ADDRESSING SURPRISAL DEFICIENCIES IN READING TIME MODELS

Marten van Schijndel    William Schuler
December 11, 2016

Department of Linguistics, The Ohio State University
• Surprisal (PCFG, N-gram) is a way to estimate text complexity
• Surprisal (PCFG, N-gram) is a way to estimate text complexity
• Experienced complexity is reflected in reading speed
Overview

- Surprisal (PCFG, N-gram) is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

Claim:
Current surprisal models inadequately estimate reading complexity
Overview

- Surprisal (PCFG, N-gram) is a way to estimate text complexity
- Experienced complexity is reflected in reading speed

Claim:
Current surprisal models inadequately estimate reading complexity

This work:
A simple tweak to fix the surprisal measures
The red apple that the girl ate ...
The red apple that the \textbf{girl} ate ...

Reading model of ‘girl’:
sentence position
Reading complexity is estimated based on region ending

The red apple that the girl ate ...

Reading model of ‘girl’:
sentence position, word length
Reading complexity is estimated based on region ending

The red apple that the \textbf{girl} ate ...

Reading model of ‘girl’:
sentence position, word length, $P(\text{girl}|\text{the})$
The red apple that the \textbf{girl} ate ...

Reading model of ‘girl’:
sentence position, word length, $P(\text{girl} | \text{the})$
Reading complexity is estimated based on region ending

The red apple that the \textbf{girl} ate ...

Reading model of ‘girl’:
sentence position, word length, P(girl|the)
The red apple that the girl ate ...

The red canoe that the girl ate ...

Reading model of ‘girl’: sentence position, word length, \( P(\text{girl}|\text{the}) \)
This study: $n$-gram and PCFG surprisal
This study: $n$-gram and PCFG surprisal

\[
N\text{-gram-surp}(\text{girl}) = -\log P(\text{girl} | \text{the})
\]
This study: \( n \)-gram and PCFG surprisal

\[
\text{PCFG-surp}(\text{girl}) = -\log P(T_6 = \text{girl} \mid T_1 \ldots T_5 = \text{The} \ldots \text{the})
\]
Cumulative $N$-gram Surprisal

The red apple that the girl ate ...
The red apple that the girl ate ...

\[ \text{cumu-n-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log P(w_i \mid w_{i-n} \ldots w_{i-1}) \]
Accumulated surprisal fixes the theoretical problem

Cumulative $N$-gram Surprisal

\[ \text{cumu-n-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} - \log P(w_i \mid w_{i-n} \ldots w_{i-1}) \]

The red apple that the girl ate ...
Accumulated surprisal fixes the theoretical problem

Cumulative $N$-gram Surprisal

$$\text{cumu-}n\text{-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log P(w_i \mid w_{i-n} \ldots w_{i-1})$$

The red \textbf{apple} that the girl ate …
Cumulative $N$-gram Surprisal

\[
\text{cumu-}n\text{-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} \! -\log P(w_i \mid w_{i-n} \ldots w_{i-1})
\]
Accumulated surprisal fixes the theoretical problem

Cumulative PCFG Surprisal

\[
\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log P(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})
\]
Accumulated surprisal fixes the theoretical problem

Cumulative PCFG Surprisal

\[
\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log P(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})
\]
Accumulated surprisal fixes the theoretical problem

Cumulative PCFG Surprisal

\[ \text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log P(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1}) \]

- The red apple
- the girl

van Schijndel, Schuler

Fixing surprisal
Cumulative PCFG Surprisal

\[
\text{Cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}}^{f_t} -\log P(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})
\]
How well does this fix work?

N-gram surprisal

- 5-grams
- Trained on Gigaword 3.0 (Graff and Cieri, 2003)
- Computed with KenLM (Heafield et al., 2013)
How well does this fix work?

N-gram surprisal

- 5-grams
- Trained on Gigaword 3.0 (Graff and Cieri, 2003)
- Computed with KenLM (Heafield et al., 2013)

PCFG surprisal

- Trained on WSJ 02-21 (Marcus et al., 1993)
- Computed with van Schijndel et al., (2013) parser
University College London (UCL) Corpus (Frank et al., 2013)

- 43 subjects
- reading short sentences from online novels
- frequent comprehension questions
Baseline mixed effects model

Fixed Factors

- sentence position
- word length
- region length
- whether the previous word was fixated
How well does this fix work?

