Approximations of Predictive Entropy Correlate with Reading Times

Marten van Schijndel and William Schuler

van-schijndel.1@osu.edu

Department of Linguistics, The Ohio State University

Introduction

Uncertainty about upcoming words affects reading times [4]. But calculating the actual amount of uncertainty (entropy) over each word is expensive.

We use the surprisal of upcoming words to sample from entropy's conditional probability distribution, which is much easier to compute than entropy. In addition, this work shows how far in advance readers experience uncertainty.

Single-Step Predictive Entropy

Single-step predictive entropy reflects the amount of uncertainty over upcoming lexical observations w_i given a preceding lexical context, $w_{1..i-1}$, a grammar, G, and a vocabulary, V:

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

H^1 predicts reading times when computed over upcoming preterminal categories [4].
Entropy of lexical items is expensive to compute because it requires estimating probabilities for every word in V at every time step, and it is less effective than entropy over preterminals because of sparse data effects.

Surprisal is a Sample of Entropy

The surprisal of a word given its context reflects how unexpected the word was in context:

$$S_G(w_i, w_{1..i-1}) \stackrel{def}{=} -\log \mathrm{P}_G(w_i \mid w_{1..i-1}) \quad (2)$$

Entropy is just the expected value of surprisal [3]:

$$\begin{aligned} &H^1_G(w_{1..i-1}) \\ &\stackrel{def}{=} \sum_{w_i \in V} -\mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) && (3) \\ &= \sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1})\, S_G(w_i, w_{1..i-1}) && (4) \\ &= \mathrm{E}[S_G(w_i, w_{1..i-1})] && (5) \end{aligned}$$

Therefore, surprisal is a single sample from the conditional probability distribution over which H^1 is computed, where the sampled observation is the occurrence that ultimately will be observed. Over several trials, future surprisal should approximate entropy since each observed occurrence should happen proportionately to its expected occurrence frequency.

Self-Paced Reading Analyses

Data

Large self-paced reading corpus of linguistically difficult sentences, which read naturally [2]

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

Baseline Model

Fixed Effects

- Sentence position
- Word length
- Back-off 5-gram surprisal

Random Structure

- All fixed effects as by-subject slopes
- Word, subject, subject×sentence intercepts

Analysis 1: Roark Parser

Predictor	$\hat{\beta}$	$\hat{\sigma}$
Syntactic H^1	2.29*	0.61
Future Roark Surprisal	3.47*	0.50

- H^1 predicts reading times (Replicates [4])
- Future surprisal fits reading times

Analysis 2: Future N-grams

Predictor	$\hat{\beta}$	$\hat{\sigma}$
Future Roark Surprisal	1.25	0.59
Future 5-gram Surprisal	4.77*	0.64

- Future n-gram surprisal is a better predictor than future Roark surprisal
- Roark uses a coarse grammar
- Roark entropy predicts preterminals; not reflected in n-grams

Analysis 3: Fine-Grained Parser [5]

Predictor	$\hat{\beta}$	$\hat{\sigma}$
Syntactic H^1	2.99*	0.73
Future 5-gram Surprisal	5.11*	0.66
Future PCFG Surprisal	2.35*	0.64

- Future PCFG surprisal is predictive
- Syntactic entropy is still predictive, too
- Maybe PCFG surprisal is distorted by tail

Analysis 4: Mode Surprisal

Predictor	$\hat{\beta}$	$\hat{\sigma}$
Future PCFG Surprisal	7.63*	1.21
Future PCFG Surprisal Mode	−0.25	0.87

- Surprisal of most likely next event not useful
- So readers do not estimate uncertainty via mode choice

Eye-Tracking Analysis

Data

University College London Corpus [1]

- 43 subjects
- 361 narrative sentences
- Presentation order randomized
- 50% of sentences followed by a question

Model

Fixed Effects

- Sentence position
- Word length
- Length of preceding saccade
- Length of future saccade
- Cumulative back-off 5-gram surprisal [6]

Random Structure

- All fixed effects as by-subject slopes
- Word, subject, subject×sentence intercepts

Analysis: Future N-grams

First Pass Predictor	$\hat{\beta}$	$\hat{\sigma}$
Future Cumulative 5-gram Surprisal	4.72*	1.10
Future Cumulative PCFG Surprisal	0.73	1.35

- Future n-gram surprisal is predictive in ET
- Future cumulative PCFG surprisal is not
- To do: Try non-cumulative PCFG surprisal

Predictive Extent Analysis

Self-Paced Reading

- Future 5-gram surprisal is predictive for the next word
- Future effect is likely entirely predictive

Eye-Tracking

- Future 5-gram surprisal is predictive for the next 2 words
- Future effect may be partially parafoveal

Conclusion

- Readers are influenced by upcoming uncertainty
- Future surprisal can estimate that uncertainty
- Uncertainty may be driven by parafovea **and** prediction
- Uncertainty is not driven by distribution mode

[1] Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. The ERP response to the amount of information conveyed by words in sentences. *Brain & Language*, 140:1–11, 2015.

[2] Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. Natural stories corpus. in prep.

[3] Brian Roark. Expected surprisal and entropy. Technical Report CSLU-11-004, Center for Spoken Language Processing, Oregon Health and Science University, Portland, OR, 2011.

[4] Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Langauge Processing*, pages 324–333, 2009.

[5] Marten van Schijndel, Andy Exley, and William Schuler. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540, 2013.

[6] Marten van Schijndel and William Schuler. Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics, 2015.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1343012.

THE OHIO STATE UNIVERSITY