Transformation schemes for context-free grammars:

structural, algorithmic, linguistic applications

Eli Shamir

Hebrew university of Jerusalem, Israel

ISCOL - Haifa university - September 2014

Abstract

Schemes transforming context free grammars (CFG) range from symbol to subtle.

- Chomsky's normal form (CNF)
- Elimination of redundant symbols, ε rules
- Greibach's normal form (GNF) (subtle)

  all rules are $A \rightarrow Tx$. T terminal (lexicalization)

GNF preserves language-derived sentences, but distorts the derivation trees. It has a short, elegant proof by Rozenkranz, based on algebraic-equations description of CFG. GNF enhances several structural applications (e.g Shamir's multi-valued homomorphism of CFG into the dyke language, Chomsky Schutzenberger's theorem… [1], the hardest context-free language [2,3].

GNF specific lexicalization became a goal in NLP parsing and compilers technology, and many schemes were developed for restricted classes of CFG (deterministic, LR(k)…).

We present a new top-down scheme called "dissection-rotation". It works perfectly for the subclass of non-expansive [NE] grammars (which coincides with the quasi-rational class [1]). The scheme has the form of a bounded operation tree (BOT), in which the nodes are labeled by {current grammar - in a product form; current operation}.

At the route of BOT the current grammar is #G (# is a sentinel playing an important role). All the leaves at the bottom of BOT are labeled by linear grammars G[i].

The main invariant of BOT: modulo cyclic rotation the derivation trees in U G[i] are exactly those of #G, and their weights (if the prescribed by the production rules) are also preserved.

A variant of the BOT scheme applies to grammars with bounded ambiguity degree, decomposing G to a bounded union of degree 1 grammars, thus providing positive answer to a question Sam Eilenberg posed (1970).

Applications: proofs of structural properties of NE grammars are facilitated, mainly by using a strong form of the Pumping Lemma, which holds for linear grammar.

The algorithems for membership or total membership weight of sentences in L(G) become as easy as for linear grammars (hence as for finite-state transducer) i.e in the parallel class NC(1). To construct actual parse trees or the optimal weight one, the dynamic programming algorithm (CYK, Earley) are invoked and improved to have quadratic time and linear space.

For NLP and possibly for compiler techniques, the BOT scheme provides a strong link to finite-state transducers which have many uses in all aspects of computational linguistics.

Ambiguity – it relation to expansiveness in grammars and resolution by cyclic rotation are studied. Consider the bible verse in book of Job chapter 6 verse 14 (six Hebrew words). Translated to English: "a friend should extend # mercy to the sufferer, $ even if he abandons God's fear."

- The ambiguity here is anaphoric, does the pronoun "he" refer to the sufferer or to the friend? The poetic beautiful answer is: to both.
- The rotated sentences, starting at the symbols # and $, resolve the ambiguity one way or the other.
- A politically loaded example: "The policeman shot # the boy $ with the gun."

References

1. J. Autebert, J. Berstel and L. Boasson, Context-free language and pushdown automata. Chap. 3 In: *handbook of formal languages* Vol 1. G. Rozenberg and A. Salomaa (eds.), Springer-Verlag 1997.

2. Y. Bar-Hillel, H. Gaifman and E. Shamir, On categorical and phrase structure grammars. *Bulletin research council of Israel*, vol. 9f (1960), 1-16.

3. S. Greibach. The hardest context-free language. *SIAM J. on computing* 3 (1973), 304-310.

4. E. Shamir. Some inherently ambiguous context-free languages. *Inf. and Control 18* (1971), 355-363.