— Beyond PARSEVAL —
Towards Improved Evaluation Measures for Parsing Systems

John Carroll[1] Anette Frank[2] Dekang Lin[3] Detlef Prescher[2] Hans Uszkoreit[2]

[1] Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton BN1 9QH
UK

[2] Language Technology Lab
DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken
Germany

[3] Department of Computing Science
University of Alberta
Edmonton, Alberta
Canada, T6G 2H1

The Current Situation in Stochastic Parsing

The earliest corpus-based approaches to stochastic parsing (e.g. Sampson et al. (1989), Fujisaki et al. (1989), Sharman et al. (1990), Black (1992)) used a variety of data resources and evaluation techniques. With the creation of the Penn Treebank of English (Marcus et al., 1993) and the parser evaluation measures established by the PARSEVAL initiative (Black, 1992), new approaches to stochastic parsing and uniform evaluation regimes emerged (Magerman (1995), Charniak (1996), Collins (1996)), leading to impressive improvements in parser accuracy (Collins (1997), Charniak (2000), Bod (2001)).

In the meantime, annotated corpora have been built for several other languages, most notably the Prague Dependency Treebank for Czech (Hajic, 1998), and the NEGRA corpus for German (Skut et al., 1997). Well-known, but smaller corpora for English are the ATIS Corpus and SUSANNE. Many more corpora are available or under construction, e.g. the Penn treebanks for Chinese and Korean, the TIGER corpus for German, as well as corpora for Bulgarian, French, Italian, Portugese, Spanish, Turkish, etc. Annotation schemes in these treebanks vary, often motivated by language-specific characteristics. For example, dependency-based annotation is generally preferred for languages with relatively free word order.

More recently, in line with increasing interest in more fine-grained syntactic and semantic representations, stochastic parsing has been applied to several higher-order syntactic frameworks, such as unification-based grammars (Johnson et al., 1999), tree-adjoining grammars (Chen et al., 1999) and combinatory categorial grammars (Hockenmaier, 2001). In parallel, due to the lack of appropriate large-scale annotated training corpora, unsupervised methods have been investigated, i.e. training of manually written (context-free or unification-based) grammars on free text (Beil et al. (1999), Riezler et al. (2000), Bouma et al. (2001)).

As opposed to the PARSEVAL measures — which are based on phrase structure tree match — most of these novel parsing approaches use other evaluation measures, such as dependency-based, valence-based, exact, or selective category match.

Challenges for Parser Evaluation

Despite the emergence of stochastic parsing approaches using alternative syntactic frameworks, the currently established paradigm for evaluating stochastic parsing still consists of the combination of Penn Treebank English (Section 23) with PARSEVAL measures.

However, in practice (especially if we count industrial labs) parsing systems using treebank grammars are not representative of the field. Moreover, a strong trend in stochastic parsing is away from treebank grammars and towards higher-level syntactic frameworks and hand-built grammars.

Research in stochastic parsing with higher-order syntactic frameworks is therefore confronted with a lack of a common evaluation metrics: neither do the PARSEVAL measures straightforwardly correspond to dependency structures or other valence-based representations, nor have these alternative approaches come up with a common, agreed-on standard for evaluation. Furthermore, no common evaluation corpora exist for many alternative languages. To some extent, this problem has been circumvented by building small theory-specific treebanks (with the obvious drawbacks for supervised training and inter-comparability). In sum, the growing field in stochastic parsing with alternative syntactic models or languages other than English faces problems in benchmarking against the established Gold Standard.

As a consequence, the best-known stochastic parsers are trained for Penn Treebank English. Yet, to validate these parsers on a broader basis, it has to be evaluated how well these stochastic models carry over to languages with e.g. free word order, intricate long-distance phenomena, pro-drop properties, and agglutinative or clitic languages. Again, this presupposes the availability of annotated corpora and evaluation schemes appropriate to cover a broad range of diverse language types.

Towards a New Gold Standard

The current situation in stochastic parsing, as well as prospects for its future development, calls for a new and uniform scheme for parser evaluation which covers both shallow and deep grammars, different syntactic frameworks, and different language types.

What is needed is an annotation scheme bridging structural differences across diverse languages and frameworks. In practice, many researchers have been using their own evaluation metrics which, despite divergences, bear some common ground, namely higher-level syntactic annotations such as grammatical relations, dependencies, or subcategorization frames (Beil et al. (1999), Carroll et al. (2000), Collins et al. (1999), Hockenmaier (2001), etc). Such basic syntactic relations build on crucial, but underlying structural constraints, yet provide more abstract, functional information.

This information is not only an appropriate level of abstraction to bridge structural differences between languages and higher-level syntactic theories, but moreover, provides a basis for evaluation of partial, more shallow analysis systems, at a higher level of representation. For example, if the evaluation is against grammatical relation rather than phrase structure information, partial parsers extracting functional relations can be evaluated within the same setup as full parsers.

