IMPROVED PRUNING AND SUPERTAGGING VIA SHIFT-REDUCE CCG PARSING

Stephen Merity, James R. Curran

ə-lab, School of Information Technologies
smerity@it.usyd.edu.au

SUMMARY

We implement the shift-reduce algorithm for the C&C statistical natural language parser. By using new features that the shift-reduce algorithm provides, supertagging accuracy is improved by 0.7%. By performing frontier pruning with these new features, parsing speed was increased by 34%.

MOTIVATION

- NLP parsers can extract the underlying structure and meaning of raw text and are vital for deep knowledge discovery and understanding.
- Improving supertagging accuracy results in higher parsing speed and higher accuracy.
- Current parsers are too slow to process web-scale resources without sacrificing accuracy.

BACKGROUND

- Combinatory Categorial Grammar (CCG; Steedman, 2000) is a lexicalised grammar formalism based on combinatory logic.
- Each word is assigned a category that encodes how the word behaves in the sentence.
- Supertagging improves parsing speed substantially by removing potential CCG categories for a word depending on the word's context.
  - e.g. the word *saw* is unlikely to be a verb if the word *the* appears immediately before it.

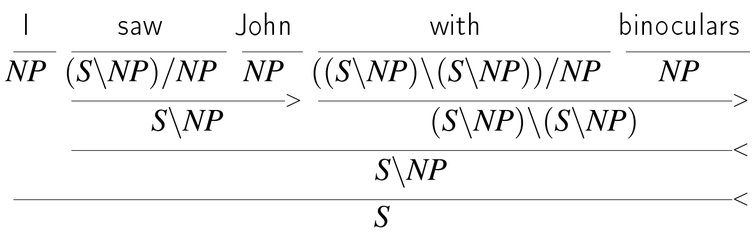

Figure 1: *Two* CCG *derivations with PP ambiguity.*

THE C&C PARSER

- The C&C CCG parser is a state-of-the-art parser trained over CCGbank, a corpus of 40,000 annotated sentences [Clark and Curran, 2007].
- State-of-the-art parsing speed due to highly efficient $C++$ codebase and CCG supertagging.

SHIFT-REDUCE PARSING

- Shift-reduce parsing allows for more accurate supertagging by incrementally parsing the sentence just like a human.
- Unfortunately shift-reduce parsing is exponential if backtracking is allowed. This is necessary due to ambiguity, such as in Figure 1.
- To ensure polynomial time worst-case complexity, a graph-structured stack (GSS; Tomita, 1988; Huang and Sagae, 2010) is employed.
  - This is the first time a graph-structured stack has been implemented for a shift-reduce CCG parser.

