Verb Behavior is not Verb Nature: Sense and Genre Biases as Sources of Subcategorization Probabilities

Douglas Roland, Daniel Jurafsky, and Laura Michaelis - Department of Linguistics - University of Colorado, Boulder

Introduction

Problem:
Verb subcategorization probability depends on the method of measurement (Merlo 1994, Gibson et al. 1996, Roland & Jurafsky 1997). Does this mean there is no fixed subcategorization probability for a given verb?

Solution:
- **Lemma Argument Probability Hypothesis**: Each lemma contains a vector of probabilistic expectations for its possible syntactic/semantic argument frames.
- **Probabilistic Combination Hypothesis**: Observed Subcategorization Probability = Lemma Argument Probability + Contextual Influence

Methodology

5 Corpora:
- Connine et al. (1984) (CFJC) single sentence production
- Garnsey et al. (1997) (Garnsey) single sentence completion
- Brown corpus (BC) Penn Treebank
- Wall Street Journal corpus (WSJ) Penn Treebank
- Switchboard corpus (SWBD) Penn Treebank

166 verbs coded for subcategorization:
- Complementation: FrameNet (Baker et al. 1999).
- 17 major categories: 0, PP, Vp, S, S0, Sw, Sfin, VPing, Vp[NS], VPNP, VP[PP], [VP NP], [VP Np], [VP S], [VP Sfin], Q, P, Passen, and Other
- Only true syntactic arguments, no adjuncts, following the distinction made in Treebank (Marcus et al. 1993).

7 verbs also hand-coded for Wordnet sense.

The following table shows the sample size for each corpus:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Token/Type</th>
<th>examples/verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFJC</td>
<td>14,000</td>
<td>(127 CFJC verbs)</td>
</tr>
<tr>
<td>Garnsey</td>
<td>3,200</td>
<td>(48 Garnsey verbs)</td>
</tr>
<tr>
<td>BC</td>
<td>21,000</td>
<td>(127 BC verbs)</td>
</tr>
<tr>
<td>WSJ</td>
<td>5,700</td>
<td>(127 WSJ verbs)</td>
</tr>
<tr>
<td>SWBD</td>
<td>10,000</td>
<td>(127 SWBD verbs)</td>
</tr>
</tbody>
</table>

How do you know if a verb is used the same way in two different corpora?

The cosine of the subcategorization probability vectors for the verb in each corpus can be used as a measure of the degree of difference (Salton & McGill 1983).

\[
\text{Cosine} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}
\]

Test tube sentences are different from wild sentences.

Word-sense Influences

Each verb sense has its own subcategorization probability.

- **Corpora have different distributions of verb sense.**
  - Sense of BC% WSI%
  - Example of the sense of charge,
  - attack 23% 0% His followers ... charged the trail, firing as they run. (BC)
  - run 8% 0% She charged off to the bedrooms. (BC)
  - appoint 6% 4% The commission is charged with designing a ... program. (WSI)
  - accuse 39% 58% Separately, a Compass shareholder filed suit, charging Compuers... (WSI)
  - bill 24% 36% Currently the government charges nothing for such filings. (WSI)
  - credit 0% 2% -
  - -
  - -
  - -
  - -

- **Verb senses have different subcategorization probabilities.**

- **Comparing between same discourse type?**
  - Written vs. Written
  - 5,200 examples
  - BC 571; BC 855; Garnsey 298; Garnsey 214; BC 80; Garnsey 63

- **Comparison between subcategorization vectors?**
  - Correlates with word type (Wordnet Sense)
  - Correlates with context (Contextual Influence)

Conclusion

- **Lemma Argument Probability Hypothesis**: Each lemma contains a vector of probabilistic expectations for its possible syntactic/semantic argument frames.
- **Probabilistic Combination Hypothesis**: Observed subcategorization probabilities are a combination of core probabilities and contextual influence.
- Psycholinguistic models and studies must take word sense and context into account.

Acknowledgments

- Thanks to Charles Clifton for providing the original data from Connine et al. (1984).
- Thanks to Sue Garnsey for providing the original data from Garnsey et al. (1997).
- This project was supported by the generosity of the NSF via NSF IIS-9733067, NSF IRI-9704046, NSF IRI-9618838, and the Committee on Research and Creative Work at the graduate school of the University of Colorado, Boulder. Many thanks to Giulia Bencani, Charles Billmore, Suzanne Gull, Adele Goldberg, Michelle Gregory, Uli Hoen, Paola Merlo, Noa Perlmutter, Bill Raymond, Philip Resnik, and to anonymous reviewers.