ACCURATE PARSING USING RERANKING

Dominick Ng, Matthew Honnibal, and James R. Curran

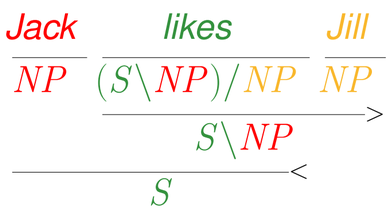
ə-lab, School of Information Technologies
dominick.ng@sydney.edu.au

SUMMARY

- **We implement a discriminative parser reranker for the C&C statistical natural language parser, which improves parsing accuracy significantly**.
- Reranking returns the best of an n-best set of parses; the process is more flexible than parsing and can consider a wide variety of features that model topology, context, and linguistic fidelity.

BACKGROUND

- Combinatory Categorial Grammar (CCG) is a lexicalised grammar formalism based on combinatory logic [Steedman, 2000].
- Each word is assigned a category that describes how it behaves in the sentence.
- Transitive verbs such as *likes* require an object NP to the right and a subject NP to the left to produce a valid sentence.



MOTIVATION

- Humans cannot possibly keep up with the amount of data generated every day. Automated analysis of information is vital.
- *Parsers* take raw text and produce representations of the underlying structure.
- The most accurate parsers only reach 85-90% accuracy on unseen sentences. We need more accurate systems in order to improve automated language analysis.
- The C&C CCG parser is a state-of-the-art system trained and tested over CCGbank, a corpus of 40,000 annotated sentences [Clark and Curran, 2007].

RERANKING

- *Reranking* is the process of taking a ranked list of parses and reordering them according to an external model.
  - features in the parser are constrained in nature by dynamic programming
  - rerankers are less constrained and can consider a wider set of features
- Charniak and Johnson [2005] showed that reranking improves parsing accuracy for a PCFG Penn Treebank-based parser.
- We adapt the features used by Charniak and Johnson, and develop novel features based on the expressive CCG formalism.
- We experiment with various combinations of features and parameters for reranking
  - values of n (10 and 50)
  - regression and classification
  - feature pruning cutoffs

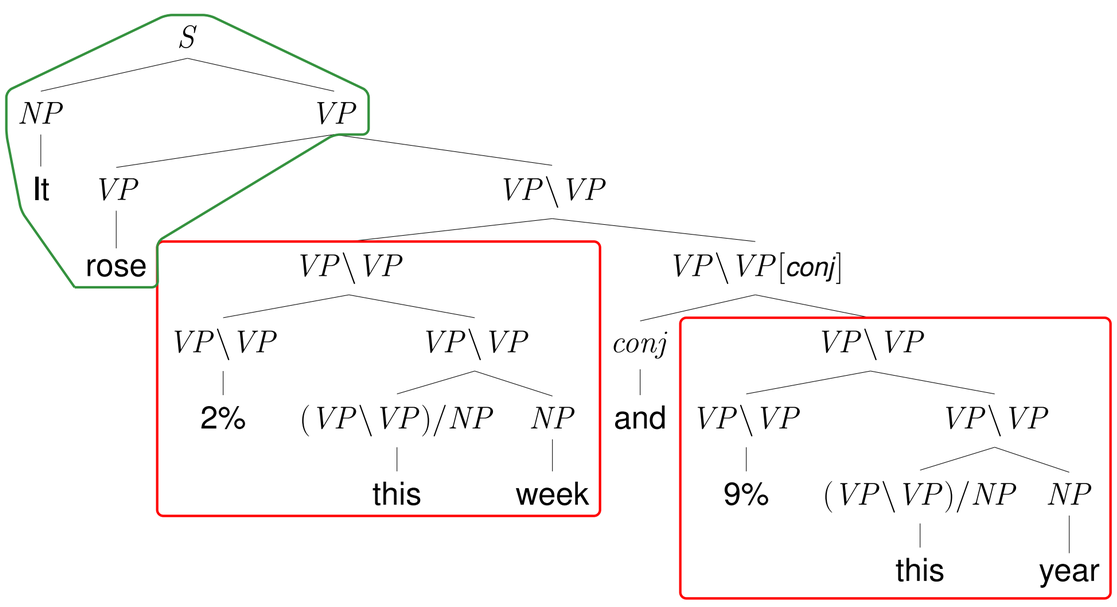


Figure 2: The correct derivation.

