

CS769 Advanced NLP

Syntactic Parsing I: Constituency Grammar

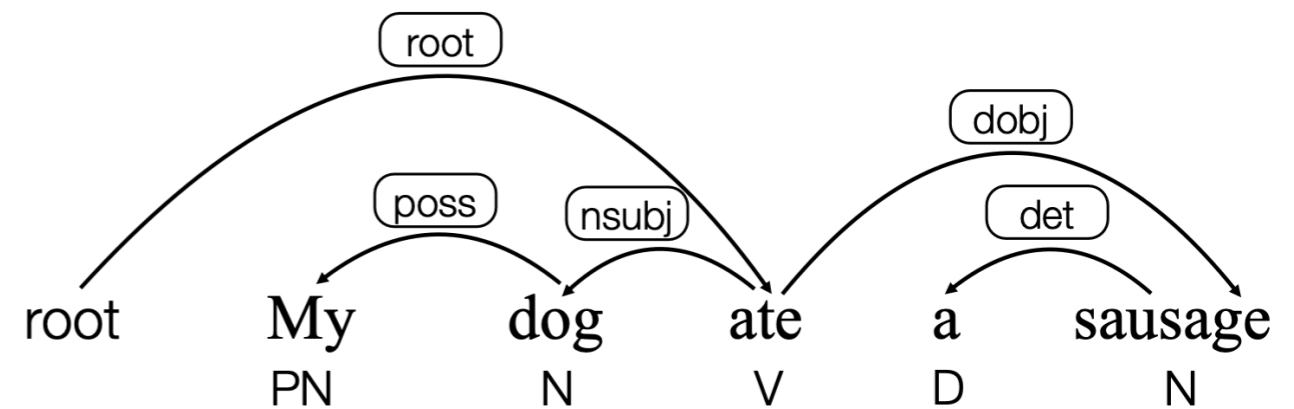
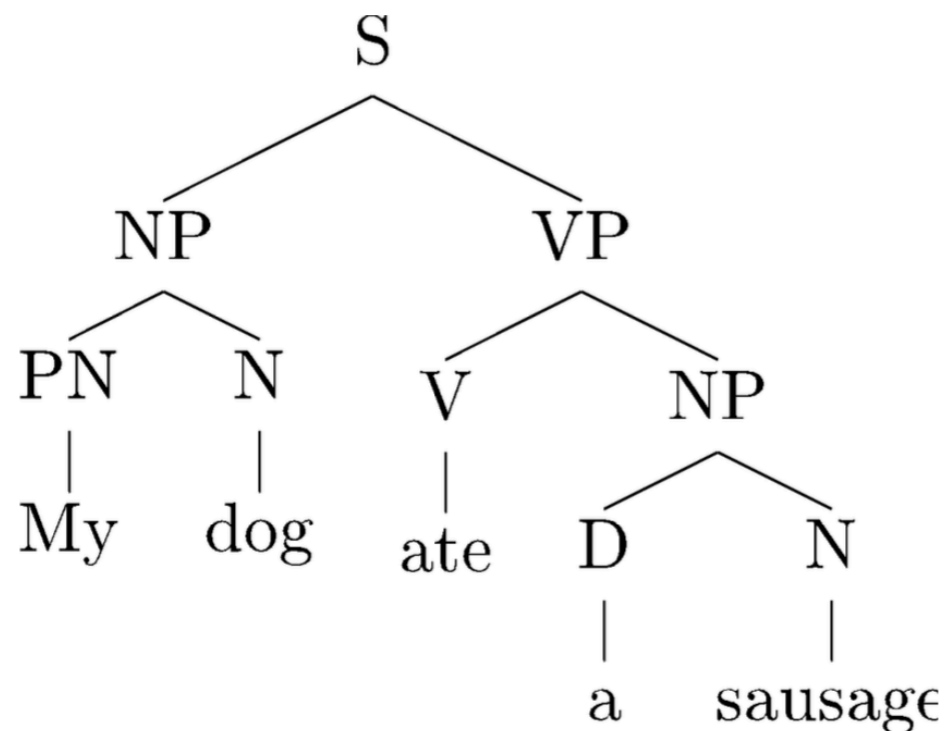
Junjie Hu



Slides adapted from Bob, Hao, Dan
<https://junjiehu.github.io/cs769-spring22/>

Syntactic Parsing

- The process of predicting **syntactic representations**
- Two types of linguistic structures:



Constituency (aka phrase structure) tree:

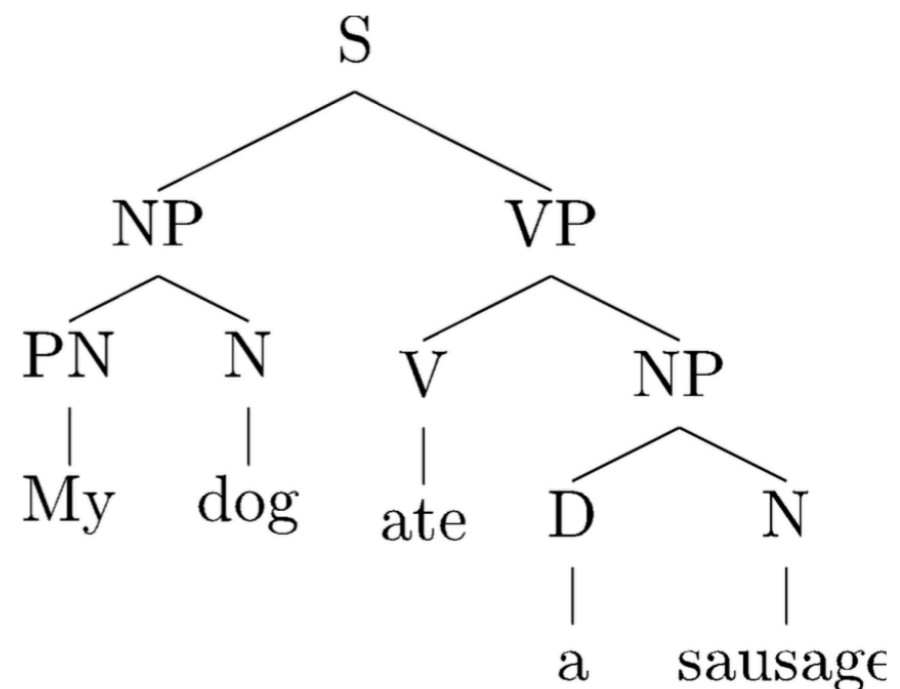
Focus on the structure of the sentence

Dependency tree:

Focus on relations between words

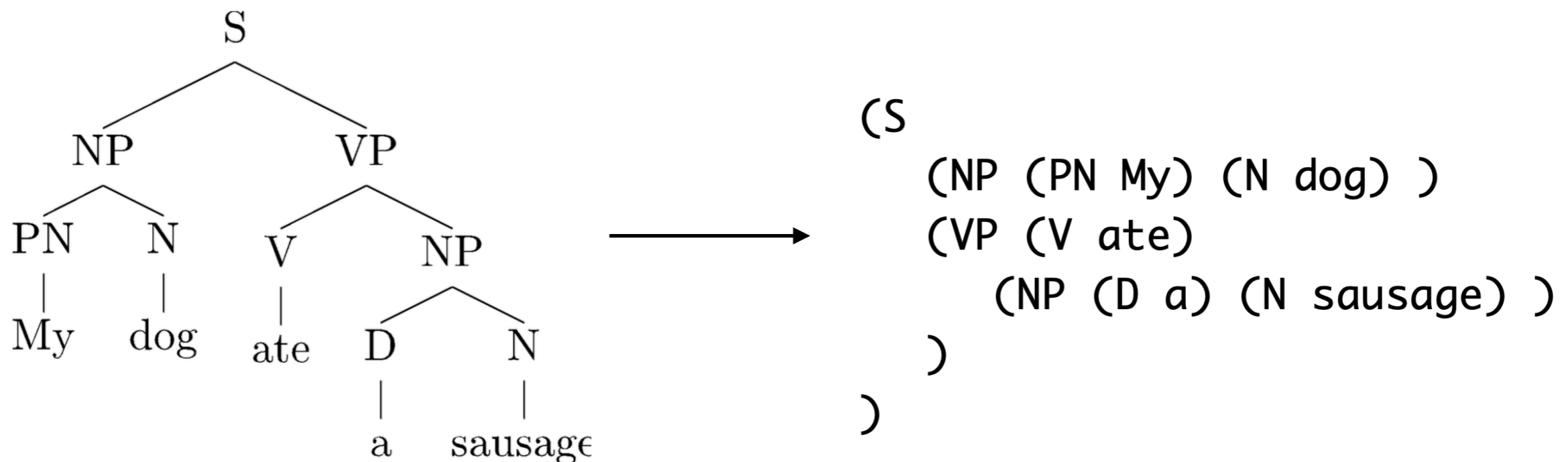
Constituency Trees

- Internal nodes (or non-terminals) correspond to phrases
 - S: a sentence
 - NP (noun phrase): My dog, a sandwich, ...
 - VP (verb phrase): ate a sausage, ...
 - PP (prepositional phrases): with a friend, in a car, ...
- Nodes immediately above words are part-of-speech tags (or preterminals).
 - PN: pronoun
 - D: determiner
 - V: verb
 - N: noun
 - P: preposition



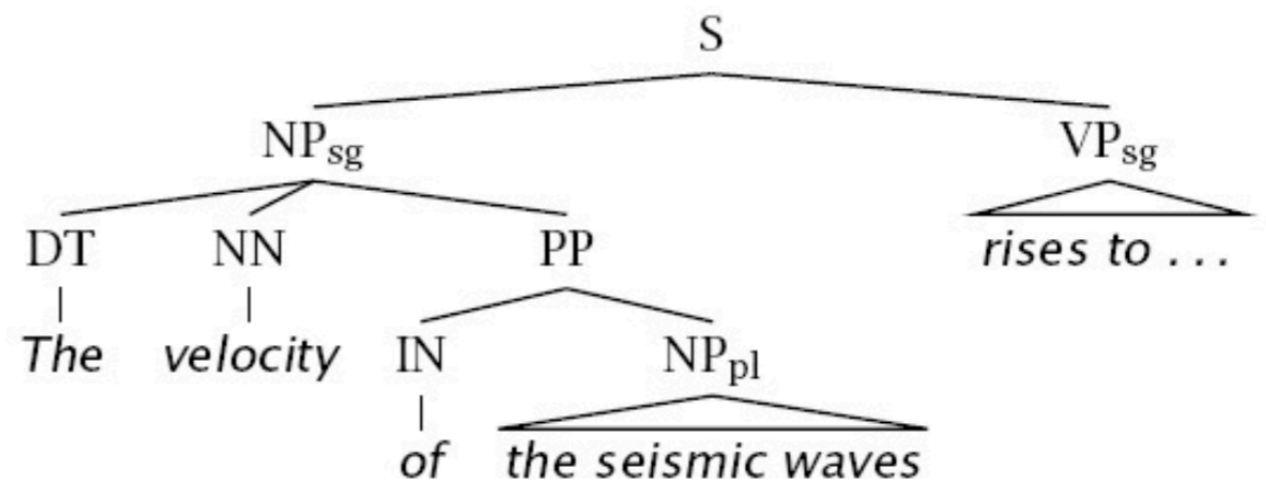
Bracketing notation

- Often convenient to represent a tree as **a bracketed sequence**:
- In principle, constituency tree can be an n-nary tree, however, it is easy to convert it to a binary tree (by adding a null non-terminal \emptyset). By convention, we often just represent the structure as a binary tree.



