On the weak generative capacity of weighted context-free grammars*

Anders Søgaard
University of Potsdam
soegaard@ling.uni-potsdam.de

Abstract

It is shown how weighted context-free grammars can be used to recognize languages beyond their weak generative capacity by a one-step constant time extension of standard recognition algorithms.

1 Introduction

Weighted context-free grammars (WCFGs) are used to disambiguate strings and thus filter out subsets of the tree languages of the underlying context-free grammars (CFGs). Weights can either be used as probabilities, i.e. higher weights are preferred, or as penalities, i.e. lower weights are preferred. The first convention, also followed by Smith and Johnson (2007), is followed here. The subsets of the tree languages that consist of the heaviest tree for each yield are called the Viterbi tree languages. String languages are the yields of tree languages, and Viterbi string languages are the yields of Viterbi tree languages.

Infante-Lopez and de Rijke (2006) show that the Viterbi tree languages strictly extend the tree languages.

The idea explored in this paper is simple. If trees must have particular weights for their yields to be recognized, weights can be used to encode non-local dependencies. Technically, the $\{r_1, \ldots, r_n\}$-language is defined as all the strings for which the heaviest, i.e. most probable, tree has weight $r_i \in \{r_1, \ldots, r_n\}$. It is shown that this class of languages includes common classes of context-sensitive languages. In other words, standard Viterbi-style recognition algorithms for WCFGs can be used to recognize these classes by a one-step look-up that checks if the weight of the heaviest tree is in $\{r_1, \ldots, r_n\}$. We say that $\{r_1, \ldots, r_n\}$-languages are $\{r_1, \ldots, r_n\}$-recognized.

Sect. 1.1 presents formal preliminaries and a Viterbi-style recognition algorithm for WCFGs. Note that for simplicity we restrict weights to be rational numbers.

Sect. 2 defines $\{r_1, \ldots, r_n\}$-languages and presents some examples of WCFGs that $\{r_1, \ldots, r_n\}$-recognize context-sensitive languages. Sect. 3 gives a rough characterization of the class of languages that can be $\{r_1, \ldots, r_n\}$-recognized by WCFGs.

Cortes and Mohri (2000) introduced a similar idea in the context of weighted finite-state automata (WFSAs) and showed that WFSAs can be used to $\{r_1, \ldots, r_n\}$-recognize context-free languages. Their results are extended in Sect. 4. It is shown that WFSAs can also be used to $\{r_1, \ldots, r_n\}$-recognize context-sensitive languages. It is shown, however, that the non-context-free languages that can be $\{r_1, \ldots, r_n\}$-recognized by WCFGs strictly extend the non-context-free languages that can be $\{r_1, \ldots, r_n\}$-recognized by WFSAs.

Sect. 5 discusses a more exact characterization of the weak generative capacity of WCFGs in this view. Coprime WCFGs (CWCFGs), i.e. a subclass of WCFGs where the weights can be partitioned into reciprocal coprimes, are introduced. It is conjectured that the infinite hierarchy of k-CWCFGs is non-collapsing, and the classes of languages that can be $\{r_1, \ldots, r_n\}$-recognized by k-CWCFGs are characterized in terms of an untraditional modifi-

*Thanks to Mark Hopkins, Daniel Quernheim and the anonymous reviewers for helpful comments.

Coling 2008: Companion volume – Posters and Demonstrations, pages 99–102
Manchester, August 2008

cation of indexed grammars.

1.1 Preliminaries

A CFG is a 4-tuple $G = \langle N, T, P, S\rangle$ where N, T are finite and disjoint sets of nonterminal and terminal symbols, P a finite set of production rules of the form $A \rightarrow \phi$ where $A \in N$ and $\phi \in (N \cup T)^*$, and $S \in N$ is the start symbol. A WCFG is a 2-tuple $G' = \langle G, \pi\rangle$ where $G = \langle N, T, P, S\rangle$ is a CFG and $\pi : P \rightarrow \{\frac{m}{n} \mid m \in \mathbb{Z}^+, n \in \mathbb{Z}^+, m, n \neq 0\}$ a (total) weight function.

