Efficient Semiring-Weighted Earley Parsing

Andreas Opedal, Ran Zmigrod, Tim Vieira, Ryan Cotterell and Jason Eisner

ETH zürich

UNIVERSITY OF CAMBRIDGE

JOHNS HOPKINS UNIVERSITY

We present a **fast** version of Earley's algorithm in the form of a unified **deduction system** that can be executed with good asymptotic complexity. We obtain further **speedups** by allowing the grammar to be presented in the form of a **weighted finite-state automaton**, which allows similar productions to share computation. We carefully generalize our methods to handle **semiring-weighted grammars**; providing a method to compute **prefix weights** for non-commutative semirings and showing how the same time and space complexity can be achieved for weighted parsing as for unweighted parsing (Eisner, 2023).

Why do we like Earley's algorithm?

- Parses in $\mathcal{O}(N^3)$ time for general grammars, in $\mathcal{O}(N^2)$ time for unambiguous grammars and in $\mathcal{O}(N)$ time for bounded-state grammars
- Incremental left-to-right sentence processing
- Can be extended to compute next-word probabilities (Stolcke, 1995)
  - Cognitive modeling (Hale, 2001)
  - Constrained generation of LLMs (Shin et al., 2021; Roy et al., 2022; Fang et al., 2023)

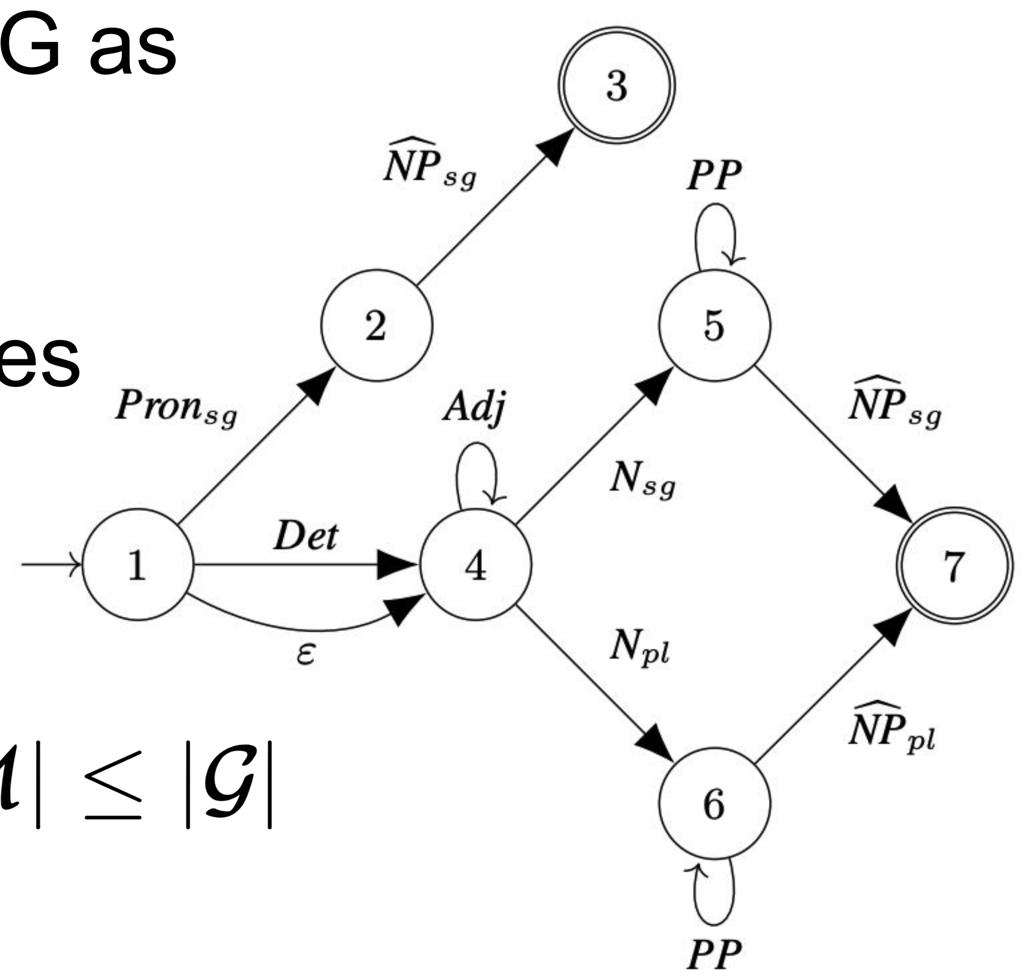
A faster Earley deduction system

	Earley	EarleyFast
Domains	$i,j,k \in \{0,\dots,N\}$ $\quad A,B \in \mathcal{N}\cup\{S'\}$ $\quad a\in\Sigma$ $\quad \rho,\mu,\nu \in (\mathcal{N}\cup\Sigma)^*$	
Items	$[i,j,A\to\mu\bullet\nu]$ $\quad [j,k,a]$ $\quad A\to\rho$	$[i,j,A\to\mu\bullet\nu]$ $\quad [j,k,a]$ $\quad A\to\rho$ $[i,j,A\to\bullet\star]$ $\quad [i,j,A\to\star\bullet]$
