Approximations of Predictive Entropy Correlate with Reading Times

Marten van Schijndel William Schuler

July 29, 2017

Department of Linguistics, The Ohio State University

Upcoming material affects processing

Angele et al. (2015)

A ¦ ch*ild ¦ XXXXXXX ¦ the ¦ fish

Angele et al. (2015)

A	child*	XXXXXXX	the	fish
A	child	annoyed*	XXX	fish

Angele et al. (2015)

A	chi*ld	XXXXXXX	the	fish
A	child	anno*yed	XXX	fish
A	child	annoyed	th*e	XXXX

Upcoming material affects processing

Angele et al. (2015)

A	child*	XXXXXXX	the	fish
A	child	annoyed*	XXX	fish
A	child	annoyed	the*	XXXX

Lexical frequency of the upcoming masked word affects processing

Angele et al. (2015)

A	child*	XXXXXXX	the	fish
A	child	annoyed*	XXX	fish
A	child	annoyed	the*	XXXX

Lexical frequency of the upcoming masked word affects processing

Hypothesis: Effect is due to uncertainty over continuations

Angele et al. (2015)

A	$\overset{*}{\text{child}}$	XXXXXXX	the	fish
A	child	$\overset{*}{\text{annoyed}}$	XXX	fish
A	child	annoyed	$\overset{*}{\text{the}}$	XXXX

Lexical frequency of the upcoming masked word affects processing

Hypothesis: Effect is due to uncertainty over continuations

Problem: Uncertainty is expensive to calculate

Shannon (1948)

$$H(X) \stackrel{def}{=} -\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \tag{1}$$

Entropy measures uncertainty

Shannon (1948)

$$H(X) \stackrel{def}{=} -\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \tag{1}$$

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

$$\mathrm{Lex}H(w_{1..i-1}) \stackrel{def}{=} -\sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{2}$$

$$\mathrm{Syn}H(w_{1..i-1}) \stackrel{def}{=} -\sum_{p_i \in G} \mathrm{P}_G(p_i \mid w_{1..i-1}) \log \mathrm{P}_G(p_i \mid w_{1..i-1}) \tag{3}$$

Entropy measures uncertainty

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

- Syn$H$ predicts self-paced reading times
- Lex$H$ is not predictive of SPR times
  (No Angele et al., 2015, effect)

Roark et al. (2009) showed

- Syn$H$ predicts self-paced reading times
- Lex$H$ is not predictive of SPR times
  (No Angele et al., 2015, effect)

But

- Small training corpus (V is poor)
- Small test corpus:
  $\sim$ 200 sentences, $\sim$ 4000 words, 23 subjects

Test data in this work

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 *lmer*

Spaces were masked

_ _

Spaces were masked

```
A -----------------------
```

Spaces were masked

```
- child -----------------
```

Spaces were masked

------- annoyed ---------

Spaces were masked

`---------------- the -----`

Spaces were masked

```
-------------------- fish.
```

Syntactic entropy predicts RTs

Predictor	$\hat{\beta}$	$\hat{\sigma}$
Syntactic H	**4.53***	**0.54**
Lexical H	-1.05	0.41

Replication of Roark et al. (2009)

Syntactic entropy predicts RTs

Predictor	$\hat{\beta}$	$\hat{\sigma}$
Syntactic H	**4.53***	**0.54**
Lexical H	−1.05	0.41

Replication of Roark et al. (2009)

But Angele et al. (2015) found a *lexical* frequency effect

Can we make Lex$H$ more tractable?

$$S_G(w_i, w_{1..i-1}) \stackrel{def}{=} -\log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{4}$$

$$\mathrm{Lex}H_G(w_{1..i-1}) \stackrel{def}{=} \sum_{w_i \in V} -\mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{5}$$

$$= \sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1})\, S_G(w_i, w_{1..i-1}) \tag{6}$$

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

Can we make Lex$H$ more tractable?

$$S_G(w_i, w_{1..i-1}) \overset{def}{=} -\log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{4}$$

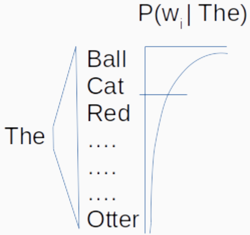
$$\mathrm{Lex}H_G(w_{1..i-1}) \overset{def}{=} \sum_{w_i \in V} -\mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{5}$$

$$= \sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1})\, S_G(w_i, w_{1..i-1}) \tag{6}$$

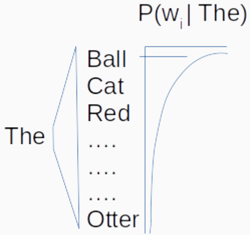
$$= \mathrm{E}[S_G(w_i, w_{1..i-1})] \tag{7}$$

We can use a corpus instead of explicitly computing the expectation

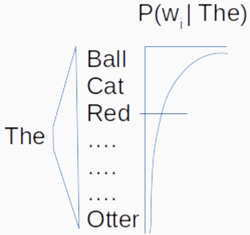
Entropy gives mean surprisal



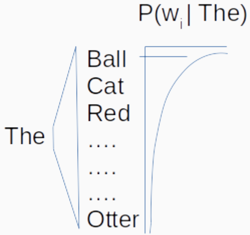
Surprisal approximates entropy in the aggregate



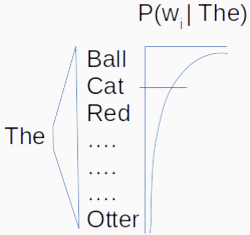
Surprisal approximates entropy in the aggregate



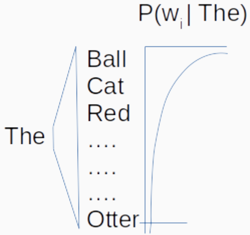
Surprisal approximates entropy in the aggregate



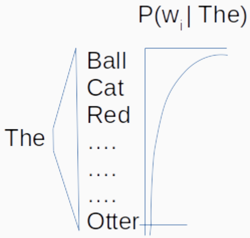
Surprisal approximates entropy in the aggregate



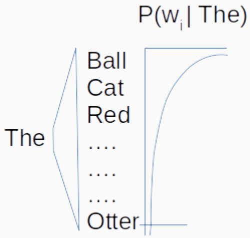
Surprisal approximates entropy in the aggregate



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)

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)

Predictor	$\hat{\beta}$	$\hat{\sigma}$
Syntactic *H*	**4.62***	**0.53**
Future Roark Surprisal	0.33	0.40
Future *N*-gram Surprisal	**4.05***	**0.58**
Future Categorial Grammar Surprisal	**4.10***	**0.74**

Why does this pre-slowing occur?

- Better encoding of w_i to help with w_{i+1}

- 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

Conclusions

- Uncertainty about upcoming words slows processing

- 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! Questions?

Thanks to:

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