Baseline mixed effects model

Fixed Factors

- sentence position
- word length
- region length
- whether the previous word was fixated

Random Factors

- All fixed factors as by-subject random slopes
- Item, subject and subject×sentence intercepts
Accumulation improves \( n \)-gram surprisal
Accumulation improves $n$-gram surprisal
Accumulation improves n-gram surprisal

Captured reading time variance

Cumu-N-grams

N-grams

Baseline

10.61 ms
6.69 ms

Fixing surprisal
Accumulation may also help PCFG surprisal

After adding cumulative $n$-gram surprisal to model:

- PCFG surprisal is not useful ($p > 0.05$)
- Cumulative PCFG surprisal is not useful ($p > 0.05$)
- Cumulative PCFG is useful with richer grammar ($p < 0.001$)
After adding cumulative $n$-gram surprisal to model:

- PCFG surprisal is not useful ($p > 0.05$)
After adding cumulative $n$-gram surprisal to model:

- PCFG surprisal is not useful ($p > 0.05$)
- Cumulative PCFG surprisal is not useful ($p > 0.05$)
After adding cumulative n-gram surprisal to model:

- PCFG surprisal is not useful ($p > 0.05$)
- Cumulative PCFG surprisal is not useful ($p > 0.05$)
- Cumulative PCFG is useful with richer grammar ($p < 0.001$)
What does accumulation model?
Subsequent regression

1

The red apple that the girl ate ...
Subsequent regression

\[
^{1}\text{The red apple that the girl ate...}
\]
Subsequent regression

1 3
The red apple that the girl ate ...

2
Subsequent regression

The red apple that the girl ate ...
Subsequent regression

The red apple that the girl ate ...
Parafoveal processing

1

The red apple that the girl ate ...
Parafoveal processing

Th(e red apple that t)he girl ate ...
Parafoveal processing

\[ \text{The red apple that the girl ate ...} \]
Possible accumulation influences

Prediction (entropy)

1

The red apple that the girl ate ...
Possible accumulation influences

Prediction (entropy)

\[1\]

The red (apple that the girl) ate ...
Possible accumulation influences

Prediction (entropy)

The red (apple that the girl) ate ...
Accumulation alternative: Successor surprisal

Cumulative surprisal only handles subsequent regression
**Accumulation alternative: Successor surprisal**

Cumulative surprisal only handles subsequent regression

Parafoval: Th(e red apple that t)he girl ate …

Prediction: The red (apple that the girl) ate …

\[\text{accumulated}\]
Accumulation alternative: Successor surprisal

Cumulative surprisal only handles subsequent regression

Parafoveal: Th(e red apple that t)he girl ate ...

Prediction: The red (apple that the girl) ate ...

Other accumulation mechanisms presuppose earlier accumulation
Successor effects influence reading times

Upcoming material influences reading times
Successor effects influence reading times

Upcoming material influences reading times

- Orthographic effects
  (Pynte, Kennedy, & Ducrot, 2004; Angele, Tran, & Rayner, 2013)
Successor effects influence reading times

Upcoming material influences reading times

- Orthographic effects
  (Pynte, Kennedy, & Ducrot, 2004; Angele, Tran, & Rayner, 2013)

- Lexical effects
  (Kliegl et al., 2006; Li et al., 2014; Angele et al., 2015)
The red apple that the girl ate ...

\[ \text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(w_i \mid w_{i-n} \ldots w_{i-1}) \]
The red apple that the girl ate ...
The red \textbf{apple} that the girl ate ...

\begin{equation}
\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(w_i \mid w_{i-n} \ldots w_{i-1})
\end{equation}
**Successor N-grams**

The red apple that the girl ate ...

\[
\text{future-} n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(w_i \mid w_{i-n} \ldots w_{i-1})
\]
The red apple that the girl ate ...

\[
\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(w_i \mid w_{i-n} \ldots w_{i-1})
\]
Future-PCFG\( (w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} \log P(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1}) \)
Future-PCFG(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(T_i = w_i | T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})
Future-PCFG\((w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(T_i = w_i \mid T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})\)
Future-PCFG(w, f_t, f_{t+1}) = \sum_{i=f_t}^{f_{t+1}} -\log P(T_i = w_i | T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1})
Successor PCFG works