A left-most derivation $t(\omega)$ for some CFG $G = \langle N, T, P, S\rangle$ is a sequence of production rules $\langle p_1, \ldots, p_m\rangle$ with $1 \leq i \leq m : p_i \in P$ such that

$$S \overset{p_1}{\Longrightarrow} \phi_1 \ldots \phi_{m-1} \overset{p_m}{\Longrightarrow} \omega$$

ω is called the yield of $t(\omega)$. The tree language $T(G)$ is the set of all left-most derivations licensed by the production rules of G. The string language of G is the set of yields:

$$L(G) = \{\omega \mid t(\omega) \in T(G)\}$$

The accumulated weight of a derivation of a string ω $\pi(t(\omega))$ is the product of the weight of all the productions in $t(\omega)$. The Viterbi tree language of a WCFG then is:

$$V(G) = \{t(\omega) \mid t(\omega) \in \arg \max_{t'(\omega) \in T(G)} (\pi(t'(\omega)))\}$$

A simple Viterbi recognition algorithm for WCFGs is presented in Figure 1 for further reference.

2 Our extension

For a set of n many rational numbers $\{r_1, \ldots, r_n\}$, the language that is $\{r_1, \ldots, r_n\}$-recognized by the WCFG G, $L_{\{r_1,\ldots,r_n\}}(G)$, is defined:

$$L_{\{r_1,\ldots,r_n\}}(G) = \{\omega \mid t(\omega) \in V(G), \pi(t(\omega)) \in \{r_1, \ldots, r_n\}\}$$

Call the class of all languages that can be $\{r_1, \ldots, r_n\}$-recognized by a WCFG for all finite and non-empty sets of rational numbers $\{r_1, \ldots, r_n\}$ for *balanced* weighted context-free languages (BWCFLs). In all our examples $\{r_1, \ldots, r_n\}$ will be a singleton set.

Note that all there is needed to do to recognize the BWCFLs is to change line 7 of the Viterbi algorithm in Figure 1 to:

if $(S, r_i) \in t(0, n), r_i \in \{r_1, \ldots, r_n\}$ **then** ...

3 Bounds on weak generative capacity

The first result of this paper is the following:

Theorem 3.1. *The BWCFLs strictly extend the context-free languages.*

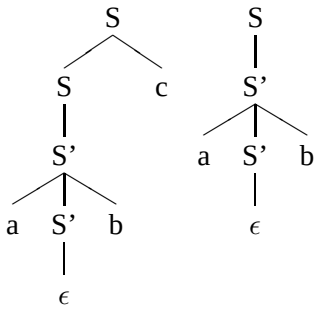
Proof. It is not difficult to see that any context-free language is a BWCFL. Simply construct a WCFG $G = \langle G', \pi\rangle$ for any CFG $G' = \langle N, T, P, S\rangle$ such that the weight associated with each production rule in P is $\frac{1}{1}$. It then holds that $L_{\{\frac{1}{1}\}}(G) = L(G')$.

The other direction is not very difficult either. It is shown that $\{a^n b^n c^n \mid n \geq 0\}$, which is non-context-free by the Bar-Hillel lemma, is a BWCFL. The language is, for instance, the set of strings $L_{\{\frac{1}{1}\}}(G)$ for the WCFG $G = \langle\langle\{S, S'\}, \{a, b, c\}, P, S\rangle, \pi\rangle$ where P is the following set of production rules, and π assigns the weights in the left column to the items in the right column:

$\frac{1}{2}$: S $\rightarrow$ Sc
$\frac{2}{1}$: S $\rightarrow$ S'
$\frac{2}{1}$: S' $\rightarrow$ aS'b
$\frac{1}{2}$: S' $\rightarrow$ ϵ

$L_{\{\frac{1}{1}\}} = \{a^n b^n c^n \mid n \geq 0\}$. Some example derivations are presented in Example 3.2. □

Example 3.2. Consider the only and thus heaviest tree for abc, resp. ab:



The weight of the left tree, whose yield is abc, is $\frac{1}{1}$. The weight of the left tree is $\frac{2}{1}$.

Consider also the $\{\frac{1}{1}\}$-language of $G = \langle\{S, D, T, T'\}, \{a, b, c, d\}, P, S\rangle$ with production rules P:

$\frac{1}{1}$: S $\rightarrow$ TD
$\frac{1}{2}$: D $\rightarrow$ dD
$\frac{1}{1}$: D $\rightarrow$ ϵ
$\frac{1}{1}$: T $\rightarrow$ aTc
$\frac{1}{1}$: T $\rightarrow$ T'
$\frac{2}{1}$: T' $\rightarrow$ bT'
$\frac{1}{1}$: T' $\rightarrow$ ϵ

