Practical problems with Chomsky-Schützenberger parsing for weighted multiple context-free grammars*

Tobias Denkinger

Faculty of Computer Science, Technische Universität Dresden,
Nöthnitzer Str. 46, 01062 Dresden, Germany
tobias.denkinger@tu-dresden.de

The Chomsky-Schützenberger theorem (short: CS-theorem) [CS63, Prop. 2] is well-known in the formal languages community. It states that the language $L(G)$ of a context-free grammar (short: CFG) G can be expressed in terms of a homomorphism h, a regular language R, and a Dyck language D: $L(G) = h(R \cap D)$. Hulden [Hul11] described a method to use this decomposition to solve the parsing problem for CFGs. The *parsing problem* is to output for a given CFG G and a given word w over the alphabet of G the set of abstract syntax trees of w in G. If G is a weighted CFG, then we can formulate the *k-best parsing problem* [HC05]: Given a weighted CFG G, a partial order $\trianglelefteq$ on the weights, a natural number n, and a word w, output n abstract syntax trees of w in G that have the smallest weights with respect to $\trianglelefteq$. Hulden [Hul11, Sec. 4.1] briefly discusses how to use the CS-theorem to solve the 1-best parsing problem for probabilistic CFGs (i.e. CFGs weighted with probabilities).

In the meantime, the CS-theorem has been generalised to CFGs weighted with unital valuation monoids [DV13, Thm. 2], to multiple context-free grammars (short: MCFGs) [YKS10, Thm. 3], and to MCFGs weighted with commutative strong bimonoids [Den15, Thm. 19], among others.

MCFGs [SMFK91] and the expressively equivalent linear context-free rewriting systems [VWJ87] are currently being studied in the natural language processing community because they can express the non-projective constituents and discontinuous dependencies that occur in natural languages [Mai10; KS09]. Based on the CS-theorem for MCFGs weighted with commutative strong bimonoids, we developed an algorithm to solve the k-best parsing problem [Den17, Alg. 3] in the spirit of Hulden [Hul11]. However, our algorithm does not necessarily terminate.

In this talk [and in Den17, Sec. 5], we introduce our parsing algorithm and identify four restrictions to the complete commutative strong bimonoid $(\mathcal{A}, +, \cdot, 0, 1)$, the $\mathcal{A}$-weighted MCFG G, and the partial order $\trianglelefteq$ that ensure its termination: **(1)** G needs to be *restricted*, i.e. there are no paths with only rules of weight 1 in a derivation tree of G on which some rule occurs more than once, **(2)** $\cdot$ needs to *respect* $\trianglelefteq$, i.e. for any $a, b \in \mathcal{A}$ holds $a \trianglelefteq a \cdot b$, **(3)** $\cdot$ needs to have *arbitrarily large powers*, i.e. for any $a, b \in \mathcal{A}$ there is a $k \in \mathbb{N}$ such that $a \trianglelefteq b^k$, and **(4)** $\mathcal{A}$ needs to be *$\trianglelefteq$-factorisable*, i.e. for each $a \in \mathcal{A}$ there are $a_1, a_2 \in \mathcal{A}$ such that $a_1 \cdot a_2 = a$. Those four requirements are met, for example, if $\mathcal{A}$ is the Viterbi semiring $([0, 1], \max, \cdot, 0, 1)$, G is proper (i.e. weights of rules with the same left-hand side sum up to 1), and $\trianglelefteq = \geq$.

*This work is an excerpt of Denkinger [Den17]

References

[CS63] N. Chomsky and M. P. Schützenberger. "The algebraic theory of context-free languages". In: *Computer Programming and Formal Systems, Studies in Logic* (1963), pp. 118–161. DOI: 10.1016/S0049-237X(09)70104-1.

[Den15] T. Denkinger. "A Chomsky-Schützenberger representation for weighted multiple context-free languages". In: *FSMNLP*. 2015. URL: http://www.aclweb.org/anthology/W10-2506.

[Den17] T. Denkinger. "Chomsky-Schützenberger parsing for weighted multiple context-free languages". In: *JLM* 5.1 (2017), pp. 3–55. DOI: 10.15398/jlm.v5i1.159.

[DV13] M. Droste and H. Vogler. "The Chomsky-Schützenberger Theorem for Quantitative Context-Free Languages". In: *DLT*. Ed. by M.-P. Béal and O. Carton. Vol. 7907. LNCS. 2013, pp. 203–214. DOI: 10.1007/978-3-642-38771-5_19.

[HC05] L. Huang and D. Chiang. "Better k-best Parsing". In: *IWPT*. Parsing '05. 2005, pp. 53–64. URL: http://dl.acm.org/citation.cfm?id=1654494.1654500.

[Hul11] M. Hulden. "Parsing CFGs and PCFGs with a Chomsky-Schützenberger Representation". In: *HLT*. Ed. by Z. Vetulani. Vol. 6562. LNCS. 2011, pp. 151–160. DOI: 10.1007/978-3-642-20095-3_14.

[KS09] M. Kuhlmann and G. Satta. "Treebank Grammar Techniques for Non-projective Dependency Parsing". In: *EACL*. 2009, pp. 478–486. DOI: 10.3115/1609067.1609120.

[Mai10] W. Maier. "Direct Parsing of Discontinuous Constituents in German". In: *NAACL HLT*. 2010, pp. 58–66. URL: http://dl.acm.org/citation.cfm?id=1868771.1868778.

[SMFK91] H. Seki et al. "On multiple context-free grammars". In: *TCS* 88.2 (1991), pp. 191–229. DOI: 10.1016/0304-3975(91)90374-B.

[VWJ87] K. Vijay-Shanker, D. J. Weir, and A. K. Joshi. "Characterizing Structural Descriptions Produced by Various Grammatical Formalisms". In: *ACL*. 1987, pp. 104–111. DOI: 10.3115/981175.981190.

[YKS10] R. Yoshinaka, Y. Kaji, and H. Seki. "Chomsky-Schützenberger-type characterization of multiple context-free languages". In: *LATA*. Ed. by A.-H. Dediu, H. Fernau, and C. Martín-Vide. 2010, pp. 596–607. DOI: 10.1007/978-3-642-13089-2_50.