Hierarchic syntax improves reading time prediction

Contact:
vanschm@ling.osu.edu

Marten van Schijndel and William Schuler
Linguistics, The Ohio State University

The Ohio State University

Introduction

Previous work has debated whether humans make use of hierarchic syntax when processing language [Frank and Bod, 2011]. The present work demonstrates:

- How to improve strong 5-gram language models,
- Hierarchic syntax improves reading time fit over a strong linear baseline,
- Hierarchic syntax is used during reading to resolve both local and long-distance structural dependencies.

Modeled Variables

Two reading time measures are computed:

The ^{1}red apple that the ^{2}girl ate ...

Given the fixation sequence: *red*, *girl*
Time from initial fixation of *girl* until:

- **First Pass**: first fixation before *red* or after *girl*.
- **Go-Past**: first fixation after *girl*.

The sequence from *red* to *girl* is called the *region*

Predictors evaluated against both reading time measures.
Results are similar for both measures.

Predictors

The following predictors are tested:

Factors	Duration Predictions R^{w4}_{w4}	R^{w6}_{w5}
n-gram	$\mathrm{P}(w_4\|w_3, w_2)$	$\mathrm{P}(w_6\|w_5, w_4)$
cumu-n-gram	$\mathrm{P}(w_4\|w_3, w_2)$	$\mathrm{P}(w_6\|w_5, w_4)\cdot\mathrm{P}(w_5\|w_4, w_3)$
surp	$-\log \mathrm{P}(w_4\|T_3)$	$-\log \mathrm{P}(w_6\|T_5)$
cumusurp	$-\log \mathrm{P}(w_4\|T_3)$	$-\log [\mathrm{P}(w_6\|T_5)\cdot\mathrm{P}(w_5\|T_4)]$

w_i: word i
R^{wj}_{wi}: region from w_i to w_j (inclusive)
T_i: set of syntactic structures that can span from w_1 to w_i

Software and Data

PCFG surprisal values were obtained using the van Schijndel et al., (2013) parser, which was trained on the WSJ corpus. *N*-gram probabilities were computed using KenLM over the 2.96 billion word Gigaword 4.0 corpus. Mixed models were fit using lme4 (1.1-7). Experiments were conducted over the Dundee corpus after filtering the first and last word of each sentence/line and all regions with more than 4 words.

Acknowledgements

Thanks to Stefan Frank for feedback and engaging discussion related to this work. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

N-gram Example

Bigram probabilities predict reading time of *girl* after *red*:

The ^{1}red apple that the ^{2}girl ate ...

$^{\#}$X: fixations X: bigram target X: bigram condition

Traditional n-gram measures fail to capture entire sequence.
Conditions are never generated;
Probability of given sequence is deficient.

Cumulative *N*-gram Example

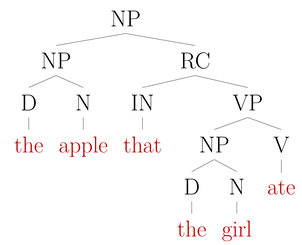
Cumu-bigram probs predict reading time of *girl* after *red*:

The ^{1}red apple that the ^{2}girl ate ...

$^{\#}$X: fixations X: bigram targets X: bigram conditions

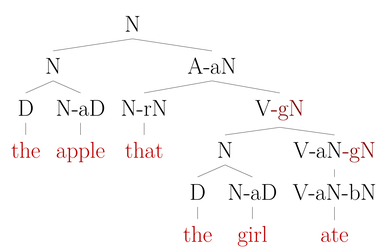
Cumulative n-gram product captures entire sequence.
Probability of given sequence is well-formed.
Reflects processing that must be done by humans.

PTB Example



Penn Treebank (PTB) grammar
Sensitive to local structure

GCG Example



Nguyen et al. (2012) generalized categorial grammar (GCG)
Sensitive to long-distance dependencies
(gap -g propagates from filler to gap)

Experiments

Experiments used linear mixed effects models with by-item and by-subject random intercepts and by-subject random slopes.
The significance of model fit differences was determined using χ^2 tests (First Pass $n = 194882$; Go-Past $n = 193709$).
All effects go in expected, usual directions (e.g., high cumu/n-grams → faster reading, high surprisal → slower reading).

1) Cumulative *N*-grams

Can n-grams reflect more complete probabilities?

- Base factors:
  - Fixed: Sentence position
  - Fixed: Word length
  - Fixed: Region length (in words)
  - Fixed: Was preceding word fixated?
  - Random: All fixed effects
  - Random: 5-gram
  - Random: Cumu-5-gram

First Pass Evaluation (AIC):

Base 2424868	
Base+*N*-gram 2424864 ($p < 0.05$)	**Base+Cumu-*n*-gram** 2424856 ($p < 0.01$)
Base+Both 2424848 ($p < 0.01$)	**Base+Both** 2424848 ($p < 0.01$)

2) Cumulative Surprisal

Can PCFG surprisal reflect more complete probabilities?

- Base contains factors from Experiment 1, plus:
  - Fixed: 5-gram
  - Fixed: Cumu-5-gram
  - Random: Surprisal (PTB PCFG)
  - Random: Cumusurp (PTB PCFG)

First Pass Evaluation (AIC):

Base 2424627	
Base+Surp 2424617 ($p < 0.01$)	Base+Cumusurp 2424627
Base+Both 2424619	**Base+Both** 2424619 ($p < 0.01$)

Results are comparable when using GCG PCFG

3) Hierarchic Syntax

Does hierarchic syntax improve over a strong linear baseline?

- Base contains factors from Experiment 1, plus:
  - Fixed: 5-gram
  - Fixed: Cumu-5-gram
  - Random: Surprisal (PTB PCFG)
  - Random: Surprisal (GCG PCFG)

First Pass Evaluation (AIC):

Base 2424592	
Base+PTB 2424587 ($p < 0.01$)	**Base+GCG** 2424589 ($p < 0.05$)
Base+Both 2424583 ($p < 0.05$)	**Base+Both** 2424583 ($p < 0.01$)

Go-Past Evaluation (AIC):

Base 2523055	
Base+PTB 2523047 ($p < 0.01$)	**Base+GCG** 2523050 ($p < 0.01$)
Base+Both 2523043 ($p < 0.01$)	**Base+Both** 2523043 ($p < 0.01$)

Results and Discussion

Results

- *N*-grams predict reading times locally **and** cumulatively.
- Cumulative surprisal does not improve reading time fit.
- PCFG surprisal predicts reading times over n-grams.
- Local surprisal predicts times over non-local surprisal.
- Non-local surprisal predicts times over local surprisal.

Conclusion

- Hierarchic structure affects reading times
- Long distance dependencies independently affect reading times
- Studies should compute n-grams for entire processed sequence

References

[Frank and Bod, 2011] Frank, S. and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*.

[Nguyen et al., 2012] Nguyen, L., van Schijndel, M., and Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*, pages 2125–2140, Mumbai, India.

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