Axioms	$A\to\rho, \forall(A\to\rho)\in\mathcal{R}$ $[k-1,k,x_k], \forall k\in\{1,\dots,N\}$ $[0,0,S'\to\bullet S]$	$A\to\rho, \forall(A\to\rho)\in\mathcal{R}$ $[k-1,k,x_k], \forall k\in\{1,\dots,N\}$ $[0,0,S\to\bullet\star]$
Goal	$[0,N,S'\to S\bullet]$	$[0,N,S\to\star\bullet]$
Rules	PRED: $\dfrac{B\to\rho}{[j,j,B\to\bullet\rho]}\,[i,j,A\to\mu\bullet B\,\nu]$	PRED1: $\dfrac{}{[j,j,B\to\bullet\star]}\,[i,j,A\to\mu\bullet B\,\nu]$ PRED2: $\dfrac{B\to\rho}{[j,j,B\to\bullet\rho]}\,[j,j,B\to\bullet\star]$
	SCAN: $\dfrac{[i,j,A\to\mu\bullet a\,\nu] \quad [j,k,a]}{[i,k,A\to\mu\,a\bullet\nu]}$	SCAN: $\dfrac{[i,j,A\to\mu\bullet a\,\nu] \quad [j,k,a]}{[i,k,A\to\mu\,a\bullet\nu]}$
		COMP1: $\dfrac{[j,k,B\to\rho\bullet]}{[j,k,B\to\star\bullet]}$
	COMP: $\dfrac{[i,j,A\to\mu\bullet B\,\nu] \quad [j,k,B\to\rho\bullet]}{[i,k,A\to\mu\,B\bullet\nu]}$	COMP2: $\dfrac{[i,j,A\to\mu\bullet B\,\nu] \quad [j,k,B\to\star\bullet]}{[i,k,A\to\mu\,B\bullet\nu]}$

- 1-1 correspondence between proof trees and derivation trees
- From $\mathcal{O}(N^3|\mathcal{G}||\mathcal{R}|)$ to $\mathcal{O}(N^3|\mathcal{G}|)$

EarleyFSA

- Encode weighted CFG as weighted FSA
- Production rules with similar right-hand sides can be partially merged (prefix and suffix sharing)
- $\mathcal{O}(N^3|\mathcal{M}|)$, with $|\mathcal{M}| \le |\mathcal{G}|$



Preliminaries for weighted parsing

- Each production rule is given a weight from the set of values of some (commutative) semiring $\langle \mathbb{W}, \oplus, \otimes, ⓪, ① \rangle$
- The weight of a derivation tree is