Future-PCFG

4.76 ms

Baseline

Captured reading time variance
Successor n-grams work better
Successor n-grams work better

PCFG surprisal may require a richer grammar
Successor $n$-grams have limited influence

Successor $n$-grams are most predictive for 2 future words ($p < 0.001$)

6% of UCL saccades ($n=3500$)
Successor $n$-grams have limited influence

Successor $n$-grams are most predictive for 2 future words ($p < 0.001$)
Successor $n$-grams have limited influence

Successor $n$-grams are most predictive for 2 future words ($p < 0.001$)

6% of UCL saccades ($n=3500$) > 2 words
Conclusions: Accumulate surprisal!

- $N$-gram surprisal should be accumulated to predict reading times

- Pre-saccade $n$-grams are limited

- PTB PCFG surprisal does not accumulate

- Richer grammars may accumulate better
N-gram surprisal should be accumulated to predict reading times

N-gram surprisal accumulates pre- and post-saccade
Conclusion: Accumulate surprisal!

- $N$-gram surprisal should be accumulated to predict reading times
- $N$-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade $n$-grams are limited
Conclusion: Accumulate surprisal!

- $N$-gram surprisal should be accumulated to predict reading times
- $N$-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade $n$-grams are limited
- PTB PCFG surprisal does not accumulate
Conclusion: Accumulate surprisal!

- \( N \)-gram surprisal should be accumulated to predict reading times
- \( N \)-gram surprisal accumulates pre- and post-saccade
  - Pre-saccade \( n \)-grams are limited
- PTB PCFG surprisal does not accumulate
- †Richer grammars may accumulate better
Thanks to:

- Stefan Frank
- National Science Foundation (DGE-1343012)
### UCL Effect Size Reference

<table>
<thead>
<tr>
<th>Model</th>
<th>Effect Size (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future $N$-grams</td>
<td>6.5*</td>
</tr>
<tr>
<td>$N$-grams</td>
<td>6.69</td>
</tr>
<tr>
<td>Cumulative GCG-PCFG†</td>
<td>8.25*</td>
</tr>
<tr>
<td>Cumulative $N$-grams</td>
<td>10.61*</td>
</tr>
</tbody>
</table>

*p < 0.001

*N*-gram model has the given effect size before adding cumulative *n*-grams.
**Cumu-N-gram Results**

<table>
<thead>
<tr>
<th>Model</th>
<th>$N$-gram vs Cumu-$N$-gram</th>
<th>Log-Likelihood</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>$-12702$</td>
<td>25476</td>
</tr>
<tr>
<td>Base+Basic</td>
<td>0.035</td>
<td>$-12689^*$</td>
<td>25451</td>
</tr>
<tr>
<td>Base+Cumulative</td>
<td>0.055</td>
<td>$-12683^*$</td>
<td>25440</td>
</tr>
<tr>
<td>Base+Both</td>
<td></td>
<td>$-12683^*$</td>
<td>25442</td>
</tr>
</tbody>
</table>

Base random: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram

Base fixed: sentpos, wlen, rlen, prevfix

Significance for the Base+Both model applies to improvement over the Base+Basic model.
## Future Surprisal Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Future-(N)-grams vs Future-PCFG</th>
<th>Log-Likelihood</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>(-12276)</td>
<td>24642</td>
</tr>
<tr>
<td>Base+Future-(N)-grams</td>
<td>0.034</td>
<td>(-12259^*)</td>
<td>24610</td>
</tr>
<tr>
<td>Base+Future-PCFG</td>
<td>0.025</td>
<td>(-12266^*)</td>
<td>24624</td>
</tr>
<tr>
<td>Base+Both</td>
<td></td>
<td>(-12259^*)</td>
<td>24612</td>
</tr>
</tbody>
</table>

Base random: sentpos, wlen, rlen, prevfix, cumu-5-gram, future-5-grams, future-PCFG

Base fixed: sentpos, wlen, rlen, prevfix, cumu-5-gram

Significance for the Base+Both model applies to improvement over the Base+Future-PCFG model.