by De Felice (2008).

Prediction Tasks

Source Sentence

But for the young, it is rather a question of the scales falling from their eyes, and having nothing to believe in any more.

Prediction Task

But for the young, it is rather a question of the scales falling from their eyes, and having nothing to believe in any more.

1. rather a question of the scales falling 2. rather a question to the scales falling 3. rather a question in the scales falling 4. rather a question from the scales falling

Prediction of Prepositions Using Web Counts

- Determine the number of hits for each 7-gram in a cohort using Yahoo’s BOSS service.
- Pick the preposition for which the most hits were found.
- In case of no hits, back off to lower n-gram orders using overlap backoff.

Results

Summary

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
<th>Informativeness/Data Tradeoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>De Felice et al.</td>
<td>79.03%</td>
<td>70.06%</td>
</tr>
<tr>
<td>Tetreault and Chodorov (2008)</td>
<td>75.50%</td>
<td>68.40%</td>
</tr>
<tr>
<td>Fields and Tetreault (2008)</td>
<td>70.60%</td>
<td>68.40%</td>
</tr>
<tr>
<td>Human agreement</td>
<td>28.94%</td>
<td>28.94%</td>
</tr>
<tr>
<td>This paper – surface-based approach</td>
<td>76.50%</td>
<td>76.50%</td>
</tr>
</tbody>
</table>

Detailed Results

<table>
<thead>
<tr>
<th>Order</th>
<th>Sum</th>
<th>1-grams (1 + prep + 3)</th>
<th>2-grams</th>
<th>3-grams (truncated 7-gram)</th>
<th>4-grams (truncated 7-gram)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-off 6</td>
<td>8080</td>
<td>8080</td>
<td>101</td>
<td>26</td>
<td>77</td>
</tr>
<tr>
<td>Back-off 5</td>
<td>7588</td>
<td>7588</td>
<td>167</td>
<td>54</td>
<td>97</td>
</tr>
<tr>
<td>Back-off 4</td>
<td>7086</td>
<td>7086</td>
<td>203</td>
<td>95</td>
<td>140</td>
</tr>
<tr>
<td>Back-off 3</td>
<td>6584</td>
<td>6584</td>
<td>239</td>
<td>107</td>
<td>170</td>
</tr>
<tr>
<td>Back-off 2</td>
<td>6082</td>
<td>6082</td>
<td>277</td>
<td>132</td>
<td>207</td>
</tr>
<tr>
<td>Back-off 1</td>
<td>5580</td>
<td>5580</td>
<td>313</td>
<td>155</td>
<td>230</td>
</tr>
<tr>
<td>Back-off 0</td>
<td>5088</td>
<td>5088</td>
<td>353</td>
<td>181</td>
<td>270</td>
</tr>
<tr>
<td>Sum</td>
<td>26402</td>
<td>26402</td>
<td>4452</td>
<td>1788</td>
<td>2768</td>
</tr>
</tbody>
</table>

Overlap Backoff

Future Work

- Use Google’s Web 1T 5-gram corpus data source instead of web counts.
- Explore backoff strategies based on linguistic generalizations.
- Vary the size of the context window based on linguistic information.
- Apply to learner language to identify and correct incorrect usage.

References


Click on the references to access the full papers.