BUILD$(t, [w_1 \dots w_n])$
1 **for** $j \leftarrow 1$ **to** n
2 **do** $t(j-1, j) \leftarrow \{(A, \alpha) \mid A \rightarrow w_j \in P, \log(\pi(A \rightarrow w_j)) = \alpha\}$
3 **for** $k \leftarrow (j-1)$ **to** 0
4 **do** $t(k, j) \leftarrow \{(A, \alpha + \beta) \mid A \rightarrow B \in P, \log(\pi(A \rightarrow B)) = \alpha,$
$(B, \beta) \in t(k, j)$, **if** $(A, \alpha') \in t(k, j)$ **then** $\alpha > \alpha'\}$
5 **for** $i \leftarrow (j-2)$ **to** 0
6 **do** $t(i, j) \leftarrow \{(A, \alpha + \beta + \kappa) \mid A \rightarrow BC \in P, \log(\pi(A \rightarrow BC)) = \alpha,$
$\exists k.(B, \beta) \in t(i, k), (C, \kappa) \in t(k, j)$, **if** $(A, \alpha') \in t(i, j)$ **then** $\alpha > \alpha'\}$
7 **if** $(S, r_i) \in t(0, n)$ **then return** *success* **else** *failure*

Figure 1: A Viterbi recognition algorithm for WCFGs

It should be relatively easy to see that $L(G) = \{a^n b^m c^n d^m \mid n \geq 0\}$.

It is not difficult to see that the BWCFLs are a subset of the context-sensitive languages. This follows from the fact that the left-most derivations in the Viterbi tree languages of WCFGs are linear in the length of the input string; in other words, BWCFLs can be recognized in nondeterministic linear space and thus by a linear bounded automaton. Since any language that can be represented by a linear bounded automaton is context-sensitive, the BWCFLs must be a subset of the context-sensitive ones.

The set of BWCFLs is also a subset of the range concatenation languages (Boullier, 1998) by the observation made in the introduction that they can be recognized in polynomial (i.e. cubic) time by standard algorithms and a one-step inspection of the weight of the heaviest tree; and by the fact that the range concatenation languages are exactly the languages that can be recognized in polynomial time (Boullier, 1998).

4 Weighted finite-state automata

Cortes and Mohri (2000) showed, in similar work, that WFSAs can be used to recognize context-free, i.e. non-regular, languages.

Example 4.1. The weighted finite-state automaton $T = \langle \{q_0, q_1\}, \{a, b\}, \delta, q_0, \{q_1\} \rangle$ with the following δ-transitions $\{\frac{1}{1}\}$-recognizes the language $L_{\{\frac{1}{1}\}}(T) = \{a^n b^n \mid n \geq 0\}$:

$$\begin{array}{rrcl} \frac{1}{2}: & \delta(q_0, a) & = & q_0 \\ \frac{1}{1}: & \delta(q_0, \lambda) & = & q_1 \\ \frac{2}{1}: & \delta(q_1, b) & = & q_1 \end{array}$$

It is not difficult to see that the strings $ab, aabb, \dots$ have derivations with weights $\frac{1}{1}$, whereas the string aab, for example, only has a derivation with weight $\frac{1}{2}$. Since $\frac{1}{2} \notin \{\frac{1}{1}\}$, $aab \notin L_{\{\frac{1}{1}\}}(T)$.

Cortes and Mohri (2000) also formulated an extension of WFSAs over cross-products of semirings that recognized certain context-sensitive, i.e. non-context-free languages, but their results can be considerably extended. The automaton in Example 4.2, for example, even recognizes a language conjectured to be outside the linear indexed languages, namely the MIX language (Gazdar, 1988).

Example 4.2. The weighted finite-state automaton $T = \langle \{q_0, q_1, q_2, q_3\}, \{a, b, c\}, \delta, q_0, \{q_0\} \rangle$ with the following δ-transitions $\{\frac{1}{1}\}$-recognizes the MIX language:

$$\begin{array}{rrcl} \frac{1}{8}: & \delta(q_0, a) & = & q_1 \\ \frac{1}{8}: & \delta(q_1, a) & = & q_2 \\ \frac{1}{8}: & \delta(q_2, a) & = & q_3 \\ \frac{1}{125}: & \delta(q_0, b) & = & q_1 \\ \frac{1}{125}: & \delta(q_1, b) & = & q_2 \\ \frac{1}{125}: & \delta(q_2, b) & = & q_3 \\ \frac{1}{729}: & \delta(q_0, c) & = & q_1 \\ \frac{1}{729}: & \delta(q_1, c) & = & q_2 \\ \frac{1}{729}: & \delta(q_2, c) & = & q_3 \\ \frac{90^3}{1}: & \delta(q_3, \lambda) & = & q_0 \end{array}$$

This example is a bit more complicated. Note that $8 \times 125 \times 729 = 90^3$. The strings $cab, bcabac, \dots$ have derivations with weights $\frac{1}{1}$, since $\frac{90^3}{8 \times 125 \times 729} = \frac{1}{1}$, whereas the string $cababa$, for instance, has no derivations with weight $\frac{1}{1}$. The string $cababa$ has exactly one derivation whose weight is $\frac{90^3}{8^2 \times 125}$.