Constituency is not always clear

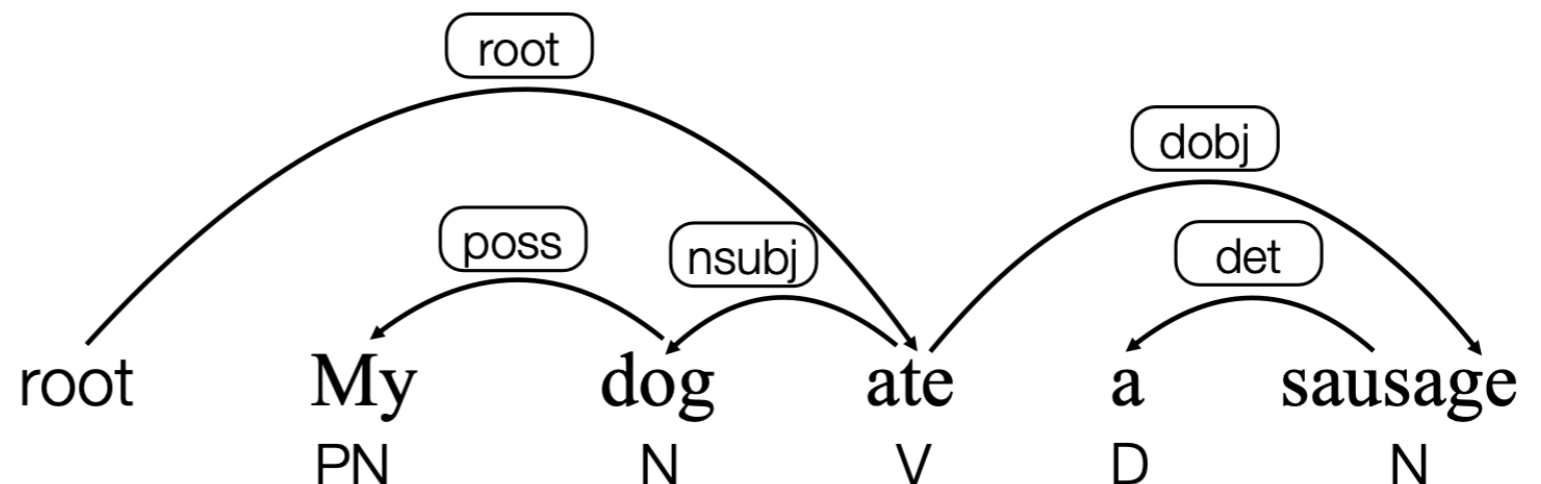
- Coordination:
 - Example: He went to and came from the store.
- Phonological reduction:
 - I will go → I'll go
 - I want to go → I wanna go
 - A le centre → au centre



La vitesse des ondes sismiques

Dependency Trees

- Nodes are words (along with part-of-speech tags)
- Directed arcs encode syntactic dependencies between words
- Labels are types of relations between words:
 - **root**: root of the tree, usually points to a verb
 - **poss**: possessive
 - **dobj**: direct object
 - **nsubj**: (noun) subject
 - **det**: determiner

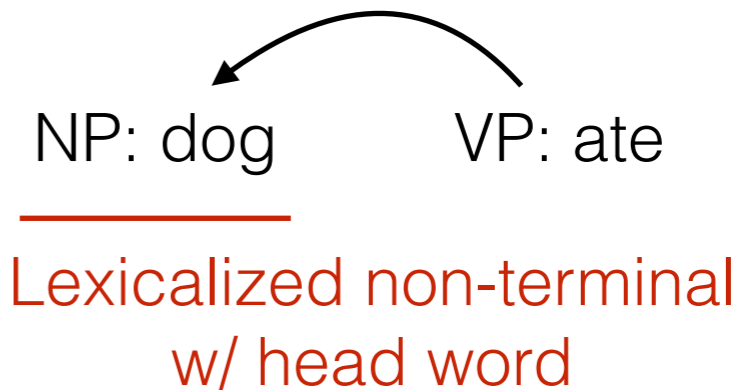
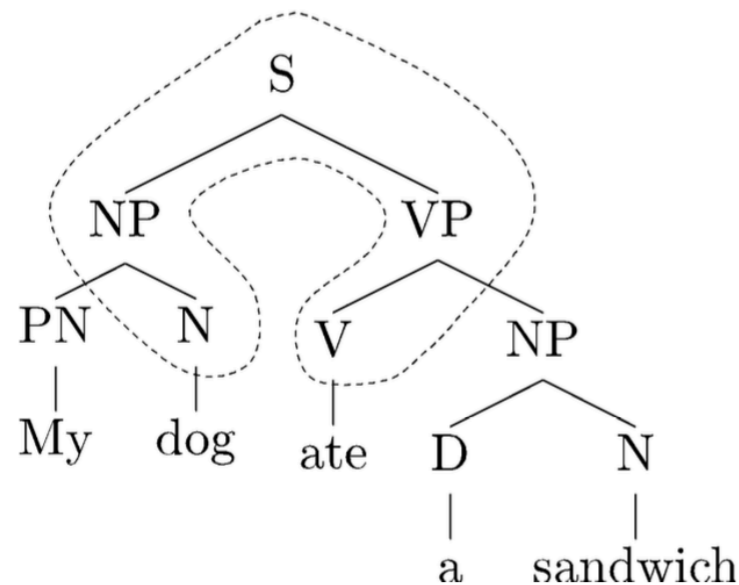


Dependency parsing

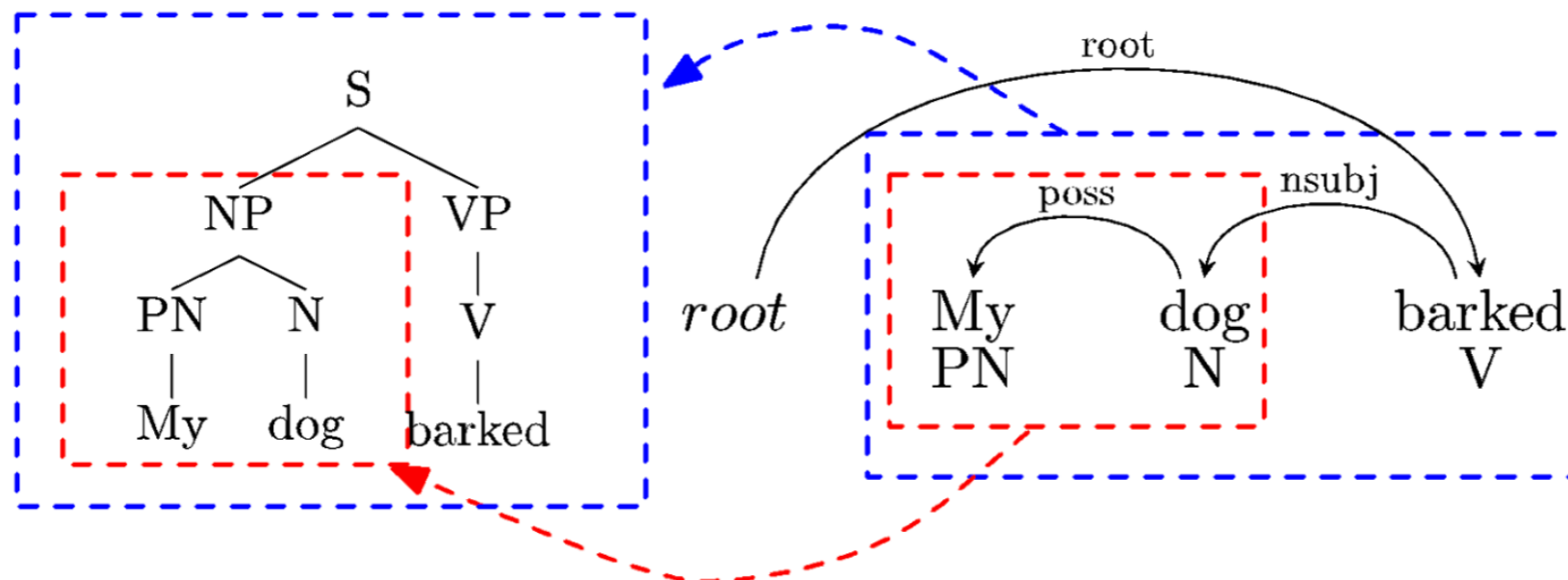
- Recover shallow semantics
- Shallow semantic information can be (approximately) derived from syntactic information
 - Subjects (nsubj) are often **agents**: *initiators / doers of an action*
 - Direct objects (dobj) are often **patients**: *affected entities*
- But not always true. Even for agents and patients, consider:
 - Mary is baking a cake in the oven
 - A cake is baking in the oven
- In general, it is not trivial even for the most shallow forms of semantics
 - e.g., prepositions: **in** can encode direction, position, temporal information, ...

Constituency \leftrightarrow Dependency

- Constituency trees can (potentially) \rightarrow dependency trees



- Dependency trees can (potentially) \rightarrow constituency trees



Context Free Grammar (CFG) & Probabilistic CFG

Context-free grammars (CFG)

- **Context-free grammars (CFG)**: a formalism for parsing.

Grammar (CFG)

ROOT \rightarrow S

S \rightarrow NP VP

NP \rightarrow DT NN

NP \rightarrow NN NNS

NP \rightarrow NP PP

VP \rightarrow VBP NP

VP \rightarrow VBP NP PP

PP \rightarrow IN NP

Lexicon

NN \rightarrow interest

NNS \rightarrow raises

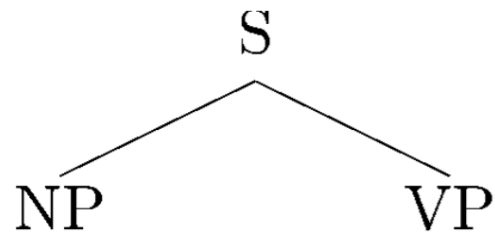
VBP \rightarrow interest

VBP \rightarrow raises

...

- Other **(non-CF)** grammar formalism: LFG, HPSG, TAG, CCG, ...

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

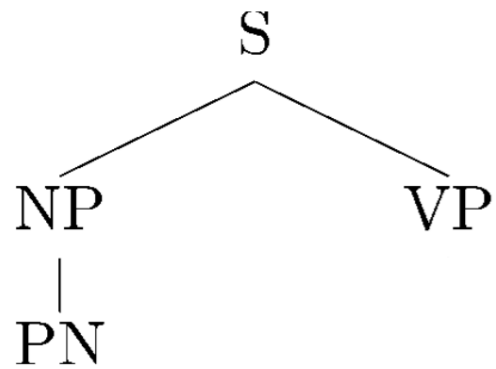
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow \text{a}$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

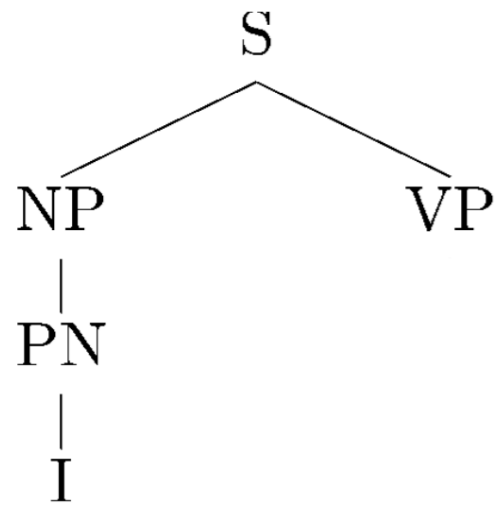
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow \text{a}$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

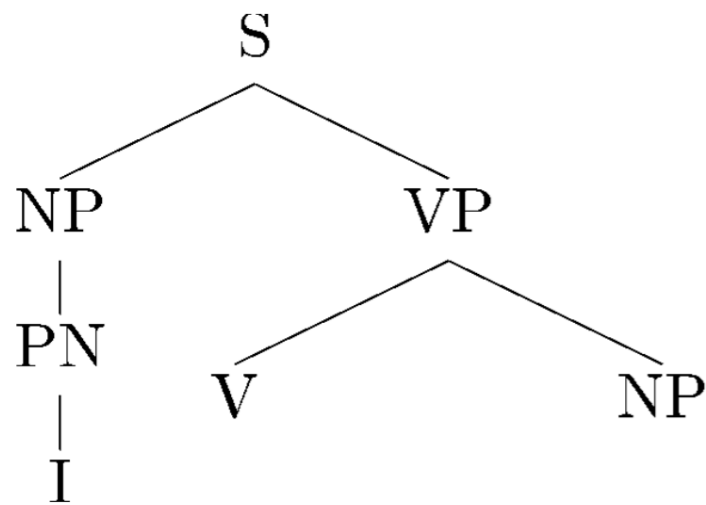
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow \text{a}$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

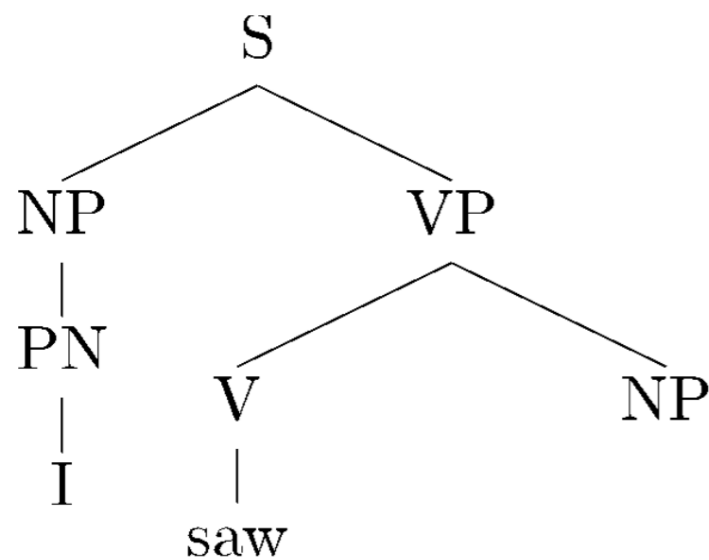
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow a$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

