Approximations of Predictive Entropy Correlate with Reading Times

Marten van Schijndel  William Schuler
July 29, 2017

Department of Linguistics, The Ohio State University
Angele et al. (2015)

A * child XXXXXXX the fish
Angele et al. (2015)

A | child | XXXXXXX | the | fish
A | child | annoyed | XXX | fish
Angele et al. (2015)

A child XXXXXXX the fish
A child annoyed XXX fish
A child annoyed the XXXX
Angele et al. (2015)

Lexical frequency of the upcoming masked word affects processing
Angele et al. (2015)

Lexical frequency of the upcoming masked word affects processing

Hypothesis: Effect is due to uncertainty over continuations
Angele et al. (2015)

Lexical frequency of the upcoming masked word affects processing

Hypothesis: Effect is due to uncertainty over continuations

Problem: Uncertainty is expensive to calculate
Entropy measures uncertainty

Shannon (1948)

\[ H(X) \overset{\text{def}}{=} - \sum_{x \in X} P(x) \log P(x) \]  

Roark et al. (2009) distinguishes two kinds of entropy (over words and preterminals)

\[ \text{Lex} H(w_1::i_1) \overset{\text{def}}{=} \sum w_{ij}^2 V P_G(w_{ij}w_1::i_1) \log P_G(w_{ij}w_1::i_1) \]  

\[ \text{Syn} H(w_1::i_1) \overset{\text{def}}{=} \sum p_{ij}^2 G P_G(p_{ij}w_1::i_1) \log P_G(p_{ij}w_1::i_1) \]
Entropy measures uncertainty

Shannon (1948)

\[
H(X) \overset{\text{def}}{=} - \sum_{x \in X} P(x) \log P(x)
\]  

Roark et al. (2009) distinguishes two kinds of entropy (over words and preterminals)

\[
\text{Lex} H(w_{1..i-1}) \overset{\text{def}}{=} - \sum_{w_i \in V} P_G(w_i \mid w_{1..i-1}) \log P_G(w_i \mid w_{1..i-1})
\]  

\[
\text{Syn} H(w_{1..i-1}) \overset{\text{def}}{=} - \sum_{p_i \in G} P_G(p_i \mid w_{1..i-1}) \log P_G(p_i \mid w_{1..i-1})
\]
Roark et al. (2009) showed

- Syn$H$ predicts self-paced reading times
- Lex$H$ is not predictive of SPR times
Roark et al. (2009) showed

- $\text{SynH}$ predicts self-paced reading times
- $\text{LexH}$ is not predictive of SPR times
  (No Angele et al., 2015, effect)
Roark et al. (2009) showed

- SynH predicts self-paced reading times
- LexH is not predictive of SPR times (No Angele et al., 2015, effect)

But

- Small training corpus (V is poor)
- Small test corpus:
  ~ 200 sentences, ~ 4000 words, 23 subjects
Natural Stories self-paced reading corpus (Futrell et al., in prep)

- 181 subjects
- 10 narrative texts
- 485 sentences (10256 words)
- Each text followed by 6 comprehension questions
- Events removed if $<100$ ms or $>3000$ ms

Parsed using Roark (2001) parser

Fitted with \textit{lmer}
A

---
- child -----------------
------ annoyed ------
Spaces were masked

--------------- the -----
Syntactic entropy predicts RTs

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\hat{\beta}$</th>
<th>$\hat{\sigma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic $H$</td>
<td>4.53*</td>
<td>0.54</td>
</tr>
<tr>
<td>Lexical $H$</td>
<td>-1.05</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Replication of Roark et al. (2009)
Syntactic entropy predicts RTs

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\hat{\beta}$</th>
<th>$\hat{\sigma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic $H$</td>
<td>4.53*</td>
<td>0.54</td>
</tr>
<tr>
<td>Lexical $H$</td>
<td>−1.05</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Replication of Roark et al. (2009)
But Angele et al. (2015) found a *lexical* frequency effect
Can we make $\text{Lex}H$ more tractable?