5 Coprime WCFGs

A 2-CWCFG is a WCFG over subsets of the rational numbers $\mathbb{C} = \{\frac{1}{n} \mid n \in \Sigma\} \cup \{\frac{n}{1} \mid n \in \Sigma\}$

	B. (2000)	WCFGs
$\{a_1^n \dots a_k^n \mid n \geq 0\}$	✓	✓
MIX	✓	✓
$\{a^n b^m c^n d^m \mid m, n \geq 0\}$	✓	✓
$\{wcw \mid w \in \{a, b\}^*\}$	✓	✓

Figure 2: Classes of languages $\{r_1, \dots, r_n\}$-recognized by WCFGs and recognized by the extension in Boullier (2000).

where Σ is an arbitrary set of coprimes ($\Sigma \subseteq \mathbb{N}^*$) such that there is a bijection from the production rules onto themselves such that if a production rule has weight $\frac{1}{1}$ it is projected onto itself, and otherwise, i.e. if it has weight $\frac{1}{m}$ with $m \neq 1$ it is projected onto a production rule with weight $\frac{m}{1}$. A k-CWCFG for $k \geq 1$ is now the extension of CWCFG where the sets of production rules the product of whose weights is 1, can be of size at most k, e.g. the WFSA in Example 4.2 is a 3-CWCFG.

The infinite hierarchy of k-CWCFGs seems to be non-collapsing. A k-CWCFG $\{r_1, \dots, r_n\}$-recognizes the language $\{a_1^n \dots a_{2k}^n \mid n \geq 0\}$, but not $\{a_1^n \dots a_{2k+1}^n \mid n \geq 0\}$. It has this property in common with k-multiple context-free grammars (Seki et al., 1991). 2-CWCFG can be shown to be weakly equivalent with the extension of linear indexed grammars (LIGs) (Gazdar, 1988) where the stack is a multiset or a bag that is globally accessible and not just along spines. The universal recognition problem for this extension of LIGs can be shown to be NP-complete by reduction of the vertex cover problem, similar to Søgaard et al. (2007). The generalization to k-CWCFG requires stacks of stacks, but is otherwise relatively straight-forward.

6 Conclusions

It was shown how weighted context-free grammars can be used to recognize languages beyond their weak generative capacity by a one-step constant time extension of standard recognition algorithms. The class of languages that can be recognized this way strictly extends the context-free languages, but is included in the cubic time recognizable ones.

Boullier (2000) defines what he calls a "cubic time extension of CFG" that recognizes generalizations of the copy language that are beyond WCFG. It remains to be seen if the set of BWCFLs is a strict subset of the set of languages that can be recognized by this formalism. They all recognize the classes of languages in Figure 2.

References

Boullier, Pierre. 1998. Proposal for a natural language processing syntactic backbone. Technical report, INRIA, Le Chesnay, France.

Boullier, Pierre. 2000. A cubic time extension of context-free grammars. *Grammars*, 3(2–3):111–131.

Cortes, Corinna and Mehryar Mohri. 2000. Context-free recognition with weighted automata. *Grammars*, 3(2–3):133–150.

Gazdar, Gerald. 1988. Applicability of indexed grammars to natural languages. In Reyle, Uwe and Christian Rohrer, editors, *Natural language parsing and linguistic theories*, pages 69–94. Reidel, Dordrecht, the Netherlands.

Infante-Lopez, Gabriel and Maarten de Rijke. 2006. A note on the expressive power of probabilistic context free grammars. *Journal of Logic, Language and Information*, 15(3):219–231.

Seki, Hiroyuki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191–229.

Smith, Noah and Mark Johnson. 2007. Weighted and probabilistic context-free grammars are equally expressive. *Computational Linguistics*, 33(4):477–491.

Søgaard, Anders, Timm Lichte, and Wolfgang Maier. 2007. On the complexity of linguistically motivated extensions of tree-adjoining grammar. In *Proceedings of Recent Advances in Natural Language Processing 2007*, Borovets, Bulgaria.