BACKGROUND

- Supervised parsers analyse grammatical structure. Accurate parsing is a vital step for most natural language processing applications.
- The defacto standard parser training corpus is based on 1 million words of 1989 newswire text that is outdated and domain-specific.
- We exploit very large corpora of unannotated and automatically parsed text for features to address this deficiency and incorporate more knowledge into the parser.

CORPORA

- Web1T (web text) and Google Books (scanned books) provide surface n-gram counts
  - Counts of 1 to 5 adjacent words
- Google Syntactic Ngrams provides dependency structure counts over parsed Google Books
  - Counts of Stanford dependency subtrees
  - Ranged with a "more accurate" baseline parser
- Bansal and Klein (2011) developed surface n-gram count features from Web1T, and improved parsing accuracy by 0.6%

RESULTS

Table 1 summarises unlabeled accuracy scores for the baseline, surface n-gram features (WEBT, BOOK), syntactic n-gram features (SYNT), and combined WEVT and SYNT.

- Google Books is comparable to Web1T for surface n-grams (despite being half the size), with accuracies being statistically indistinguishable between the two.
- Surface and syntactic n-gram features produce similar accuracy improvements on newswire (0.7%) and web text (1.0%)
- Combining the two feature sets in a single model yields up to 1.0% improvement on newswire and 1.4% on web text.

- The Syntactic Ngrams corpus is very noisy — syntactic n-gram features perform best with a minimum frequency cutoff of 10,000
- Worse performance at lower frequencies suggest parser errors are being masked by the large volume of text
- Figure 4 shows that syntactic n-grams perform best on verb phrases and conjunctions, but do worse on noun and prepositional phrases — known challenges for parsers.

CONCLUSION

- Combined surface and syntactic n-gram features outperform either in isolation.
- We achieve up to 1.4% improved accuracy across LTH and Stanford dependencies, and on newswire and web text.

Acknowledgements

Supported by Australian Research Council Discovery grant DP1097291, an Australian Postgraduate Award, a University of Sydney Vice-Chancellor’s Research Scholarship, a Google Australia PhD Fellowship, and a Fulbright Scholarship.

References