Intersection for weighted formalisms

Mark-Jan Nederhof
School of Computer Science
University of St Andrews
North Haugh, St Andrews
Fife, KY16 9SX
United Kingdom

Abstract

The paradigm of parsing as intersection has been used throughout the literature to obtain elegant and general solutions to numerous problems involving grammars and automata. The paradigm has its origins in (Bar-Hillel et al., 1964), where a general construction was used to prove closure of context-free languages under intersection with regular languages. It was pointed out by (Lang, 1994) that such a construction isolates the parsing problem from the recognition problem. The latter can be solved by a reduction of the outcome of intersection.

The paradigm has been extended in various ways, by considering more powerful formalisms, such as tree adjoining grammars (Vijay-Shanker and Weir, 1993), simple RCGs (Bertsch and Nederhof, 2001), tree grammars (Nederhof, 2009), and probabilistic extensions of grammatical formalisms (Nederhof and Satta, 2003). Different applications have been identified, such as computation of distances between languages (Nederhof and Satta, 2008), and parameter estimation of probabilistic models (Nederhof, 2005).

The lecture will focus on another application, namely the computation of prefix probabilities (Nederhof and Satta, 2011c) and infix probabilities (Nederhof and Satta, 2011a) and will address novel generalisations to linear context-free rewriting systems (Nederhof and Satta, 2011b).

References

Y. Bar-Hillel, M. Perles, and E. Shamir. 1964. On formal properties of simple phrase structure grammars. In Y. Bar-Hillel, editor, *Language and Information: Selected Essays on their Theory and Application*, chapter 9, pages 116–150. Addison-Wesley, Reading, Massachusetts.

E. Bertsch and M.-J. Nederhof. 2001. On the complexity of some extensions of RCG parsing. In *Proceedings of the Seventh International Workshop on Parsing Technologies*, pages 66–77, Beijing, China, October.

B. Lang. 1994. Recognition can be harder than parsing. *Computational Intelligence*, 10(4):486–494.

M.-J. Nederhof and G. Satta. 2003. Probabilistic parsing as intersection. In *8th International Workshop on Parsing Technologies*, pages 137–148, LORIA, Nancy, France, April.

M.-J. Nederhof and G. Satta. 2008. Computation of distances for regular and context-free probabilistic languages. *Theoretical Computer Science*, 395:235–254.

M.-J. Nederhof and G. Satta. 2011a. Computation of infix probabilities for probabilistic context-free grammars. In *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1213–1221, Edinburgh, Scotland, July.

M.-J. Nederhof and G. Satta. 2011b. Prefix probabilities for linear context-free rewriting systems. In *Proceedings of the 12th International Conference on Parsing Technologies*, Dublin, Ireland, October.

M.-J. Nederhof and G. Satta. 2011c. Prefix probability for probabilistic synchronous context-free grammars. In *49th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 460–469, Portland, Oregon, June.

M.-J. Nederhof. 2005. A general technique to train language models on language models. *Computational Linguistics*, 31(2):173–185.

M.-J. Nederhof. 2009. Weighted parsing of trees. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 13–24, Paris, France, October.

K. Vijay-Shanker and D.J. Weir. 1993. The use of shared forests in tree adjoining grammar parsing. In *Sixth*

Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, pages 1–2,
Blois (France), July 12-15, 2011. ©2011 Association for Computational Linguistics

Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, pages 384–393, Utrecht, The Netherlands, April.