Online Statistics for a Unification-Based Dialogue Parser

Micha Elsner, Mary Swift, James Allen, and Daniel Gildea
Department of Computer Science
University of Rochester
Rochester, NY 14627
{melsner,swift,allen,gildea}@cs.rochester.edu

Abstract

We describe a method for augmenting unification-based deep parsing with statistical methods. We extend and adapt the Bikel parser, which uses head-driven lexical statistics, to dialogue. We show that our augmented parser produces significantly fewer constituents than the baseline system and achieves comparable bracketing accuracy, even yielding slight improvements for longer sentences.

1 Introduction

Unification parsers have problems with efficiency and selecting the best parse. Lexically-conditioned statistics as used by Collins (1999) may provide a solution. They have been used in three ways: as a postprocess for parse selection (Toutanova et al., 2005; Riezler et al., 2000; Riezler et al., 2002), a preprocess to find more probable bracketing structures (Swift et al., 2004), and online to rank each constituent produced, as in Tsuruoka et al. (2004) and this experiment.

The TRIPS parser (Allen et al., 1996) is a unification parser using an HPSG-inspired grammar and hand-tuned weights for each rule. In our augmented system (Aug-TRIPS), we replaced these weights with a lexically-conditioned model based on the adaptation of Collins used by Bikel (2002), allowing more efficiency and (in some cases) better selection. Aug-TRIPS retains the same grammar and lexicon as TRIPS, but uses its statistical model to determine the order in which unifications are attempted.

2 Experiments

We tested bracketing accuracy on the Monroe corpus (Stent, 2001), which contains collaborative emergency-management dialogues. Aug-TRIPS is comparable to TRIPS in accuracy, but produces fewer constituents (Table 1). The Bikel parser has slightly higher precision/recall than either TRIPS or Aug-TRIPS, since it can choose any bracketing structure regardless of semantic coherence, while the TRIPS systems must find a legal pattern of feature unifications. Aug-TRIPS also has better precision/recall when parsing the longer sentences (Table 2).

(training=9282)	Bikel	Aug-TRIPS	TRIPS
Recall	79.40	76.09	76.77
Precision	79.40	77.08	78.20
Complete Match	42.00	46.00	65.00
% Constit. Reduction	-	36.96	0.00

Table 1: Bracketing accuracy for 100 random sentences ≥ 2 words.

	> 7 Aug-TRIPS	> 7 TRIPS
Recall	73.25	71.00
Precision	74.78	73.44
Complete Match	22.50	37.50

Table 2: Bracketing accuracy for the 40 sentences $>$ 7 words.

Since our motivation for unification parsing is to reveal semantics as well as syntax, we next evaluated Aug-TRIPS's production of correct interpretations at the sentence level, which require complete correctness not only of the bracketing structure but of the sense chosen for each word and the thematic

Proceedings of the Ninth International Workshop on Parsing Technologies (IWPT), pages 198–199,
Vancouver, October 2005.

roles of each argument (Tetreault et al., 2004).

For this task, we modified the probability model to condition on the senses in our lexicon rather than words. For instance, the words "two thousand dollars" are replaced with the senses "number number-unit money-unit". This allows us to model lexical disambiguation explicitly. The model generates one or more senses from each word with probability $P(sense|word, tag)$, and then uses sense statistics rather than word statistics in all other calculations. Similar but more complex models were used in the PCFG-sem model of Toutanova et al. (2005) and using WordNet senses in Bikel (2000).

We used the Projector dialogues (835 sentences), which concern purchasing video projectors. In this domain, Aug-TRIPS makes about 10% more interpretation errors than TRIPS (Table 3), but when parsing sentences on which TRIPS itself makes errors, it can correct about 10% (Table 4).

(training=310)	TRIPS	Aug-TRIPS
Correct	26	21
Incorrect	49	54
% Reduction in Constituents	0%	45%

Table 3: Sentence-level accuracy on 75 random sentences.

(training=396)	TRIPS	Aug-TRIPS
Correct	0	8
Incorrect	54	46
% Reduction in Constituents	0%	46%

Table 4: Sentence-level accuracy on 54 TRIPS error sentences

Our parser makes substantially fewer constituents than baseline TRIPS at only slightly lower accuracy. Tsuruoka et al. (2004) achieved a much higher speedup (30 times) than we did; this is partly due to their use of the Penn Treebank, which contains much more data than our corpora. In addition, however, their baseline system is a classic HPSG parser with no efficiency features, while our baseline, TRIPS, is designed as a real-time dialogue parser which uses hand-tuned weights to guide its search and imposes a maximum chart size.

Acknowledgements Our thanks to Will DeBeaumont and four anonymous reviewers.

References

James F. Allen, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski. 1996. A robust system for natural spoken dialogue. In *Proceedings of the 1996 Annual Meeting of the Association for Computational Linguistics (ACL'96)*.

Daniel Bikel. 2000. A statistical model for parsing and word-sense disambiguation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

Daniel Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Human Language Technology Conference (HLT)*, San Diego.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Stefan Riezler, Detlef Prescher, Jonas Kuhn, and Mark Johnson. 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM training. In *Proceedings of the 38th Annual Meeting of the ACL*, Hong Kong.

Stefan Riezler, Tracy H. King, Richard Crouch, and John T. Maxwell. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation. In *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia.

Amanda J. Stent. 2001. *Dialogue Systems as Conversational Partners*. Ph.D. thesis, University of Rochester.

Mary Swift, James Allen, and Daniel Gildea. 2004. Skeletons in the parser: Using a shallow parser to improve deep parsing. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, Geneva, Switzerland, August.

Joel Tetreault, Mary Swift, Preethum Prithviraj, Myroslava Dzikovska, and James Allen. 2004. Discourse annotation in the Monroe corpus. In *ACL workshop on Discourse Annotation*, Barcelona, Spain, July.

Kristina Toutanova, Christopher D. Manning, Dan Flickinger, and Stephan Oepen. 2005. Stochastic HPSG parse disambiguation using the Redwoods corpus. *Journal of Logic and Computation*.

Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Tsujii. 2004. Towards efficient probabilistic HPSG parsing: Integrating semantic and syntactic preference to guide the parsing. In *Proceedings of IJCNLP-04 Workshop: Beyond Shallow Analyses- Formalisms and Statistical Modeling for Deep Analyses*, Sanya City, China.