Contributions of Propositional Content and Syntactic Categories in Sentence Processing

Byung-Doh Oh William Schuler

The Ohio State University

March 4, 2021
34th Annual CUNY Conference

Background

Expectation-based theories of sentence processing

- Processing difficulty is determined by predictability in context
- Can be quantified via *surprisal* (Shannon, 1948)

This work: A left-corner parser that incorporates both information about *propositional content* and *syntactic category labels* in generating surprisal estimates

Why propositional content?

- Comprehension entails building a coherent mental representation of propositional content (Kintsch, 1988)
- Propositional content rather than surface form stored during processing (Bransford & Franks, 1971; Jarvella, 1971)
- Parsing decisions are informed by semantic interpretation (Brown-Schmidt et al., 2002; Tanenhaus et al., 1995)

Left-corner Parser with Propositional Content

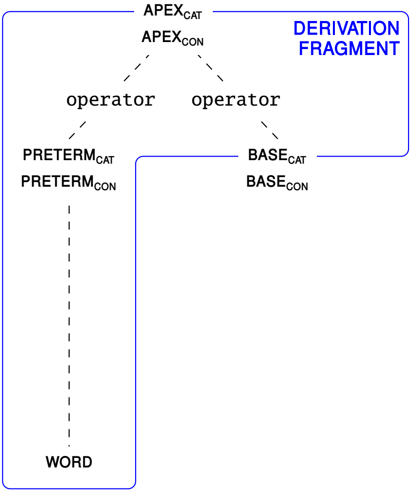
Each node in the parse tree has a *predicate context vector*

(Levy & Goldberg, 2014)

- Each element has the form of $predicate_{role}$, representing argument structure (e.g. $pour_2$)
- Argument structure derived from generalized categorial grammar reannotation (Bach, 1981; Nguyen et al., 2012)

The left-corner parser generates a predicate context vector for each word and propagates it along the parse tree

Full Model Overview



APEX_{CAT}
APEX_{CON}
DERIVATION FRAGMENT
operator
operator
$\text{PRETERM}_{\text{CAT}}$
$\text{PRETERM}_{\text{CON}}$
BASE_{CAT}
BASE_{CON}
WORD

Lexical phase

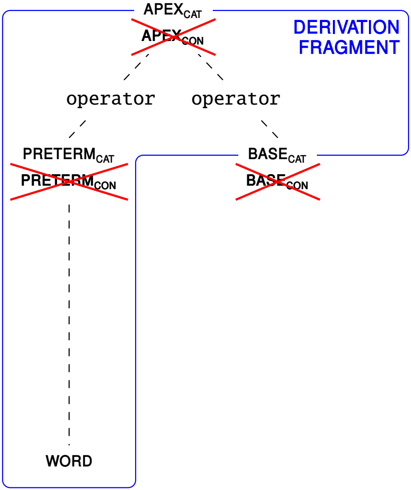
- Attach?
- Preterminal?
- Word?

Grammatical phase

- Attach?
- Operators?
- Apex?
- Base?

- Parsing decisions condition on both CONTENT and CATEGORY information
- For surprisal estimation, beam search is utilized to calculate prefix probabilities of a given word sequence (*FullSurp*)

Ablated Model 1: Content-ablated Model



Lexical phase

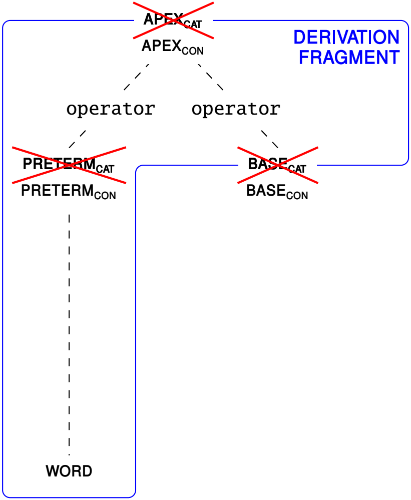
- Attach?
- Preterminal?
- Word?

Grammatical phase

- Attach?
- Operators?
- Apex?
- Base?

- Parsing decisions do not condition on CONTENT information
- Used to generate *NoConSurp* estimates

Ablated Model 2: Category-ablated Model



~~APEX$_{\text{CAT}}$~~

APEX$_{\text{CON}}$

DERIVATION FRAGMENT

operator operator

~~PRETERM$_{\text{CAT}}$~~

PRETERM$_{\text{CON}}$

~~BASE$_{\text{CAT}}$~~

BASE$_{\text{CON}}$

WORD

Lexical phase

- Attach?
- Preterminal?
- Word?

Grammatical phase

- Attach?
- Operators?
- Apex?
- Base?

- Parsing decisions do not condition on CATEGORY information
- Used to generate *NoCatSurp* estimates

Training Setup

Full, content-ablated, category-ablated models trained on WSJ02-21 (Marcus et al., 1993)

- 39,832 sentences
- 950,028 words
- Reannotated to generalized categorial grammar (Nguyen et al., 2012)
- Each variant trained with three random seeds for initialization

FullSurp, *NoConSurp*, and *NoCatSurp* estimated using beam search

Psycholinguistic Evaluation

Does propositional content or syntactic category information contribute to predicting human behavioral responses?

Evaluation on Natural Stories Corpus (Futrell et al., 2018)

- Self-paced reading times from 181 participants
- 485 sentences
- 10,245 words

Series of likelihood ratio tests based on linear mixed-effects models

- Full LME model: *NoConSurp* or *NoCatSurp* + *FullSurp*
- Base LME model: *NoConSurp* or *NoCatSurp* only

LRT Results

NoConSurp only vs. *NoConSurp* + *FullSurp*

	FullSurp		
NoConSurp	1	2	3
1	*ConvFail*	0.035^{*}	0.018^{*}
2	0.004^{**}	*ConvFail*	0.047^{*}
3	0.003^{**}	0.058	0.036^{*}

NoCatSurp only vs. *NoCatSurp* + *FullSurp*

	FullSurp		
NoCatSurp	1	2	3
1	*ConvFail*	$<0.001^{***}$	*ConvFail*
2	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
3	*ConvFail*	$<0.001^{***}$	$<0.001^{***}$

- Suggests a differential role of propositional content and syntactic category information in broad-coverage sentence processing
- Future work could aim to localize the influence of these information

Thank you for listening!

Source code:
`https://github.com/modelblocks/modelblocks-release`

References I

Bach, E. (1981). Discontinuous constituents in generalized categorial grammars. *Proceedings of the Annual Meeting of the Northeast Linguistic Society (NELS)*, *11*, 1–12.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, *2*, 331–350.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2002). Reference resolution in the wild: Online circumscription of referential domains in a natural interactive problem-solving task. *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, 148–153.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2018). The Natural Stories Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 76–82.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, *10*, 409–416.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 302–308.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

References II

Nguyen, L., van Schijndel, M., & Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. *Proceedings of the 24th International Conference on Computational Linguistics*, 2125–2140.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.