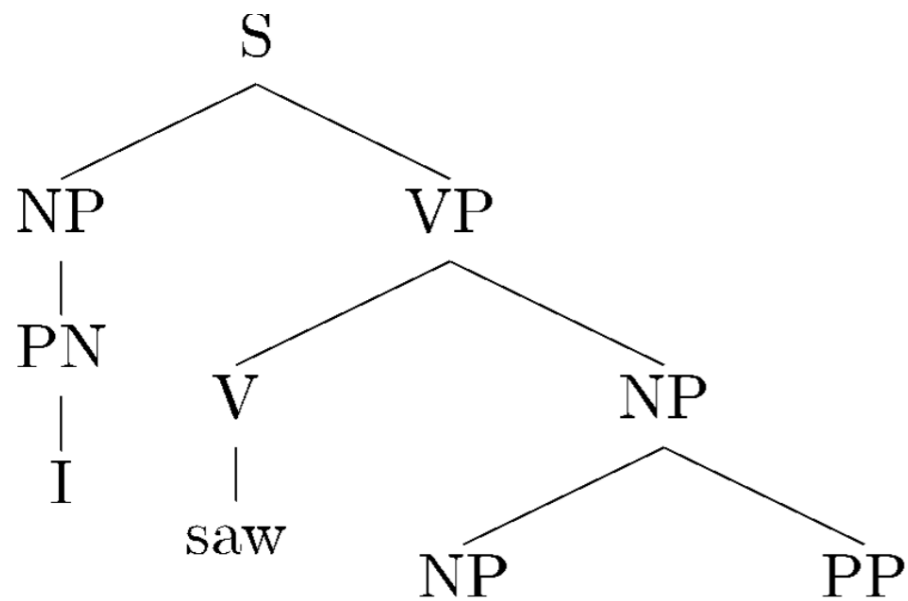
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow a$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

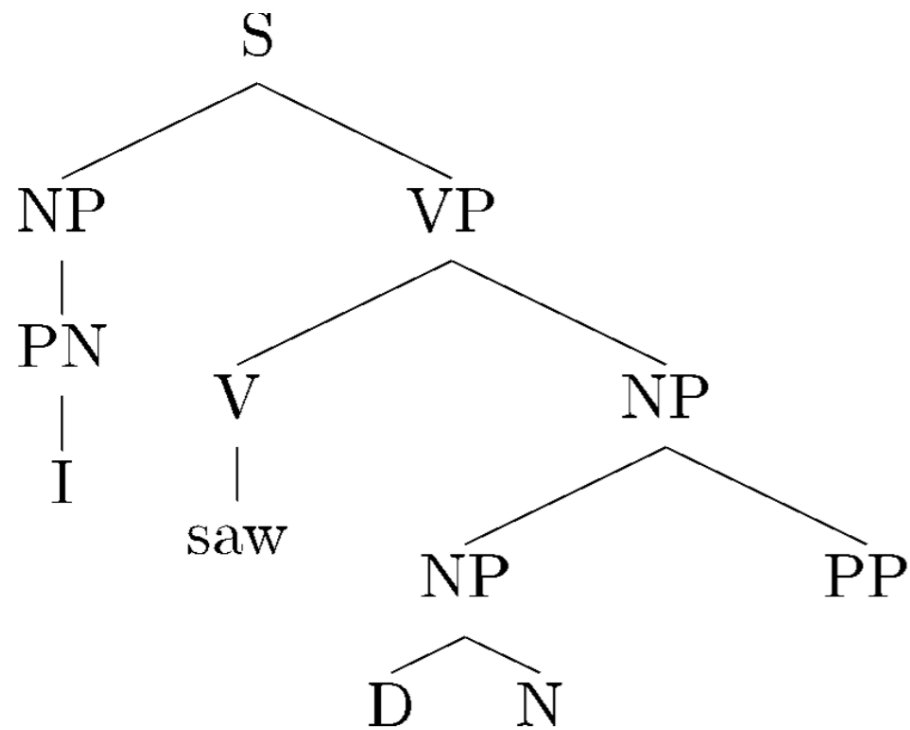
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow a$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

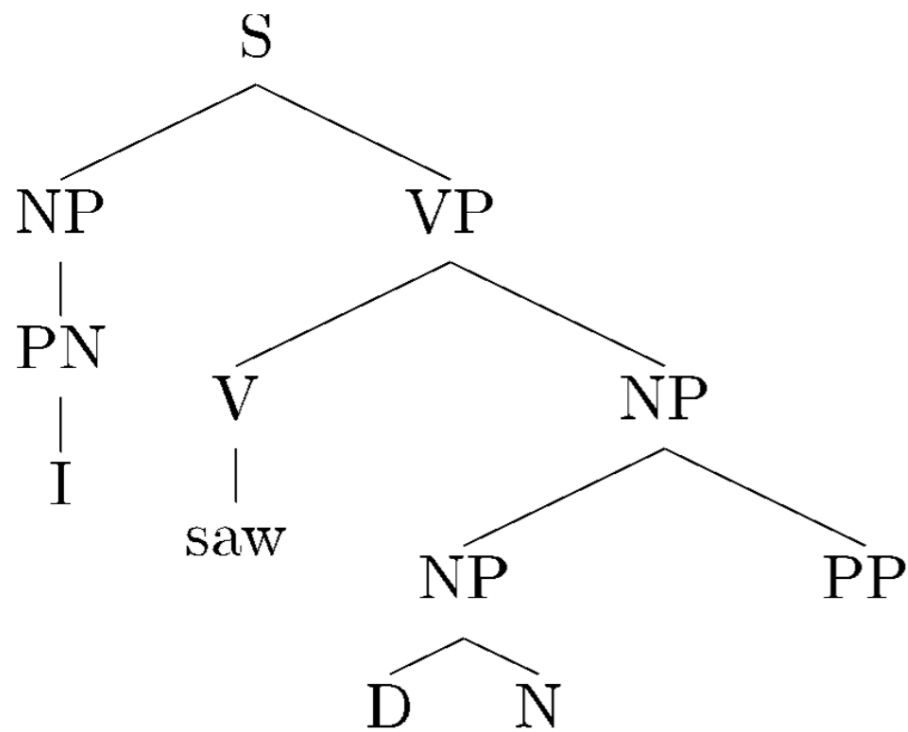
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow a$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

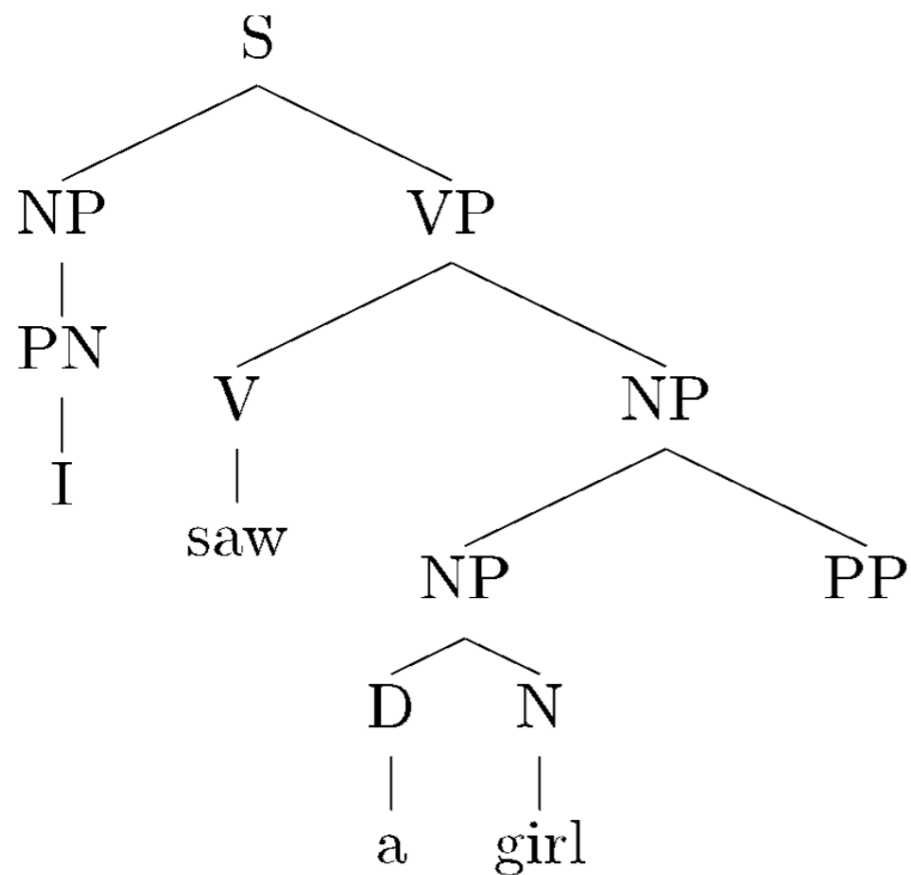
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow a$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

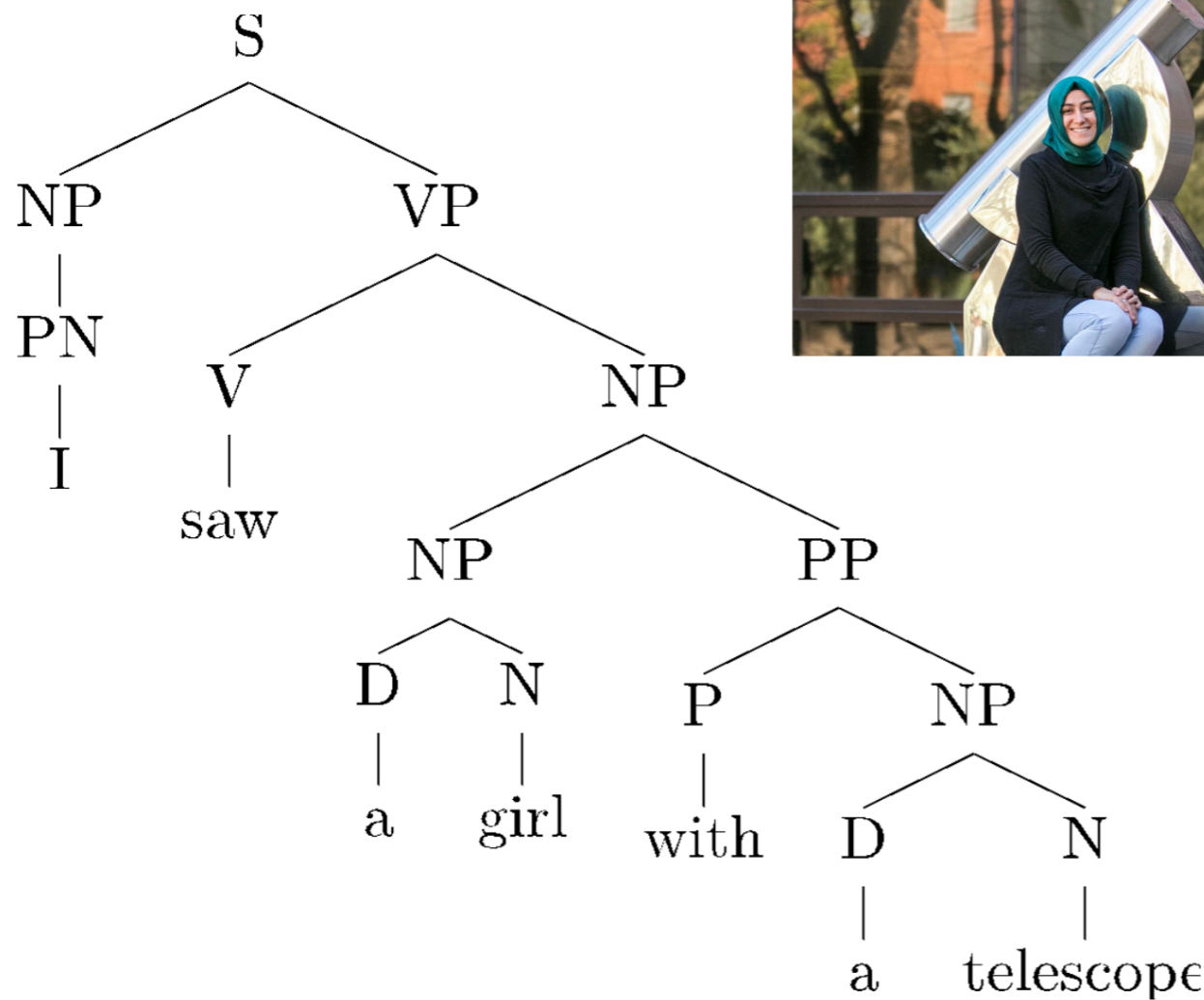
$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow a$