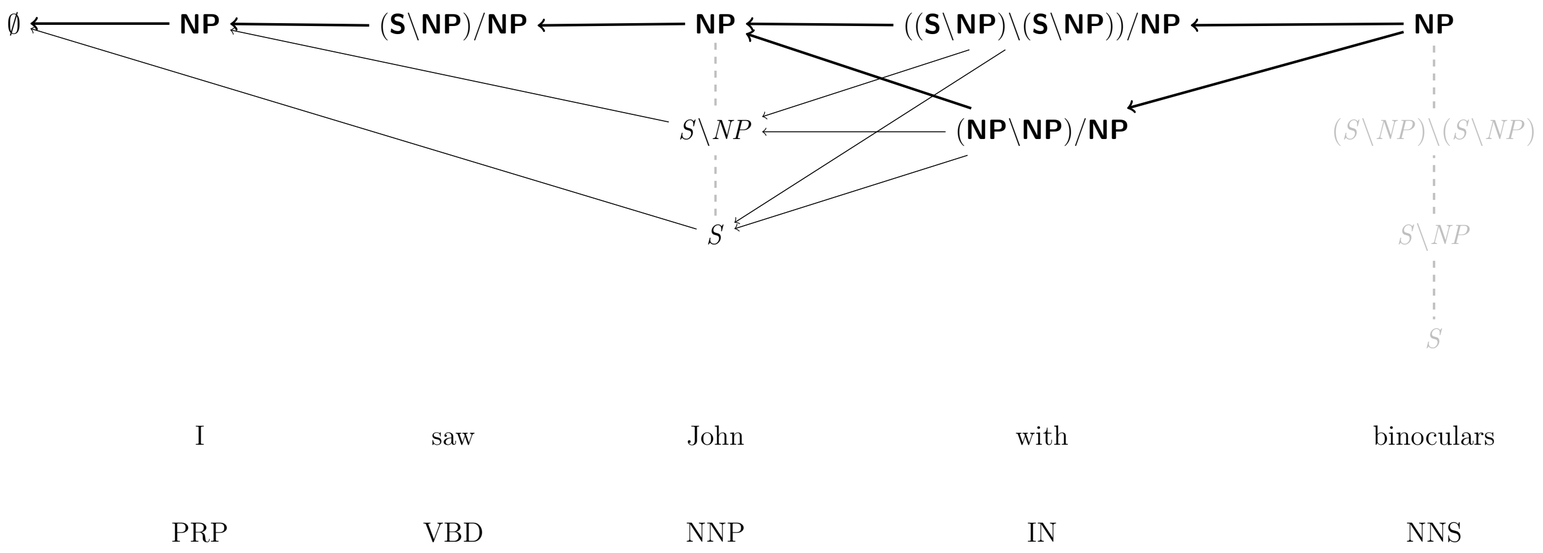

Figure 2: *A graph-structured stack representing an incomplete parse of sentences in Figure 1.*

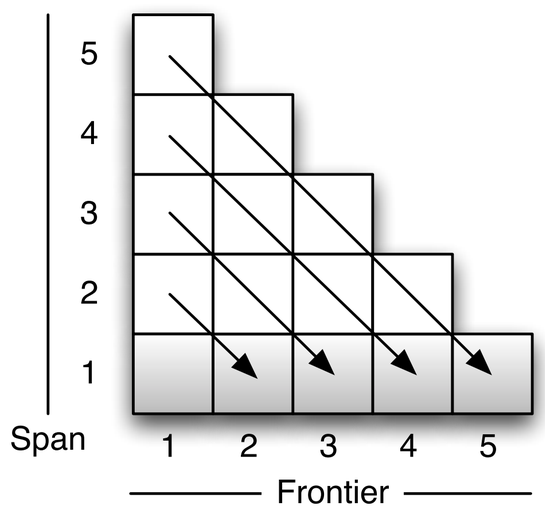

Figure 3: *An illustration of the relation between the chart in* CKY *and the graph-structured stack in* SR*.*

IMPROVED SUPERTAGGING

- Better supertagging accuracy means lower parsing ambiguity and higher parsing speeds.
- By providing better context about the current parse state, better supertags are selected. This is only now possible due to the incremental nature of the shift-reduce algorithm.

Model	Accuracy (words)	Accuracy (sents)
Baseline	93.77	42.19
Frontier Features	**94.47**	**42.23**

Table 1: *Supertagging accuracy on Section 23 of CCGbank, calculated across both individual words and entire sentences.*

FRONTIER PRUNING

- Reduces the search space the parser needs to consider by removing partial derivations that are unlikely to be in the highest-scoring derivation.
  - Increases parsing speed by not having to consider partial derivations that won't be used.
  - Prevents pruning useful partial derivations by using the same features as the base parser.
- Frontier pruning increases the shift-reduce parser speed to be competitive with the C&C parser with a small accuracy penalty.

Model	Coverage (%)	F-score (%)	Speed (sents/sec)
CKY C&C	99.34	**86.79**	**96.3**
SR C&C	**99.58**	86.78	71.3
Frontier Pruning	99.38	86.51	95.4

Table 2: *Parsing accuracy and speed on Section 23 of CCGbank.*

INDUSTRY APPLICATIONS

- Incremental shift-reduce parsing enables higher accuracy and lower response times in both *speech recognition* and *predictive text editing*.
- Improved parsing speed and supertagging accuracy allow for improved understanding of text resources such as the *Sydney Morning Herald*.
- Higher parsing speeds can allow for high accuracy processing of resources traditionally considered too large for practical use.

FUTURE WORK

- Perform part-of-speech tagging using the frontier features for improved accuracy.
- Fully integrating the new supertagger features into the parser for improved speed & accuracy.
- Implement low-level optimisation tweaks on the shift-reduce C&C parser to make it competitive with the highly optimised CKY C&C parser.

CONCLUSION

This is the first time a graph-structured stack has been implemented in a shift-reduce CCG parser. We have shown that by using the new features that the shift-reduce algorithm provides, supertagging accuracy can be significantly improved. These features also allow for a novel method of pruning that improves baseline parsing speed to similar levels as the highly optimised CKY C&C parser.

Acknowledgements

This work was supported by Australian Research Council Discovery grants DP1097291 and a University of Sydney Merit Scholarship.

References

Stephen Clark and James R. Curran. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. *Computational Linguistics*, 33:493–552, 2007.

Liang Huang and Kenji Sagae. Dynamic Programming for Linear-Time Incremental Parsing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1077–1086, 2010.

Mark Steedman. *The Syntactic Process*. MIT Press, Cambridge, Massachusetts, 2000.

Masaru Tomita. Graph-structured Stack and Natural Language Parsing. *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics*, pages 249–257, 1988.