1 Background

Previous usage of predictability-based measures of sentence comprehension difficulty:

- Empirical observations that Cloze probability (Taylor, 1953) affects reading times (Ehrlich and Rayner, 1981) and event-related potentials (ERPs; Kutas and Hillyard, 1980, 1984)

  My brother came inside to... chat? eat? play? rest?
  The children went outside to... chat? eat? play? rest?

- Use of root mean-squared (RMS) word-prediction error to evaluate neural-net learning of natural language sentences (Elman, 1990, 1991; Christiansen and Chater, 1999; MacDonald and Christiansen, 2002; Rohde, 2003)

- Predictability is implicated in mathematical models of word reading, but usually on an absolute probability scale (Reichle et al., 1998; Engbert et al., 2005; see McDonald et al., 2005 for an exception)

2 Hale 2001

Proposals:

1. Probabilistic context-free grammars (PCFGs) are a good model of how human sentence comprehension works.

2. A probabilistic Earley parser is a good model of online eager sentence comprehension for PCFGs
3. The cognitive effort associated with a word in a sentence can be measured by the word’s negative log conditional probability:

\[ \log \frac{1}{P(w_i|w_{1..i-1})} \]

Results from this proposal:

1. Garden path sentences: *the horse raced past the barn fell*

![Graph showing log probability of words in a sentence.]

2. Subject/object RC ambiguities: one of the best-established processing asymmetries is the English subject/object RC asymmetry:

   The reporter who attacked the senator <\textit{easier}
   The reporter who the senator attacked

![Graph showing log probability for subject and object relative clauses.]

Caveat: these results are with a tiny, mostly hand-crafted grammar. Example using the entire Brown corpus:
3 Levy 2007

A different derivation of surprisal:

1. Let the probability distribution over complete structures $T$ (e.g., context-free trees) given a string prefix $w_1...i$ be denoted as $P_i(T)$.

2. The relative entropy, or Kullback-Leibler divergence (Cover and Thomas, 1991), $D(q||p)$ between two probability distributions is
   - a natural (though asymmetric) measure of probabilistic distance;
   - can be thought of as the penalty incurred for using the distribution $p$ to encode the finer-grained distribution $q$.

3. It turns out that the relative entropy between distributions before and after a word $w$ is equivalent to the surprisal of $w$:

   $$D(P_{k+1}||P_k) = \log \frac{1}{P_k(w_{k+1})}$$

4. If we think of probabilistic distance as the amount of work involved in reranking the candidate set $T$, then surprisal is also a measure of reranking work.

Results:

1. Constrained syntactic contexts.
   German verb-final clauses (Konieczny, 2000):

   (1)  a. Er hat den Abgeordneten begleitet, und . . .
       He has the delegate escorted, and . . .
       “He escorted the delegate, and . . .”
b. Er hat den Abgeordneten ans Rednerpult begleitet, und . . .
   He has the delegate to the lectern escorted, and . . .
   “He escorted the delegate to the lectern, and . . .”

   c. Er hat den Abgeordneten an das große Rednerpult begleitet, und . . .
   He has the delegate to the big lectern escorted, and . . .
   “He escorted the delegate to the large lectern, and . . .”

<table>
<thead>
<tr>
<th>Average RT (ms)</th>
<th>Surprisal</th>
<th>DLT prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>no PP</td>
<td>514</td>
<td>15.99 faster</td>
</tr>
<tr>
<td>short PP</td>
<td>477</td>
<td>15.41 slower</td>
</tr>
<tr>
<td>long PP</td>
<td>463</td>
<td>15.35 slower</td>
</tr>
</tbody>
</table>

2. Verb identity versus verb location (Jaeger et al., 2005):

   (2) a. The player [that the coach met at 8 o’clock] bought the house . . .
   b. The player [that the coach met by the river at 8 o’clock] bought the house . . .
   c. The player [that the coach met near the gym by the river at 8 o’clock] bought the house . . .

   Number of PPs intervening between embedded and matrix verb
   
<table>
<thead>
<tr>
<th>DLT prediction</th>
<th>1 PP</th>
<th>2 PPs</th>
<th>3 PPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surprisal</td>
<td>13.87</td>
<td>13.54</td>
<td>13.40</td>
</tr>
<tr>
<td>Mean Reading Time (ms)</td>
<td>510 ± 34</td>
<td>410 ± 21</td>
<td>394 ± 16</td>
</tr>
</tbody>
</table>

3. Facilitative ambiguity:

   (3) (Traxler et al., 1998)
   a. The daughter$_i$ of the colonel$_j$ who shot herself$_{i/*j}$ on the balcony had been very depressed.
   b. The daughter$_i$ of the colonel$_j$ who shot himself$_{i/j}$ on the balcony had been very depressed.
   c. The son$_i$ of the colonel$_j$ who shot himself$_{i/j}$ on the balcony had been very depressed.

   The ambiguous form can derive probability mass from both attachments; the unambiguous form can only derive mass from one attachment.

4 Other developments

- Other views of surprisal
Smith (2006) has shown that surprisal can be derived as a highly general optimization of a time/resource tradeoff, assuming only a scale-free property (that the cost of a unit $U$ can be derived as the sum of the costs of the subunits $u_1 \ldots n$ that make it up)

This works because joint events are characterized by products of probabilities, and the log of a product is the sum of logs

- Surprisal and sentence production:
  - With some extra (empirically testable) assumptions, surprisal can lead to the idea of uniform information density (UID)
  - Under UID, optimal communication involves smoothing out the surprisal profile of an utterance

References