$D \rightarrow \text{the}$

CFG for Syntactic Parsing



Grammar (CFG)

$S \rightarrow NP VP$

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow VP PP$

$NP \rightarrow NP PP$

$NP \rightarrow D N$

$NP \rightarrow PN$

$PP \rightarrow P NP$

Lexicon

$N \rightarrow \text{girl}$

$N \rightarrow \text{telescope}$

$N \rightarrow \text{sandwich}$

$PN \rightarrow I$

$V \rightarrow \text{saw}$

$V \rightarrow \text{ate}$

$P \rightarrow \text{with}$

$P \rightarrow \text{in}$

$D \rightarrow a$

$D \rightarrow \text{the}$

Probabilistic context-free grammars (PCFG)

- **CFG**: A 4-tuple (N, Σ, R, S) :
 - N : a set of non-terminal symbols
 - Σ : a set of terminal symbols (disjoint from N)
 - S : a designated start symbol and a member of N
 - R : a set of rules, each of the form $A \rightarrow s$, where A is a non-terminal, s is a string of symbols, $A \in N, s \in (\Sigma \cup N)^*$

$$\begin{array}{l} S \rightarrow A, \\ A \rightarrow BC, \\ A \rightarrow \alpha, \end{array}$$

$$A \in N$$

$$A \in N, B, C \in N \cup \Sigma$$

$$\alpha \in \Sigma$$

Without loss of generality, only consider binary branching; Chomsky Normal Form

- **PCFG** adds a top-down production probability per rule.
 - Model the probability of each rule: $P(A \rightarrow s)$

$$\forall A \rightarrow s \in R : 0 \leq P(A \rightarrow s) \leq 1$$

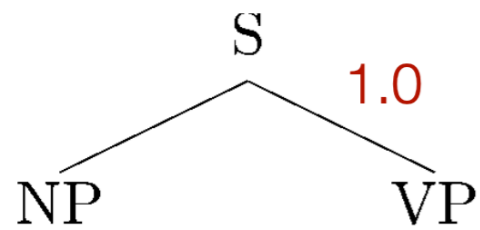
$$\forall A \in N : \sum_{s \text{ where } A \rightarrow s \in R} P(A \rightarrow s) = 1$$

PCFG (Example)

S → NP VP	1.0	(NP a girl) (VP ate a sandwich)	N → <i>girl</i>	0.2
VP → V	0.2		N → <i>telescope</i>	0.7
VP → V NP	0.4	(V ate) (NP a sandwich)	N → <i>sandwich</i>	0.1
VP → VP PP	0.4	(VP saw a girl) (PP with a telescope)	PN → <i>I</i>	1.0
NP → NP PP	0.3	(NP a girl) (PP with a sandwich)	V → <i>saw</i>	0.5
NP → D N	0.5	(D a) (N sandwich)	V → <i>ate</i>	0.5
NP → PN	0.2		P → <i>with</i>	0.6
PP → P NP	1.0	(P with) (NP a sandwich)	P → <i>in</i>	0.4
			D → <i>a</i>	0.3
			D → <i>the</i>	0.7

Now we can score a tree as a product of probabilities corresponding to the used rules!

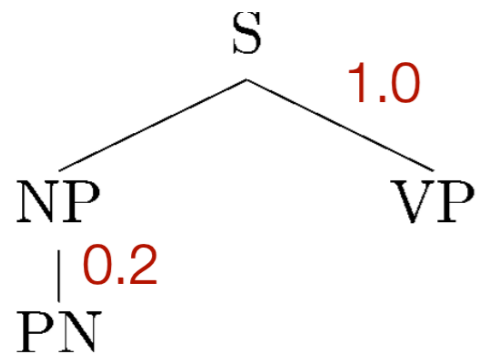
PCFG (Example)



$S \rightarrow NP VP$	1.0	$N \rightarrow girl$	0.2
		$N \rightarrow telescope$	0.7
$VP \rightarrow V$	0.2	$N \rightarrow sandwich$	0.1
$VP \rightarrow V NP$	0.4	$PN \rightarrow I$	1.0
$VP \rightarrow VP PP$	0.4	$V \rightarrow saw$	0.5
		$V \rightarrow ate$	0.5
$NP \rightarrow NP PP$	0.3	$P \rightarrow with$	0.6
$NP \rightarrow D N$	0.5	$P \rightarrow in$	0.4
$NP \rightarrow PN$	0.2	$D \rightarrow a$	0.3
		$D \rightarrow the$	0.7
$PP \rightarrow P NP$	1.0		

$P(T) = 1.0^*$

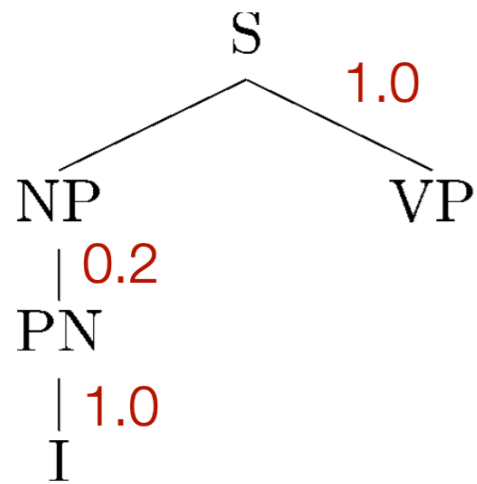
PCFG (Example)



S → NP VP	1.0	N → <i>girl</i>	0.2
		N → <i>telescope</i>	0.7
VP → V	0.2	N → <i>sandwich</i>	0.1
VP → V NP	0.4	PN → <i>I</i>	1.0
VP → VP PP	0.4	V → <i>saw</i>	0.5
		V → <i>ate</i>	0.5
NP → NP PP	0.3	P → <i>with</i>	0.6
NP → D N	0.5	P → <i>in</i>	0.4
NP → PN	0.2	D → <i>a</i>	0.3
		D → <i>the</i>	0.7
PP → P NP	1.0		

$$P(T) = 1.0 * 0.2 *$$

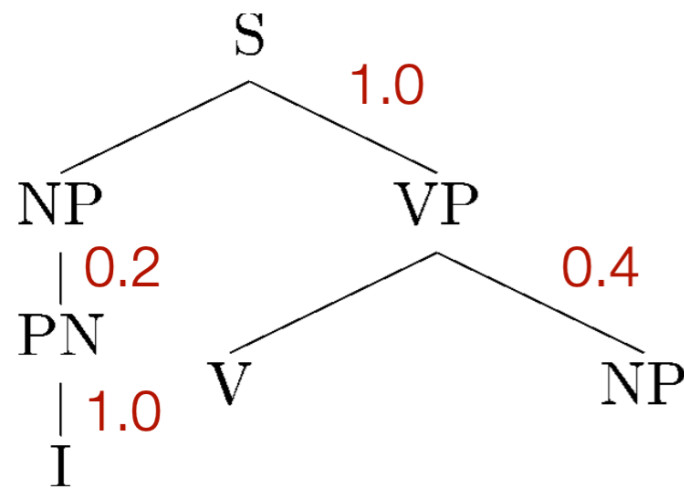
PCFG (Example)



$$P(T) = 1.0 * 0.2 * 1.0 *$$

$S \rightarrow NP VP$	1.0	$N \rightarrow girl$	0.2
		$N \rightarrow telescope$	0.7
$VP \rightarrow V$	0.2	$N \rightarrow sandwich$	0.1
$VP \rightarrow V NP$	0.4	$PN \rightarrow I$	1.0
$VP \rightarrow VP PP$	0.4	$V \rightarrow saw$	0.5
		$V \rightarrow ate$	0.5
$NP \rightarrow NP PP$	0.3	$P \rightarrow with$	0.6
$NP \rightarrow D N$	0.5	$P \rightarrow in$	0.4
$NP \rightarrow PN$	0.2	$D \rightarrow a$	0.3
		$D \rightarrow the$	0.7
$PP \rightarrow P NP$	1.0		

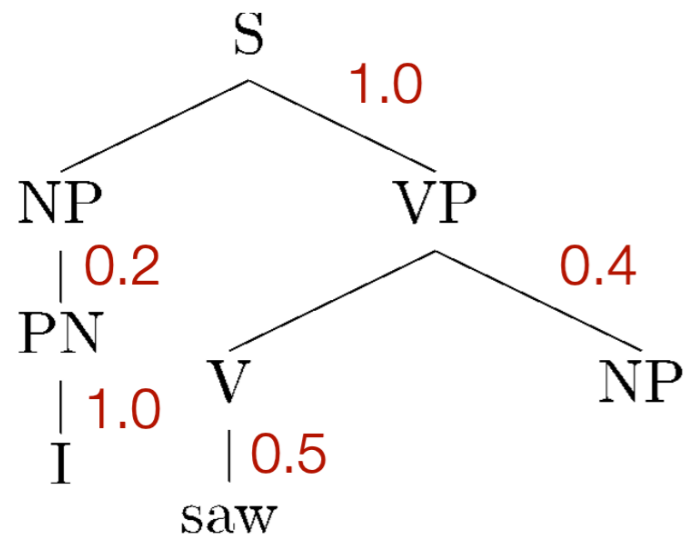
PCFG (Example)



$S \rightarrow NP VP$	1.0	$N \rightarrow girl$	0.2
		$N \rightarrow telescope$	0.7
$VP \rightarrow V$	0.2	$N \rightarrow sandwich$	0.1
$VP \rightarrow V NP$	0.4	$PN \rightarrow I$	1.0
$VP \rightarrow VP PP$	0.4	$V \rightarrow saw$	0.5
		$V \rightarrow ate$	0.5
$NP \rightarrow NP PP$	0.3	$P \rightarrow with$	0.6
$NP \rightarrow D N$	0.5	$P \rightarrow in$	0.4
$NP \rightarrow PN$	0.2	$D \rightarrow a$	0.3
		$D \rightarrow the$	0.7
$PP \rightarrow P NP$	1.0		

$$P(T) = 1.0 * 0.2 * 1.0 * 0.4 *$$

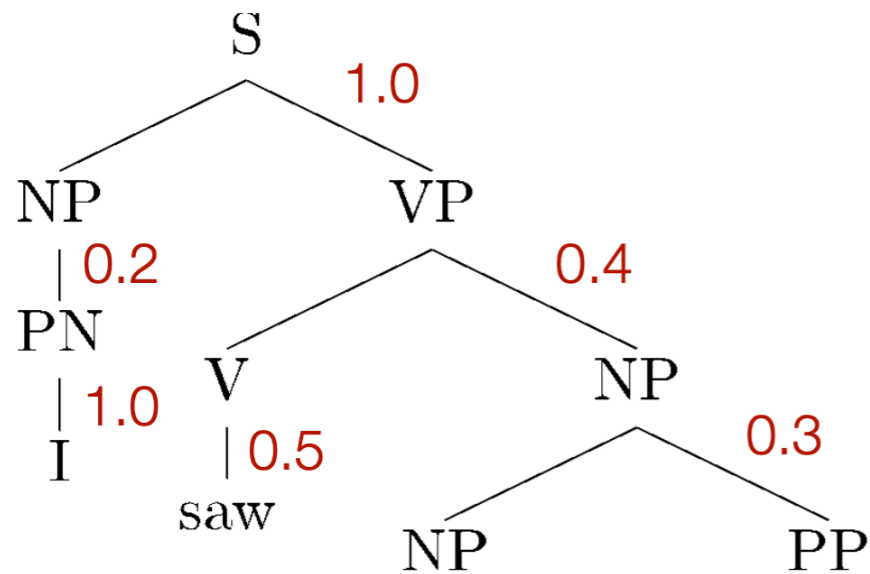
PCFG (Example)