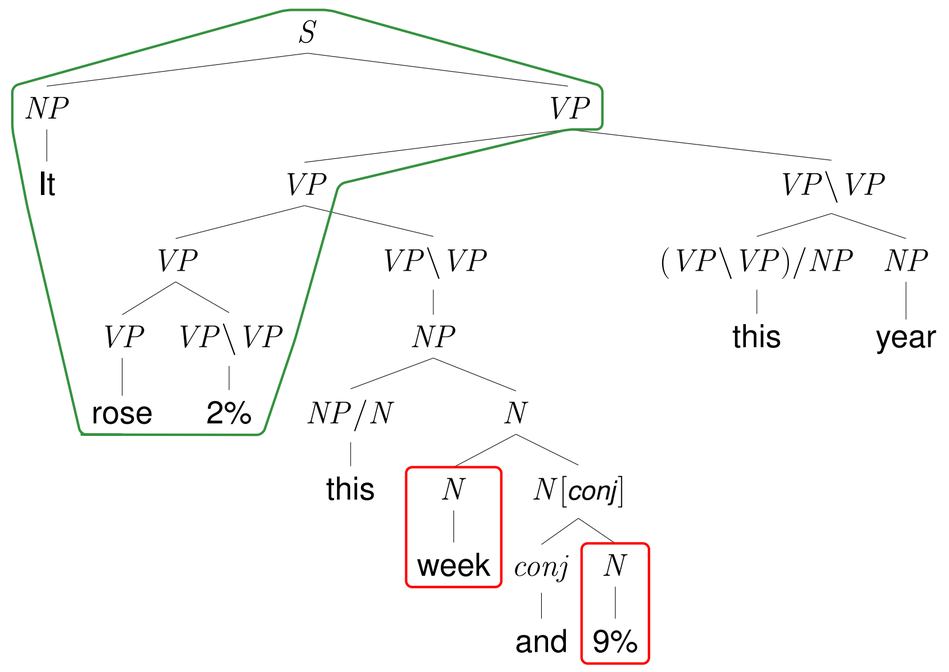


Figure 3: An incorrect derivation featuring a conjunction error.

Figure 4: Two CCG derivations for, *It rose 2% this week and 9% this year*. Highlighted differences are examples of reranking features.

FEATURES

Tree Topology - describes the overall shape of the parse tree and attempts to capture general conventions of English, e.g.
- preference for right-branching trees
- parallelism in conjunctions

Local Context - adds context that is difficult to capture in the parser model, e.g.
- ancestor paths of nodes in the tree
- words at the edges of larger constituents

CCG - features that may indicate an overly complicated or undesirable derivation based on the grammar, e.g.
- combinations of unary rule applications
- heads of conjunctive structures

C&C - the features used in the parser model, to give the reranker the same evidence used by the parser itself
- word-category pairs
- rule applications and dependencies

RESULTS

- All experiments improved over choosing a parse at random from the n-best list.
- Regression over 10-best parses without pruning achieved a significant improvement in parsing accuracy.
- More features generally improved performance, and the novel features that we developed performed best in isolation.
  - Charniak and Johnson's features alone did not improve performance for the more expressive CCG formalism.

	LP	LR	LF	AF
Baseline	87.19	86.32	86.75	84.80
Random	85.40	84.46	84.93	83.00
Class+CJ	87.21	86.06	86.63	84.81
Class+CCG	87.17	86.18	86.67	84.75
Class+ALL	87.32	86.32	86.82	84.85
Regress+CJ	86.96	85.99	86.47	84.58
Regress+CCG	87.27	86.41	86.83	85.08
Regress+ALL	**87.60**	**86.67**	**87.13**	**85.22**

Table 1: *10-best reranking performance on Section 00 of CCGbank*

	LP	LR	LF	AF
Baseline	87.19	86.32	86.75	84.80
Random	83.90	82.58	83.24	81.50
Class+CJ	86.93	85.87	86.40	84.69
Class+CCG	87.17	86.10	86.63	84.62
Class+ALL	87.38	86.29	86.83	84.91
Regress+CJ	86.49	85.64	86.07	84.22
Regress+CCG	87.08	86.15	86.61	84.65
Regress+ALL	**87.28**	**86.30**	**86.79**	**84.89**

Table 2: *50-best reranking performance on Section 00 of CCGbank*

FUTURE WORK

- Fully integrating the reranker into the parser for improved speed.
- Drawing features from wider corpora, such as the North American News Corpus and Wikipedia.

CONCLUSION

We have shown that reranking with expressive formalisms improves parsing accuracy. Our experiments also demonstrate that a wide variety of features is necessary for reranking to positively impact results.

Acknowledgements

This work was supported by Australian Research Council Discovery grants DP0665973 and DP1097291, the Capital Markets Cooperative Research Centre, and a University of Sydney Merit Scholarship.

References

Eugene Charniak and Mark Johnson. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of ACL-05*, pages 173–180, Ann Arbor, Michigan, USA, June 2005.

Stephen Clark and James R. Curran. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. *Computational Linguistics*, 33(4):493–552, 2007.

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