Latent Variable Grammars for Discontinuous Parsing

Invited Talk

Kilian Gebhardt

Technische Universität Dresden, Germany

Abstract Latent variable context-free grammars are powerful models for predicting the syntactic structure of sentences (Matsuzaki, Miyao, and Tsujii 2005; Petrov, Barrett, et al. 2006; Petrov and Klein 2007). When trained on annotated corpora, the resulting latent variables can be shown to capture different distributions for, e.g., NPs in subject and object position. Several languages (and in consequence also syntactic treebanks for these languages) such as Dutch (Lassy van Noord 2009), German (NeGra, Skut et al. 1997; TiGer Brants et al. 2004), but also English (Penn Treebank, Marcus, Santorini, and Marcinkiewicz 1993, Evang and Kallmeyer 2011) contain structures that cannot be adequately modelled by context-free grammars. In consequence, a class of more power grammar formalisms called mildly context-sensitive has been studied (cf. Kallmeyer 2010). Although parsing with these models is polynomial in the length of the input sentence (Seki et al. 1991), it has for a long time been regarded prohibitively slow. However, in recent years it was shown that the application of mildly-context sensitive grammars is feasible in coarse-to-fine parsing approaches (van Cranenburgh 2012; Ruprecht and Denkinger 2019).

In this talk I consider how both the latent variable approach and mildly context-sensitive grammars can be joined and applied to discontinuous treebanks:

1. A large class of latent variable grammars can be captured as a probabilistic regular tree grammar combined with an algebra. I show how the training methodology of latent variable PCFG can be generalized for this class.

2. I recall two mildly context-sensitive grammar formalisms: linear context-free rewriting systems (LCFRS, Vijay-Shanker, Weir, and Joshi 1987) and hybrid grammars (Nederhof and Vogler 2014; Gebhardt, Nederhof, and Vogler 2017). In particular, I consider the induction of hybrid grammars, which can be parametrized such that the polynomial complexity of parsing is of bounded degree. This way also hybrid grammars that are structurally equivalent to finite state automata can be obtained.

3. I analyse different trends when training latent variable LCFRS and hybrid grammars on different discontinuous treebanks and applying them for parsing.

Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing, pages 4–6
Dresden, Germany, September 23-25, 2019.

References

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit (2004). "TIGER: Linguistic Interpretation of a German Corpus". In: *Research on Language and Computation* 2 (4), pp. 597–620. DOI: 10.1007/s11168-004-7431-3.

van Cranenburgh, Andreas (2012). "Efficient parsing with Linear Context-Free Rewriting Systems". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 460–470. URL: https://www.aclweb.org/anthology/E12-1047.

Evang, Kilian and Laura Kallmeyer (2011). "PLCFRS Parsing of English Discontinuous Constituents". In: *Proceedings of the 12th International Conference on Parsing Technologies*. Dublin, Ireland, pp. 104–116. ISBN: 978-1-932432-04-6. URL: https://www.aclweb.org/anthology/W11-2913.

Gebhardt, Kilian, Mark-Jan Nederhof, and Heiko Vogler (2017). "Hybrid Grammars for Parsing of Discontinuous Phrase Structures and Non-Projective Dependency Structures". In: *Computational Linguistics* 43 (3), pp. 465–520. DOI: 10.1162/COLI_a_00291.

Kallmeyer, Laura (2010). *Parsing Beyond Context-Free Grammars*. 1st. Springer Publishing Company, Incorporated. ISBN: 364214845X, 9783642148453.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". In: *Computational Linguistics* 19 (2), pp. 313–330. URL: http://aclweb.org/anthology/J93-2004.

Matsuzaki, Takuya, Yusuke Miyao, and Jun'ichi Tsujii (2005). "Probabilistic CFG with Latent Annotations". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ann Arbor, Michigan, pp. 75–82. DOI: 10.3115/1219840.1219850.

Nederhof, Mark-Jan and Heiko Vogler (2014). "Hybrid Grammars for Discontinuous Parsing". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland, pp. 1370–1381. URL: https://www.aclweb.org/anthology/C14-1130.

van Noord, Gertjan (2009). "Huge Parsed Corpora in LASSY". In: *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*. Groningen, The Netherlands.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein (2006). "Learning Accurate, Compact, and Interpretable Tree Annotation". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pp. 433–440. DOI: 10.3115/1220175.1220230.

Petrov, Slav and Dan Klein (2007). "Improved Inference for Unlexicalized Parsing". In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York, pp. 404–411. URL: https://www.aclweb.org/anthology/N07-1051.

Ruprecht, Thomas and Tobias Denkinger (2019). “Implementation of a Chomsky-Schützenberger n-best parser for weighted multiple context-free grammars”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 178–191. DOI: 10.18653/v1/N19-1016.

Seki, Hiroyuki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami (1991). “On multiple context-free grammars”. In: *Theoretical Computer Science* 88 (2), pp. 191–229. ISSN: 0304-3975. DOI: 10.1016/0304-3975(91)90374-B.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit (1997). “An Annotation Scheme for Free Word Order Languages”. In: *Fifth Conference on Applied Natural Language Processing*. Washington, DC, USA: Association for Computational Linguistics, pp. 88–95. DOI: 10.3115/974557.974571.

Vijay-Shanker, Krishnamurti, David J. Weir, and Aravind K. Joshi (1987). “Characterizing Structural Descriptions Produced by Various Grammatical Formalisms”. In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. Stanford, California, USA, pp. 104–111. DOI: 10.3115/981175.981190.

Speaker’s homepage: https://wwwtcs.inf.tu-dresden.de/~kilian/