Starting from this state of affairs, one of the aims of the workshop will be to provide a forum for researchers in the field to discuss (define and agree on) a new, uniform evaluation metric which provides a basis for comparison between different parsing systems, syntactic frameworks and stochastic models, and how well they extend to languages of different types.

Definition of a new evaluation standard could be restrictive and flexible at the same time: flexible in that training can exploit fine-grained annotations of richer syntactic frameworks; and restrictive in that diverging analyses are then to be mapped to uniform (more coarse-grained) annotations for standardized evaluation.

Starting an Initiative

A previous LREC-hosted workshop on parser evaluation in 1998 in Granada brought together a number of people advocating parser evaluation based on dependencies or grammatical relations (Carroll and Briscoe (1998), Lin (1998), Bangalore et al. (1998)). The consensus of the concluding discussion at that workshop was that there is much common ground between these approaches, and that they constitute a viable alternative to the PARSEVAL measures.

In the meantime, as described above, many more corpora are under construction and novel stochastic parsing schemes are being developed, which call for an initiative for establishing a new, agreed-on evaluation standard for parsing which allows for comparison and benchmarking across alternative models and different language types.

The workshop is intended to bring together four parties: researchers in stochastic parsing, builders of annotated corpora, representatives from different syntactic frameworks, and groups with interests in and proposals for parser evaluation. As a kick-off initiative, the workshop should lead to collaborative efforts to work out a new evaluation metric, and to start initiatives for building or deriving sufficiently large evaluation corpora, and possibly, large training corpora according to the new metric.

In conclusion, stochastic parsing has now developed to a stage where new methods are emerging, both in terms of underlying frameworks and languages covered. These need to be brought together by means of a new evaluation metric to prepare the new generation of stochastic parsing.

Workshop Programme

The workshop comprises thematic papers focussing on benchmarking of stochastic parsing, parser evaluation, design of annotation schemes covering different languages, and different frameworks, as well as creation of high-quality evaluation corpora.

Intended as a forum for discussion, the workshop programme consists of paper presentations with discussion sessions and a panel, where important results of the workshop are summarized and discussed.

In the final session we intend to wrap-up, and plan a kick-off initiative leading to concrete action plans and the creation of working groups, as well as planning for future coordination. To maintain the momentum of this initiative we will work towards setting up a parsing competition based on new standard evaluation corpora and evaluation metric.

References

Srinivas Bangalore, Anoop Sarkar, Christine Doran, and Beth Ann Hockey. 1998. Grammar and parser evaluation in the xtag project. In *Workshop on the Evaluation of Parsing Systems*, LREC, Granada.

Franz Beil, Glenn Carroll, Detlef Prescher, Stefan Riezler, and Mats Rooth. 1999. Inside-outside estimation of a lexicalized PCFG for German. In *Proceedings of ACL'99*, College Park, MD.

Ezra Black. 1992. Meeting of interest group on evaluation of broad-coverage grammars of English. LINGUIST List 3.587, http://www.linguistlist.org/issues/3/3-587.html.

Rens Bod. 2001. What is the minimal set of fragments that achieves maximal parse accuracy? In *Proceedings of ACL-2001*.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of dutch. In *Computational Linguistics in The Netherlands 2000*.

John Carroll and Ted Briscoe. 1998. A survey of parser evaluation methods. In *Workshop on the Evaluation of Parsing Systems*, LREC, Granada.

Eugene Charniak. 1996. Tree-bank grammars. Technical Report CS-96-02, Brown University.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, Seattle, WA.

J. Chen, S. Bangalore, and K. Vijay-Shanker. 1999. New models for improving supertag disambiguation. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*.

M. Collins, J. Hajic, L. Ramshaw, and Ch. Tillman. 1999. A Statistical Parser for Czech. In *Proceedings of ACL 99*.

Michael Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, Santa Cruz, CA.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97)*, Madrid.

T. Fujisaki, F. Jelinek, J. Cocke, E. Black, and T. Nishino. 1989. A probabilistic method for sentence disambiguation. In *Proceedings of the 1st International Workshop on Parsing Technologies*.

J. Hajic. 1998. Building a syntactically annotated corpus: The prague dependency treebank. Issues of Valency and Meaning. Studies in Honour of Jarmila Panevova.

Julia Hockenmaier. 2001. Statistical parsing for ccg with simple generative models. In *Student Research Workshop of the 39th ACL/10th EACL*.

Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. 1999. Estimators for stochastic "unification-based" grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, MD.

D. Lin. 1998. Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*, LREC, Granada.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*, Cambridge, MA.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

S. Riezler, D. Prescher, J. Kuhn, and M. Johnson. 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM training. In *Proc. of ACL-2000*.

G. Sampson, R. Haigh, and E. Atwell. 1989. Natural language analysis by stochastic optimization: a progress report on project april. *Journal of Experimental and Theoretical Artificial Intelligence*.

R. Sharman, F. Jelinek, and R. Mercer. 1990. Generating a grammar for statistical training. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC.