Efficient Parsing with Linear Context-Free Rewriting Systems

Andreas.van.Cranenburgh@huygens.knaw.nl

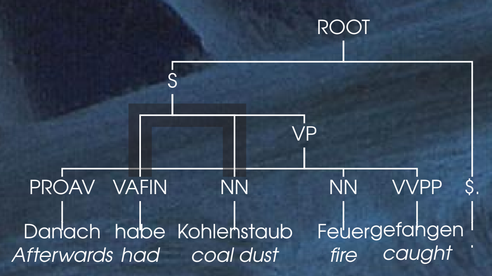
Institute for Logic, Language and Computation
University of Amsterdam

Problem

- Linear Context-Free Rewriting Systems (LCFRS) subsume a variety of mildly context-sensitive grammar formalisms.
- String rewriting LCFRSs can be used as a discontinuous treebank grammar.
- However, parsing LCFRS is too complex when $|w| > 30$ (Kallmeyer and Maier, 2010; van Cranenburgh et al., 2011).

Discontinuity

A discontinously annotated tree from the German Negra corpus:



$\text{ROOT}(ab) \rightarrow \text{S}(a)\ \text{\$.}(b)$
$\text{S}(abcd) \rightarrow \text{VAFIN}(b)\ \text{NN}(c)\ \text{VP}_2(a, d)$
$\text{VP}_2(a, bc) \rightarrow \text{PROAV}(a)\ \text{NN}(b)\ \text{VVPP}(c)$
$\text{PROAV(Danach)} \rightarrow \epsilon$
$\text{VAFIN(habe)} \rightarrow \epsilon$
$\text{NN(Kohlenstaub)} \rightarrow \epsilon$
$\text{NN(Feuer)} \rightarrow \epsilon$
$\text{VVPP(gefangen)} \rightarrow \epsilon$
$\text{\$.}(.) \rightarrow \epsilon$

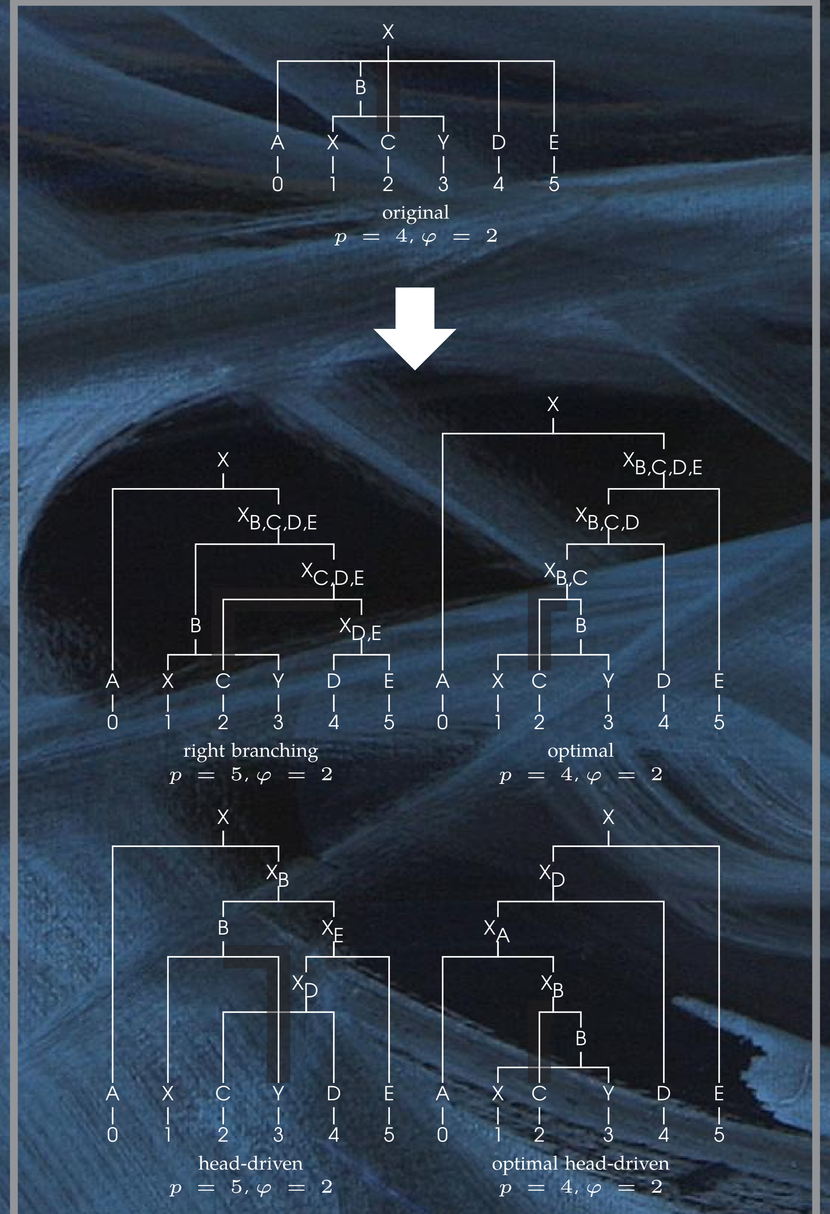
fanout (φ): number of components that a non-terminal rewrites (i.e., amount of discontinuity)

parsing complexity (p): sum of fanouts in LHS and RHS of a rule (i.e., number of comparisons needed to apply a rule)

Contributions

- Optimal binarizations (Gildea, 2010) do not eliminate the problem.
- Instead we solve the problem with a PCFG-approximation in a coarse-to-fine approach.
- Punctatuation re-attachement without spurious discontinuities. Proper evaluation ignores punctuation & ROOT node.

Binarization

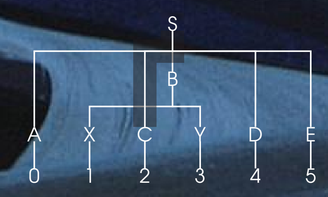


Four binarization strategies. C is the head node. Underneath each tree is the maximum parsing complexity and fan-out among its productions.

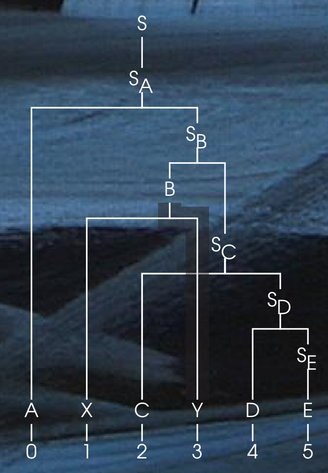
Method

PCFG-approximation of LCFRS grammar (after Barthélemy et al., 2001):

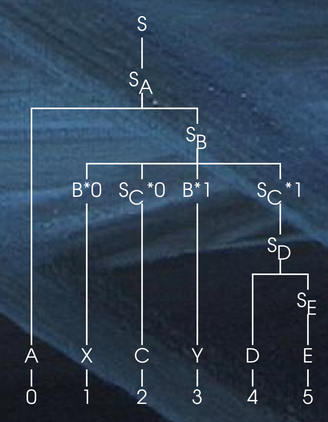
1. Original (discontinuous) tree



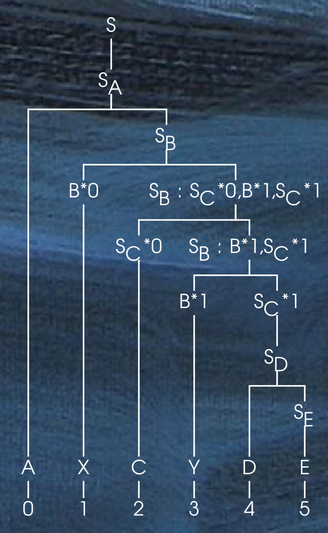
2. Binarize discontinuous tree, optionally with Markovization



3. Split discontinuous nodes into components, marked with indices



4. A binary normal form is applied; all productions are either unary, binary, or lexical.



Results

	Markovization	φ, p	binarization	parsing
right branching	v=1, h=∞	4, 8	2 s	246 s
optimal	v=1, h=∞	4, 8	46 s	194 s
head-driven	v=1, h=2	4, 9	3 s	2860 s
optimal head-driven	v=1, h=2	4, 8	29 s	717 s

Table: The effect of binarization strategies on parsing efficiency, with sentences from the development section of NEGRA-25. Longer sentences infeasible.

	words	F_1	EX
PLCFRS, dev set	≤ 25	72.37	36.58
Split-PCFG, dev set	≤ 25	70.74	33.80
Split-PCFG, dev set	≤ 40	66.81	27.59
CFG-CTF, PLCFRS, dev set	≤ 40	67.26	27.90
CFG-CTF, Disco-DOP, dev set	≤ 40	74.27	34.26
CFG-CTF, Disco-DOP, test set	≤ 40	72.33	33.16
CFG-CTF, Disco-DOP, dev set	∞	73.32	33.40
CFG-CTF, Disco-DOP, test set	∞	71.08	32.10

Table: Results on NEGRA-25 and NEGRA-40 with the CFG-CTF method.
Disco-DOP: A discontinuous tree-substitution grammar (van Cranenburgh et al., 2011).

References

François Barthélemy, Pierre Boullier, Philippe Deschamp, and Éric de la Clergerie. 2001. Guided parsing of range concatenation languages. In *Proc. of ACL*, pages 42–49.

Daniel Gildea. 2010. Optimal parsing strategies for linear context-free rewriting systems. In *Proceedings of NAACL HLT 2010.*, pages 769–776.

Laura Kallmeyer and Wolfgang Maier. 2010. Data-driven parsing with probabilistic linear context-free rewriting systems. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 537–545.

Andreas van Cranenburgh, Remko Scha, and Federico Sangati. 2011. Discontinuous data-oriented parsing: A mildly context-sensitive all-fragments grammar. In *Proceedings of SPMRL*, pages 34–44.

Painting: Christine Bittremieux (2007), Untitled. 70×100 cm.
Detail. Oil on canvas. `www.bittremieux.nl`