$S_G(w_i, w_{1..i-1}) \overset{\text{def}}{=} -\log P_G(w_i | w_{1..i-1})$ \hspace{1cm} (4)

$\text{Lex}H_G(w_{1..i-1}) \overset{\text{def}}{=} \sum_{w_i \in V} -P_G(w_i | w_{1..i-1}) \log P_G(w_i | w_{1..i-1})$ \hspace{1cm} (5)

$= \sum_{w_i \in V} P_G(w_i | w_{1..i-1}) S_G(w_i, w_{1..i-1})$ \hspace{1cm} (6)

$= \mathbb{E}[S_G(w_i, w_{1..i-1})]$ \hspace{1cm} (7)
Can we make Lex\( H \) more tractable?

\[
S_G(w_i, w_{1..i-1}) \overset{\text{def}}{=} -\log P_G(w_i \mid w_{1..i-1}) \quad (4)
\]

\[
\text{Lex}H_G(w_{1..i-1}) \overset{\text{def}}{=} \sum_{w_i \in V} -P_G(w_i \mid w_{1..i-1}) \log P_G(w_i \mid w_{1..i-1}) \quad (5)
\]

\[
= \sum_{w_i \in V} P_G(w_i \mid w_{1..i-1}) S_G(w_i, w_{1..i-1}) \quad (6)
\]

\[
= E[S_G(w_i, w_{1..i-1})] \quad (7)
\]

We can use a corpus instead of explicitly computing the expectation.
Entropy gives mean surprisal
Surprisal approximates entropy in the aggregate.
Surprisal approximates entropy in the aggregate.
Surprisal approximates entropy in the aggregate.
Surprisal approximates entropy in the aggregate.
Surprisal approximates entropy in the aggregate.
Ex: The boy annoyed the fish.
We can treat large corpora as our samplers.
We can try:

- Future Roark surprisal  
  (same distribution as SynH)
We can try:

- Future Roark surprisal (same distribution as SynH)
- Future 5-gram Surprisal (similar to what Angele et al., observed)
POSSIBLE ENTROPY APPROXIMATIONS

We can try:

- Future Roark surprisal
  (same distribution as SynH)
- Future 5-gram Surprisal
  (similar to what Angele et al., observed)
- Future categorial grammar surprisal
  (tests how specific syntactic prediction is)
### Uncertainty over both words and syntax

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\hat{\beta}$</th>
<th>$\hat{\sigma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic $H$</td>
<td>4.62*</td>
<td>0.53</td>
</tr>
<tr>
<td>Future Roark Surprisal</td>
<td>0.33</td>
<td>0.40</td>
</tr>
<tr>
<td>Future $N$-gram Surprisal</td>
<td>4.05*</td>
<td>0.58</td>
</tr>
<tr>
<td>Future Categorial Grammar Surprisal</td>
<td>4.10*</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Why does this pre-slowing occur?

- Better encoding of $w_i$ to help with $w_{i+1}$
Why does this pre-slowing occur?

- Better encoding of $w_i$ to help with $w_{i+1}$
- A kind of Uniform Information Density (UID; Jaeger, 2010)
  - Optimizes per-millisecond informativity
Conclusions

- Uncertainty about upcoming words slows processing.
- That influence can be detected prior to any expectation violation.
- Future surprisal can efficiently approximate that uncertainty.
- Syntactic uncertainty is fine-grained.
Conclusions

• Uncertainty about upcoming words slows processing
Conclusions

- Uncertainty about upcoming words slows processing
- That influence can be detected prior to any expectation violation
Conclusions

• Uncertainty about upcoming words slows processing
• That influence can be detected prior to any expectation violation
• Future surprisal can efficiently approximate that uncertainty
Conclusions

• Uncertainty about upcoming words slows processing
• That influence can be detected prior to any expectation violation
• Future surprisal can efficiently approximate that uncertainty
• Syntactic uncertainty is fine-grained
Thanks to:

- The reviewers for their very helpful comments
- National Science Foundation (DGE-1343012)