$$w(T) \overset{\text{def}}{=} \bigotimes_{(A\to\rho)\in T} w(A\to\rho)$$

- A **weighted recognizer** returns

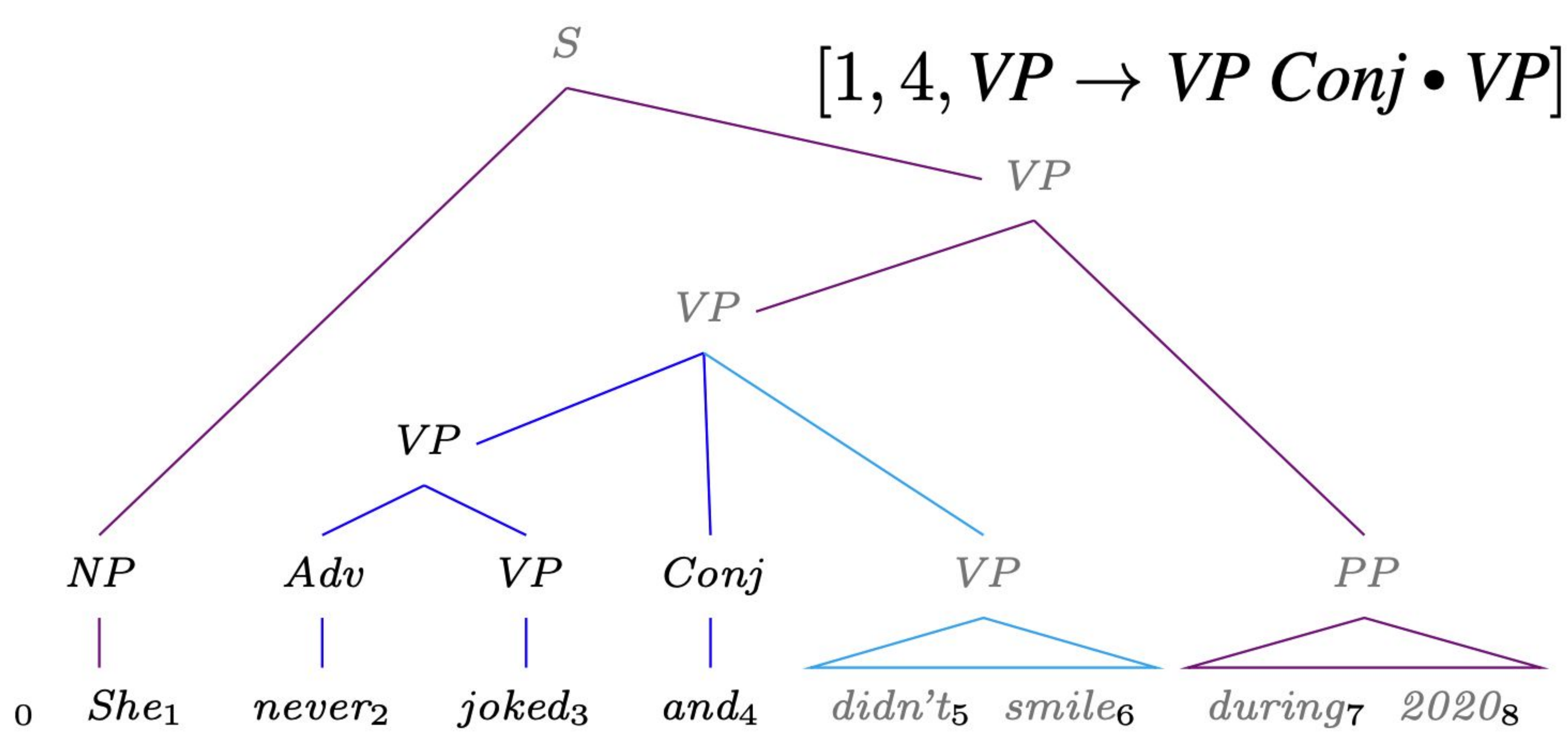
$$\mathrm{Z}_{\mathbf{x}} \overset{\text{def}}{=} w\left(S \overset{*}{\Rightarrow} \mathbf{x}\right) \overset{\text{def}}{=} \bigoplus_{T\in\mathcal{T}_{\mathbf{x}}} w(T)$$

- A **parser** returns all derivation trees of **x**
- **Prefix weights**:

$$w\left(S \overset{*}{\Rightarrow}_L \mathbf{y}\right) \overset{\text{def}}{=} \bigoplus_{\mathbf{z}\in\Sigma^*} w\left(S \overset{*}{\Rightarrow} \mathbf{yz}\right)$$

Semiring-weighted deductive parsing

- The **incomplete inside weight** of an item **V** = $[i,j,A\to\mu\bullet\nu]$ is the total weight of all its proofs
- Computed by $\oplus$-summing over the antecedent weights in all one-step proofs



- The **future inside weight** of **V** is the total weight of all ways to prove $[i,j,A\to\mu\,\nu\bullet]$ from **V**
- Requires computing **free sums**:

$$Z_A \overset{\text{def}}{=} \bigoplus_{T\in\mathcal{T}^A} \bigotimes_{B\to\rho\in T} w(B\to\rho)$$

- The **prefix outside weight** of **V** is the total weight of all ways to prove the goal item from $[i,j,A\to\mu\,\nu\bullet]$