S → NP VP	1.0	N → <i>girl</i>	0.2
		N → <i>telescope</i>	0.7
VP → V	0.2	N → <i>sandwich</i>	0.1
VP → V NP	0.4	PN → <i>I</i>	1.0
VP → VP PP	0.4	V → <i>saw</i>	0.5
		V → <i>ate</i>	0.5
NP → NP PP	0.3	P → <i>with</i>	0.6
NP → D N	0.5	P → <i>in</i>	0.4
NP → PN	0.2	D → <i>a</i>	0.3
		D → <i>the</i>	0.7
PP → P NP	1.0		

$$P(T) = 1.0 * 0.2 * 1.0 * 0.4 * 0.5 *$$

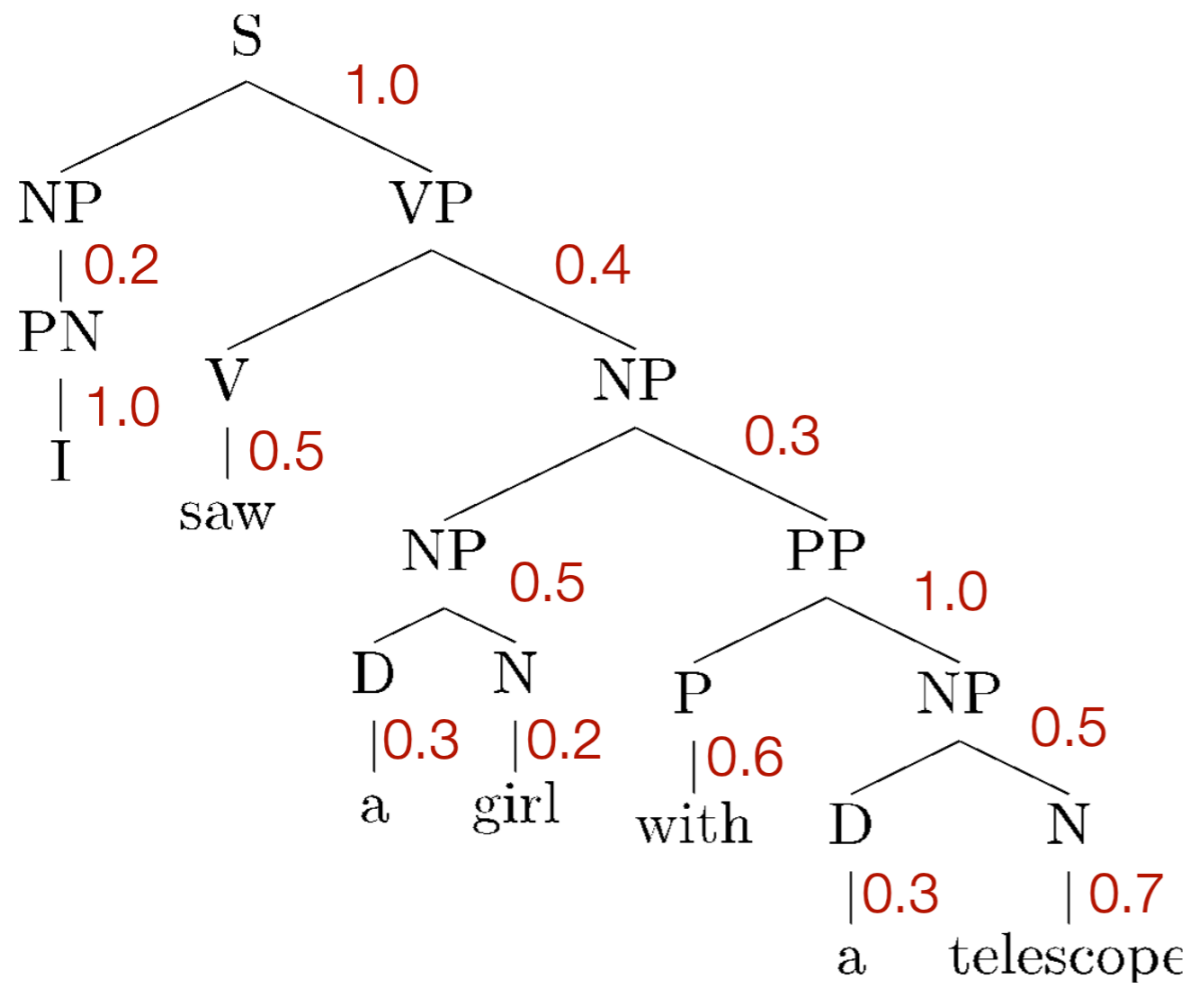
PCFG (Example)



S → NP VP	1.0	N → <i>girl</i>	0.2
		N → <i>telescope</i>	0.7
VP → V	0.2	N → <i>sandwich</i>	0.1
VP → V NP	0.4	PN → <i>I</i>	1.0
VP → VP PP	0.4	V → <i>saw</i>	0.5
		V → <i>ate</i>	0.5
NP → NP PP	0.3	P → <i>with</i>	0.6
NP → D N	0.5	P → <i>in</i>	0.4
NP → PN	0.2	D → <i>a</i>	0.3
		D → <i>the</i>	0.7
PP → P NP	1.0		

$$P(T) = 1.0 * 0.2 * 1.0 * 0.4 * 0.5 * 0.3 *$$

PCFG (Example)



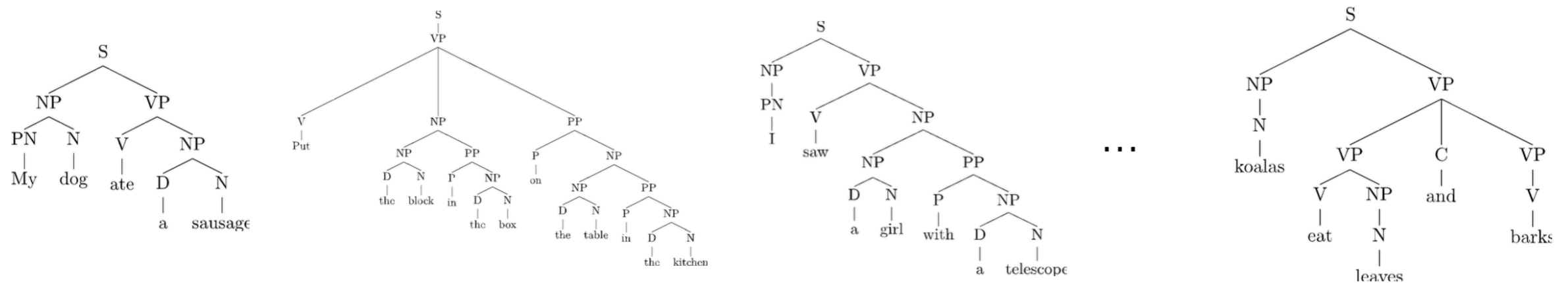
S → NP VP	1.0	N → <i>girl</i>	0.2
		N → <i>telescope</i>	0.7
VP → V	0.2	N → <i>sandwich</i>	0.1
VP → V NP	0.4	PN → <i>I</i>	1.0
VP → VP PP	0.4	V → <i>saw</i>	0.5
		V → <i>ate</i>	0.5
NP → NP PP	0.3	P → <i>with</i>	0.6
NP → D N	0.5	P → <i>in</i>	0.4
NP → PN	0.2	D → <i>a</i>	0.3
		D → <i>the</i>	0.7
PP → P NP	1.0		

$$P(T) = 1.0 * 0.2 * 1.0 * 0.4 * 0.5 * 0.3 * 0.5 * 0.3 * 0.2 * 1.0 * 0.6 * 0.5 * 0.3 * 0.7 = 2.26e-5$$

PCFG Supervised Learning & Decoding

PCFG Supervised Learning

- A treebank: a collection of sentences annotated with constituency trees
 - Penn Treebank: (X, T) pairs



- PCFG: a generative model, maximizing the joint probability of a sentence given a tree.
 - If we constraint the search space to be all valid trees that can generate the sentence, this becomes:

$$\max P(X, T) = \max P(X|T)P(T) = \max_{T \in \text{GEN}(X)} P(T)$$

PCFG Supervised Learning

- Estimate probability of each rule by maximum likelihood estimation:

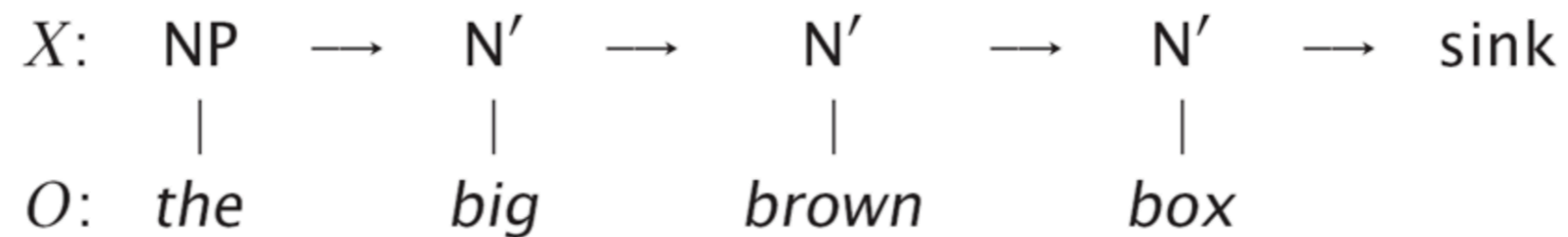
$$P(T) = \sum_{A \rightarrow s \in R} P(A \rightarrow s), \quad T \in \text{GEN}(X)$$

$$P(A \rightarrow s) = \frac{\text{Count}(A \rightarrow s)}{\text{Count}(A)} \quad \begin{array}{l} \# \text{ times the rule was used in the data} \\ \# \text{ times the nonterminal was used in the data} \end{array}$$

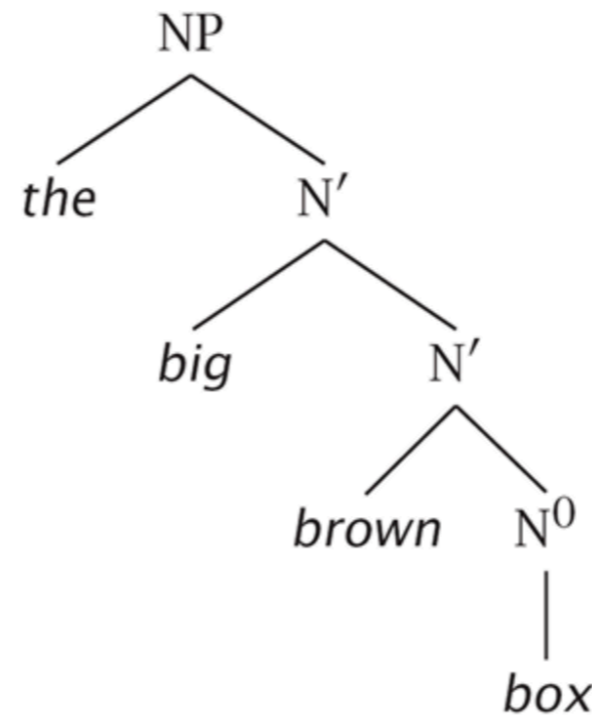
- Smoothing is helpful (esp. for rules that produce one word)
- If we don't have training data, use EM algorithm to estimate the probability

HMM vs PCFG

HMM: Linear Markov Chain

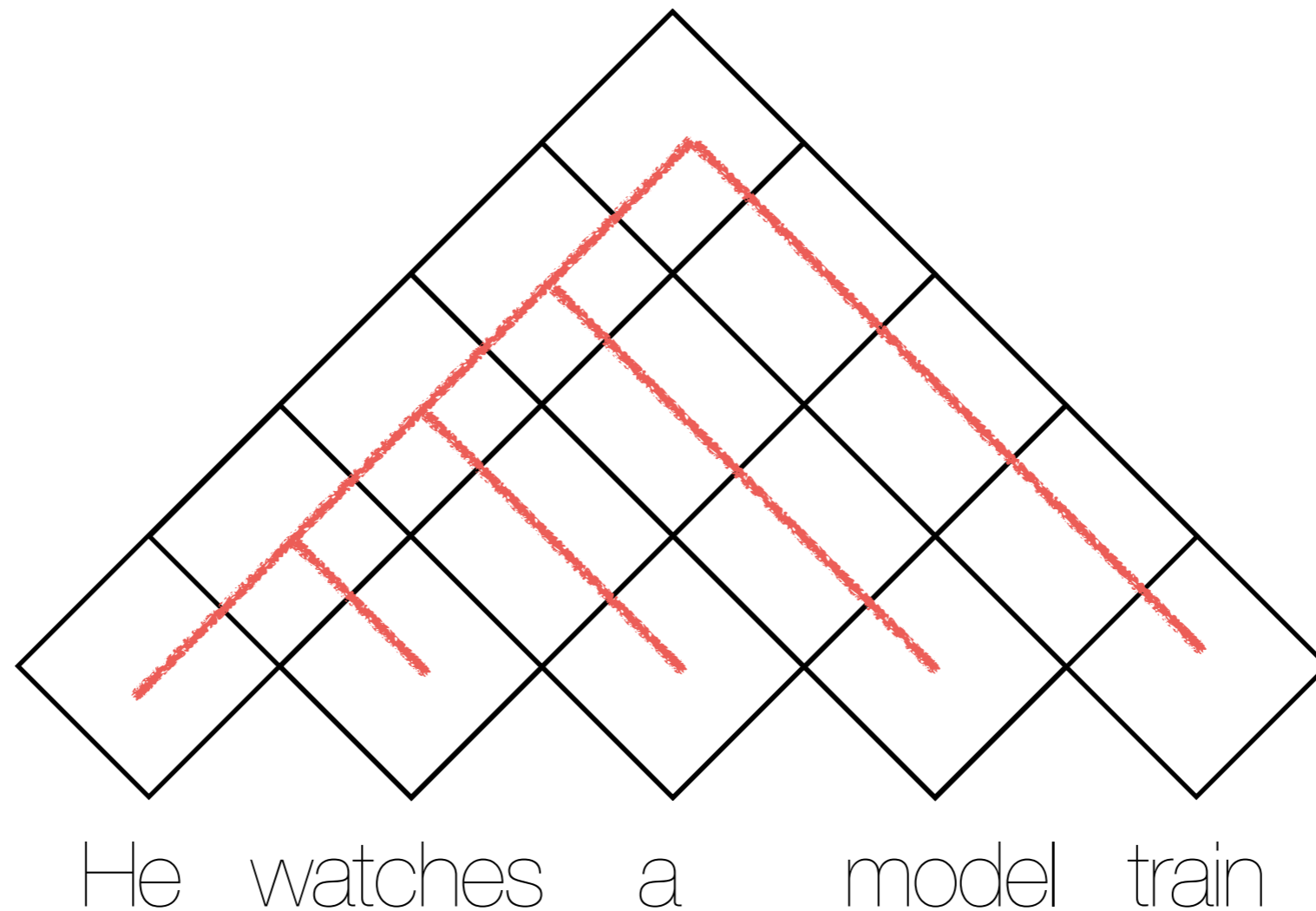


PCFG: tree



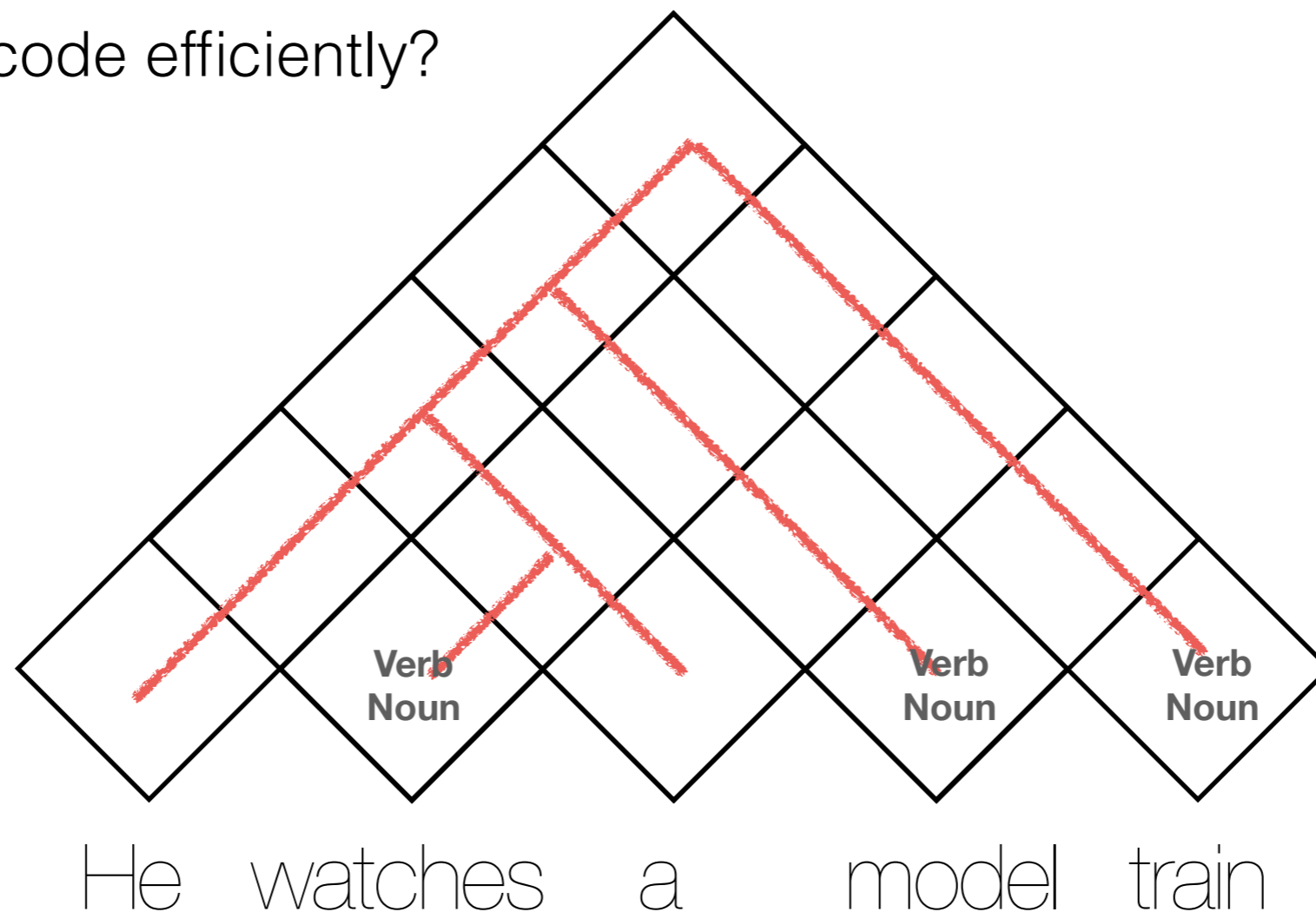
PCFG Decoding

- Brute force solution: enumerate all possible binary trees, score them, find the tree with maximum score



PCFG Decoding

- Brute force solution: enumerate all possible binary trees, score them, find the tree with maximum score
- For a sentence of n words, there are $(n-1)!$ possible binary trees. Each word may have more than 1 possible POS tags
- How to decode efficiently?



PCFG Decoding: CYK Algorithm

Bottom-up Dynamic Programming

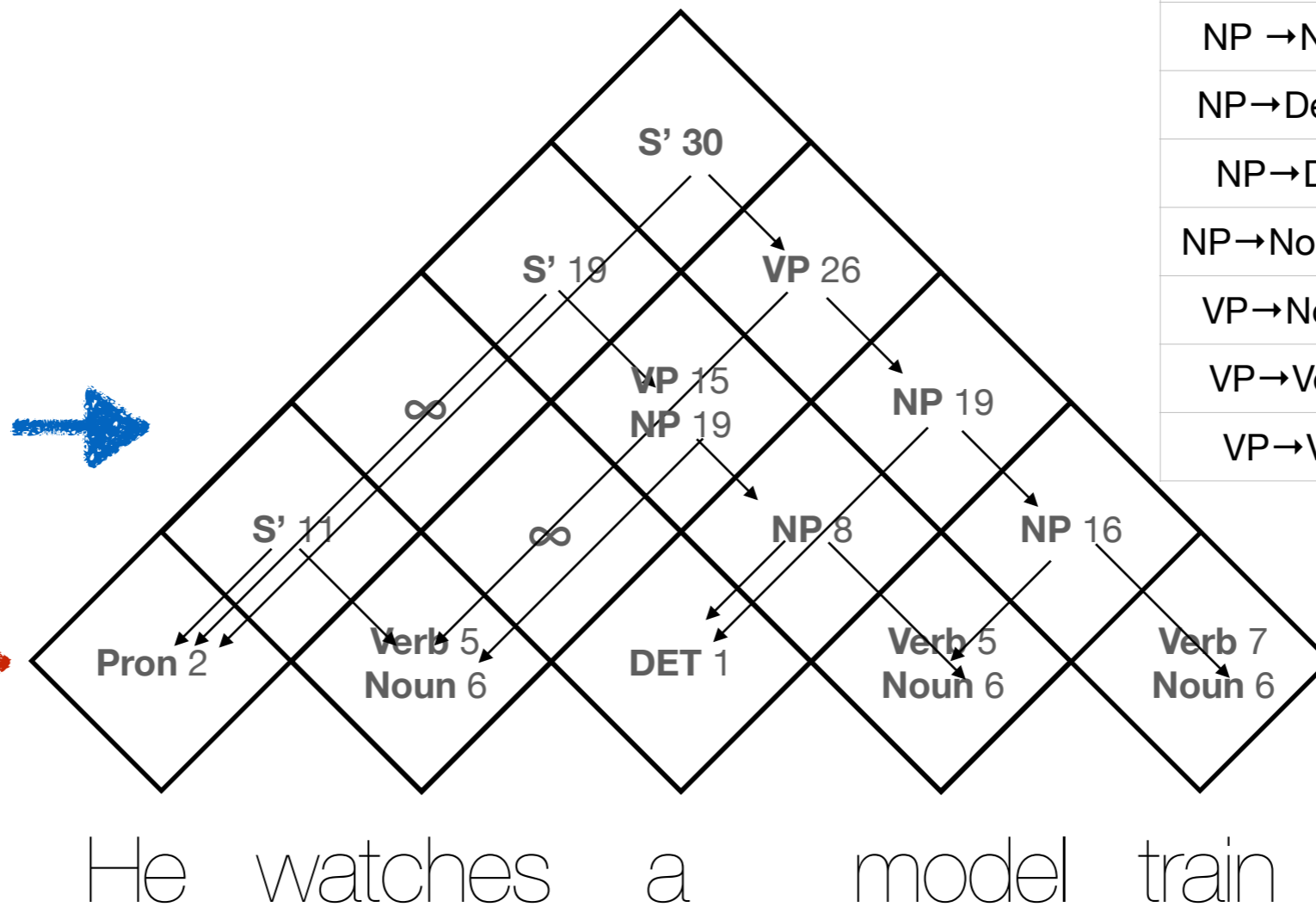
Remember to store back-pointer!

Binary Rule	-log prob
$S' \rightarrow \text{Pron Verb}$	4
$S' \rightarrow \text{Pron VP}$	2
$S' \rightarrow \text{NP VP}$	2
$\text{NP} \rightarrow \text{NP Verb}$	5
$\text{NP} \rightarrow \text{Det Noun}$	2
$\text{NP} \rightarrow \text{Det NP}$	2
$\text{NP} \rightarrow \text{Noun Noun}$	4
$\text{VP} \rightarrow \text{Noun NP}$	5
$\text{VP} \rightarrow \text{Verb NP}$	2
$\text{VP} \rightarrow \text{VP NP}$	2

Binary rule
 $A \rightarrow BC$

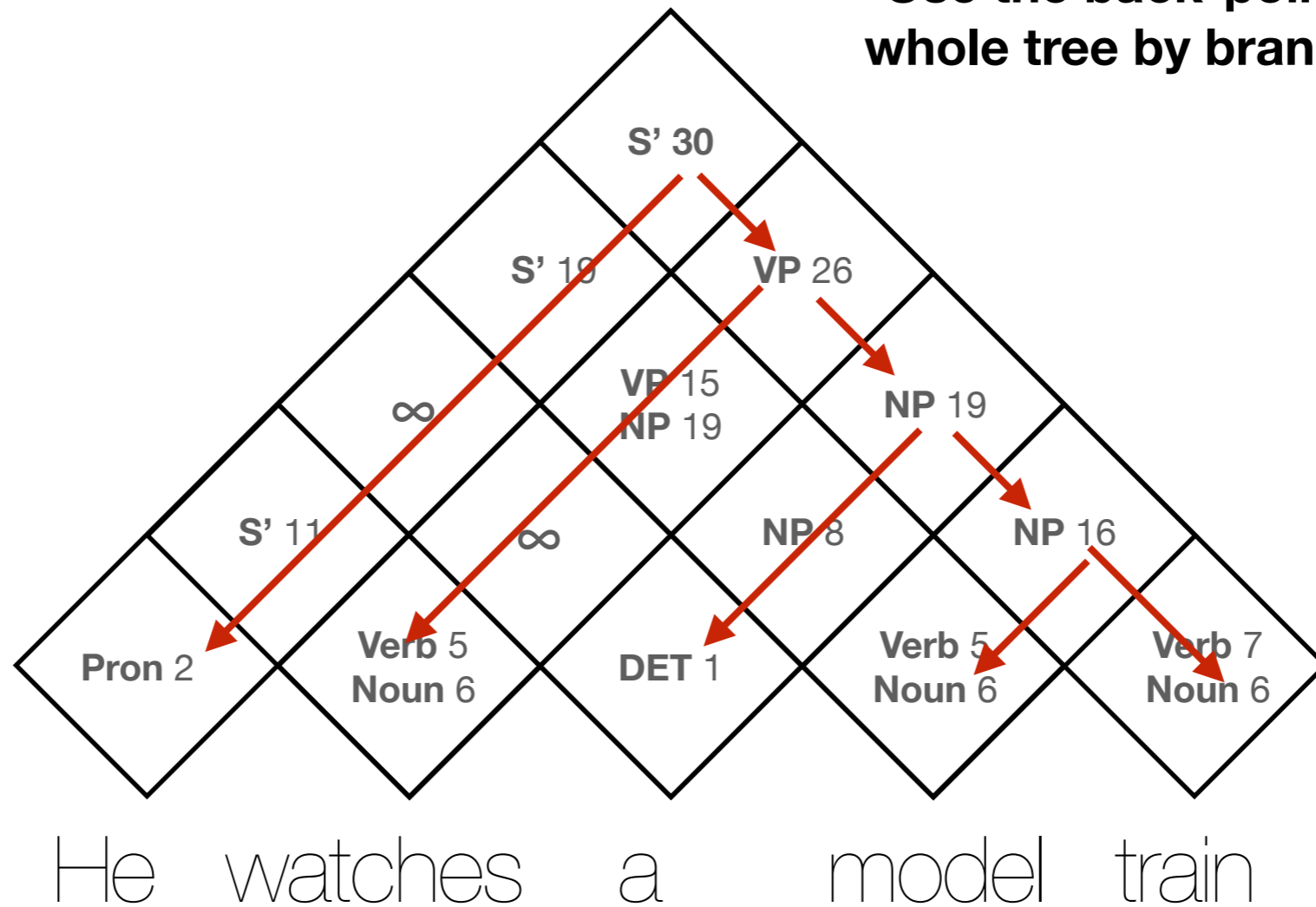
Lexical rule
(Unary rule)

$A \rightarrow \alpha, \alpha \in \Sigma$



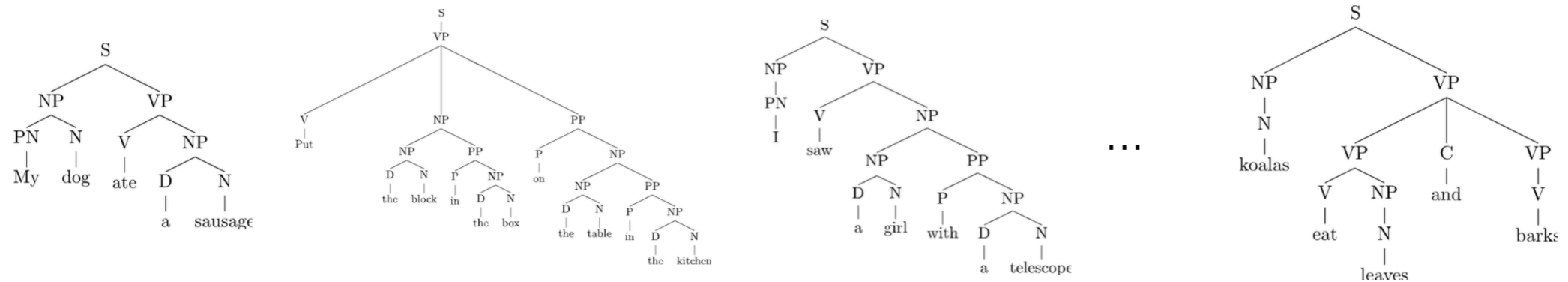
PCFG Decoding: CYK Algorithm

Use the back-pointer to build the whole tree by branching from root.



PCFG CYK Decoding

- A treebank: a collection of sentences annotated with constituency trees
 - Penn Treebank



- Estimate probability of each rule by maximum likelihood estimation:

$$P(A \rightarrow s) = \frac{\text{Count}(A \rightarrow s)}{\text{Count}(A)}$$

times the rule was used in the data
times the nonterminal was used in the data

- Smoothing is helpful (esp. for rules that produce one word)

PCFG Decoding: CYK Algorithm

- The score (or unnormalized probability) of a constituent with a non-terminal is often called inside probability
 - Computed by dynamic programming

$$s_{\text{label}}(i, j, A) = \max_{k, B, C} P(A \rightarrow BC) \times s_{\text{label}}(i, k, B) \times s_{\text{label}}(k + 1, j, C)$$

Semiring Conversion

- The score (or unnormalized probability) of a constituent with a non-terminal is often called inside probability

- Computed by dynamic programming
- Numerically unstable

$$s_{\text{label}}(i, j, A) = \max_{k, B, C} P(A \rightarrow BC) \times s_{\text{label}}(i, k, B) \times s_{\text{label}}(k + 1, j, C)$$

- Define the minimum cost score, and rewrite the scores

$$s'_{\text{label}}(i, j, A) = -\log s_{\text{label}}(i, j, A)$$

$$s'_{\text{label}}(i, j, A) = \min_{k, B, C} (-\log P(A \rightarrow BC) + s'_{\text{label}}(i, k, B) + s'_{\text{label}}(k + 1, j, C))$$

Semiring Parsing

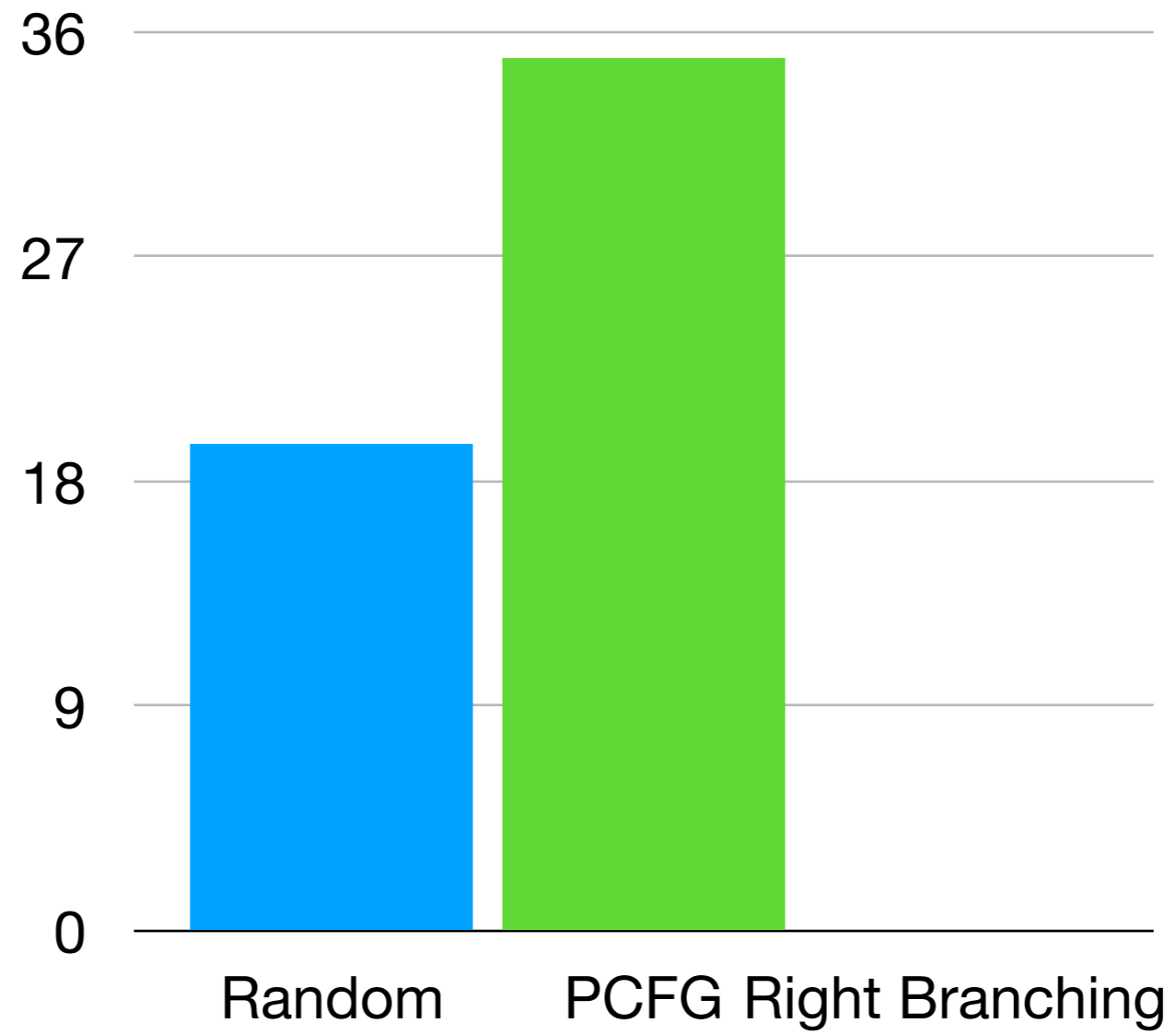
“Add” \oplus

“Multiply” \otimes

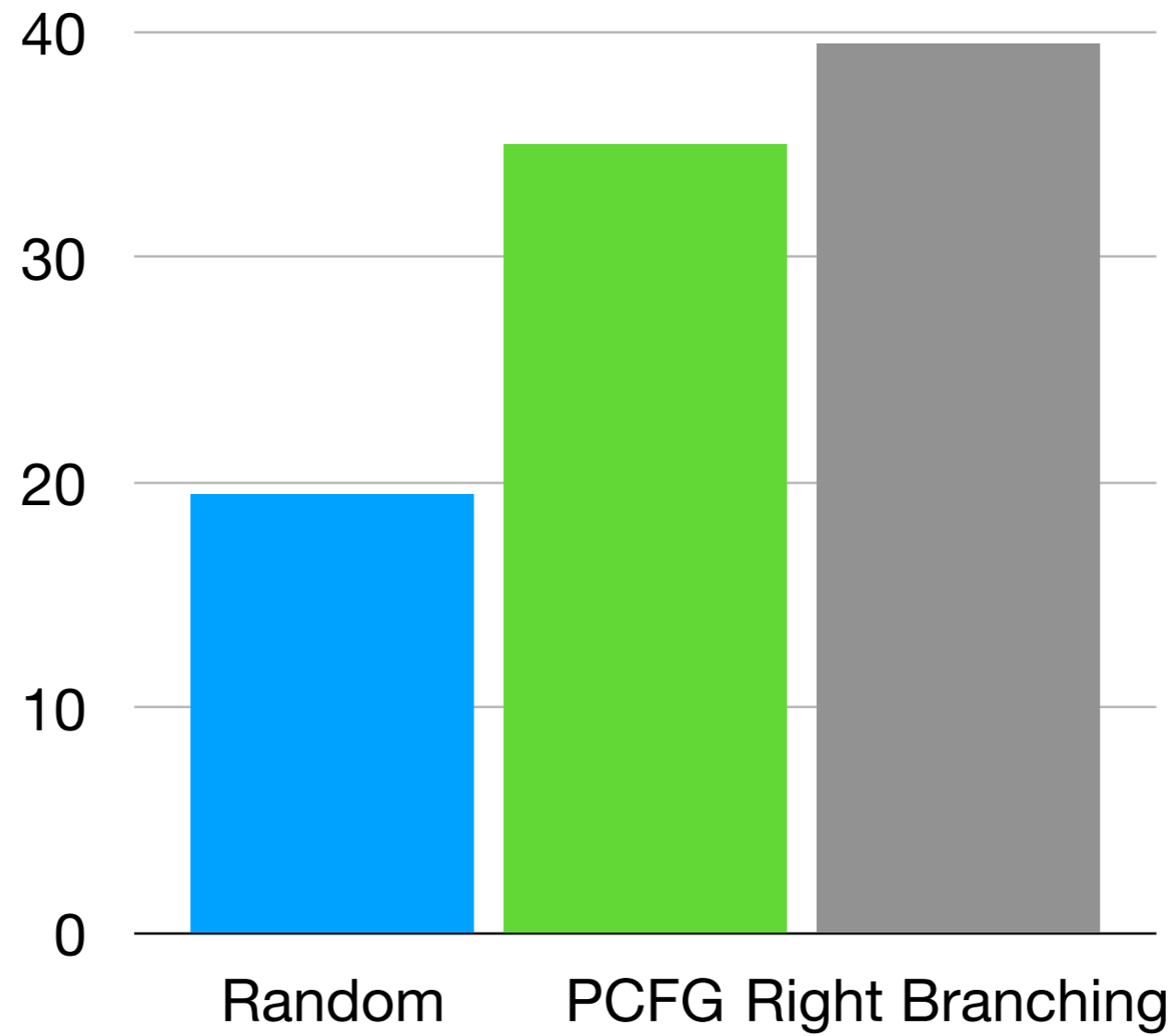
$$s_{\text{label}}(i, j, A) = \max_{k, B, C} P(A \rightarrow BC) \times s_{\text{label}}(i, k, B) \times s_{\text{label}}(k + 1, j, C)$$

$$s'_{\text{label}}(i, j, A) = \min_{k, B, C} (-\log P(A \rightarrow BC) + s'_{\text{label}}(i, k, B) + s'_{\text{label}}(k + 1, j, C))$$

	weights	\oplus	\otimes	$\mathbf{0}$	$\mathbf{1}$
total prob	[0, 1]	+	x	0	1
max prob	[0, 1]	max	x	0	1
min -logp	[0, ∞]	min	+	∞	0
log prob	$[-\infty, 0]$	logsumexp	+	$-\infty$	0
recognizer	T/F	or	and	F	T



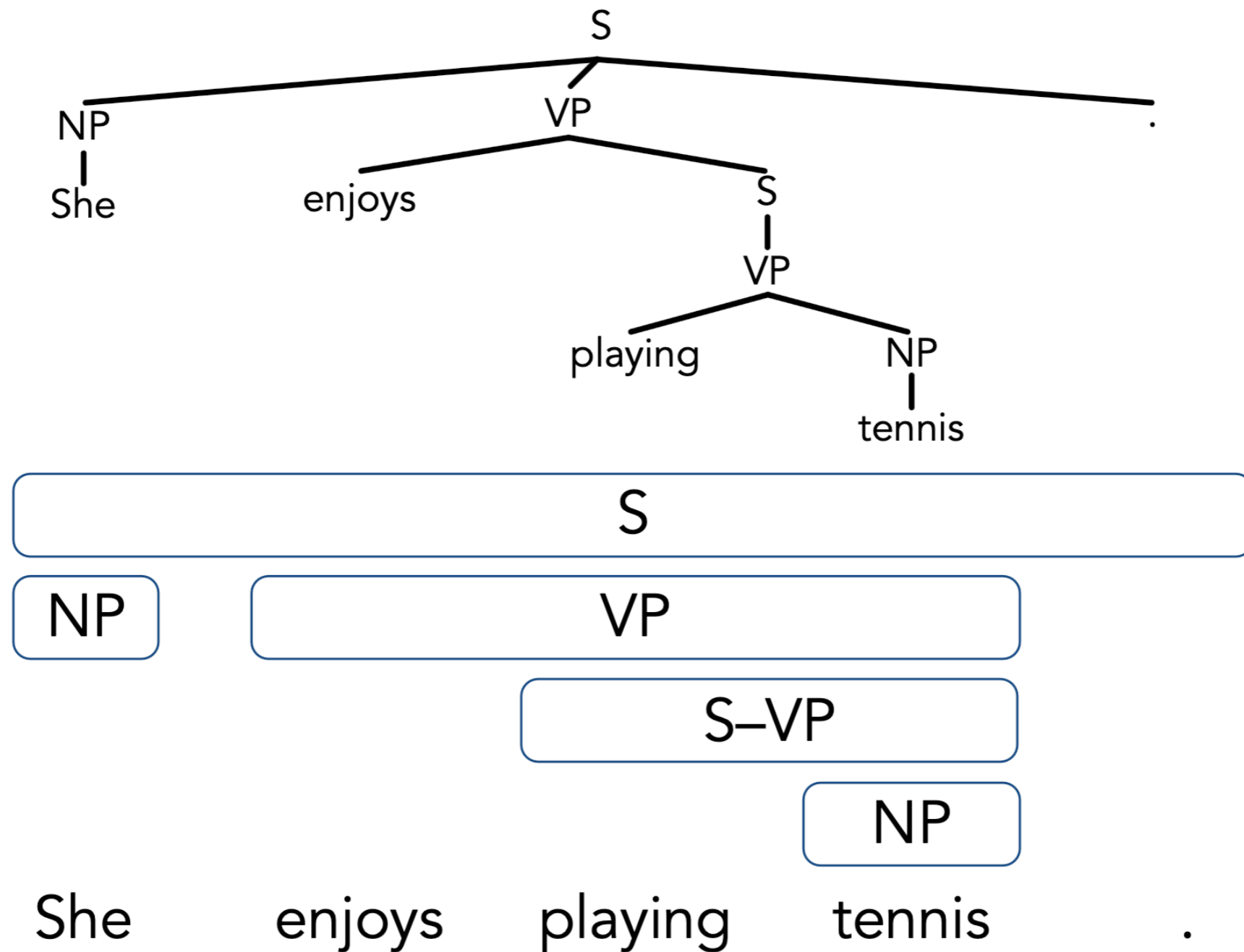
Great News: It works (better than random)



Bad News: It is even worse than right branching. Why?

Supervised Parsing: Span-based Neural Models

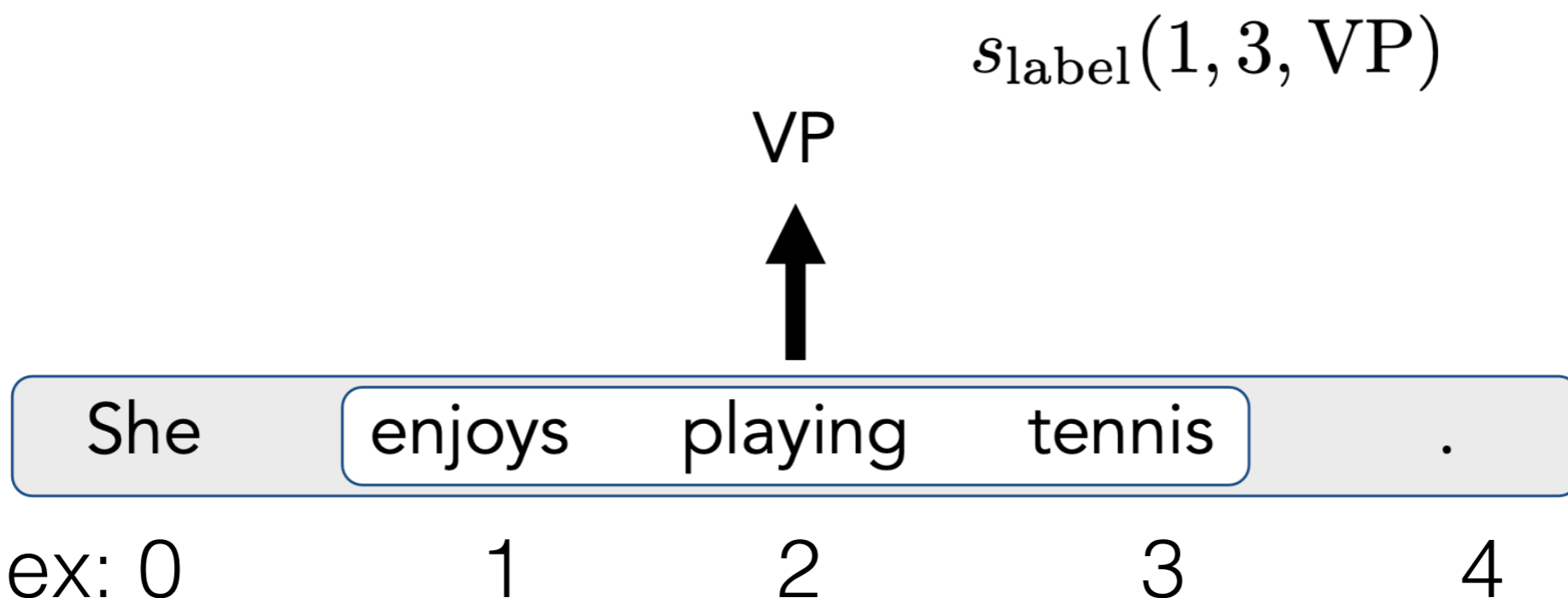
Span-Based Parsing



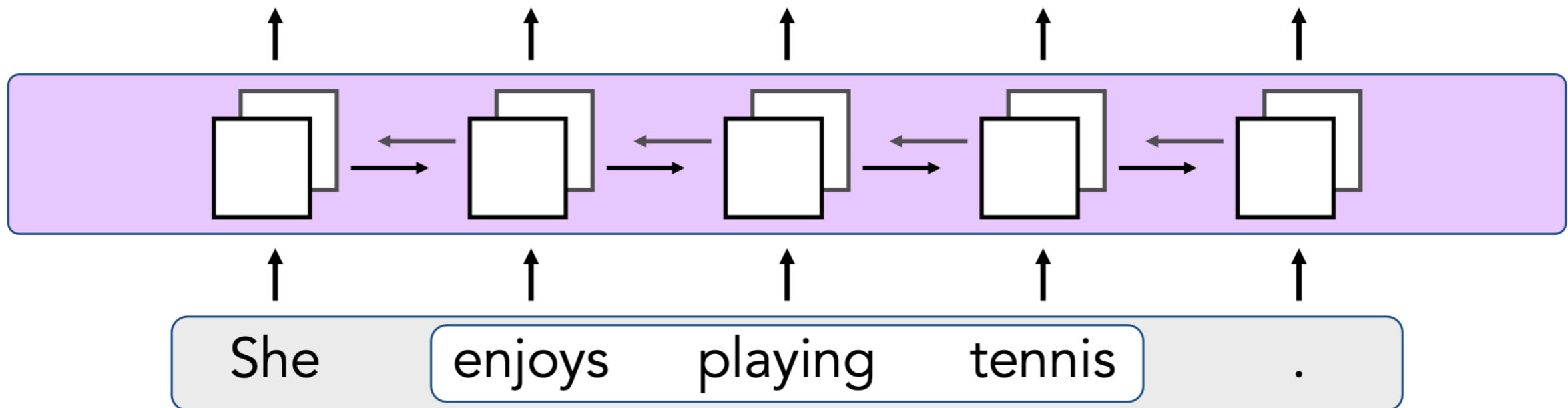
Span-Based Parsing

$$s_{\text{label}}(i, j, \ell)$$

Scoring a span from the i -th word to j -th word being the label of ℓ

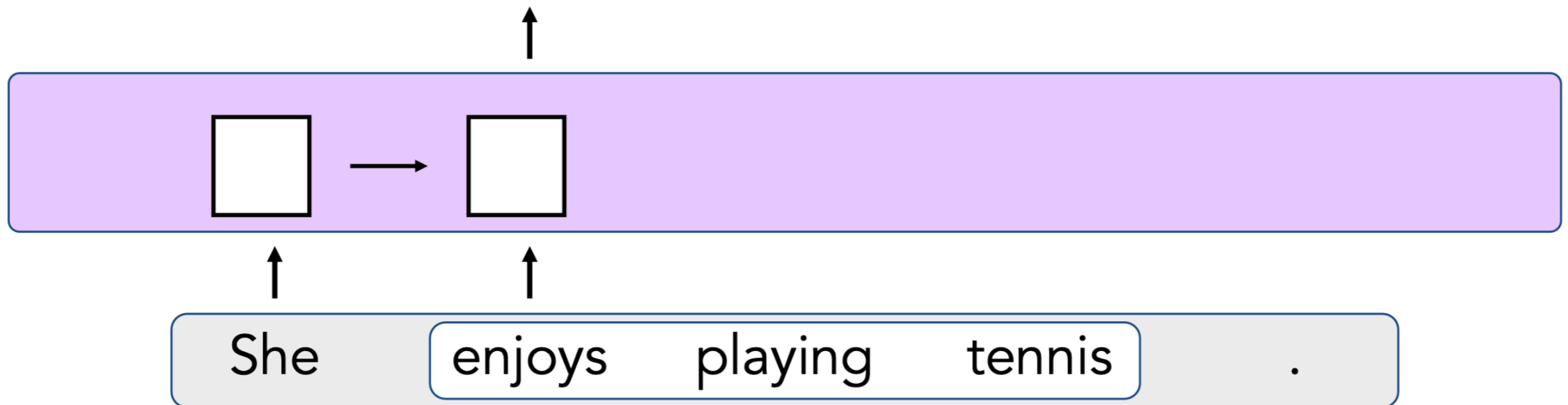


Span-Based Parsing

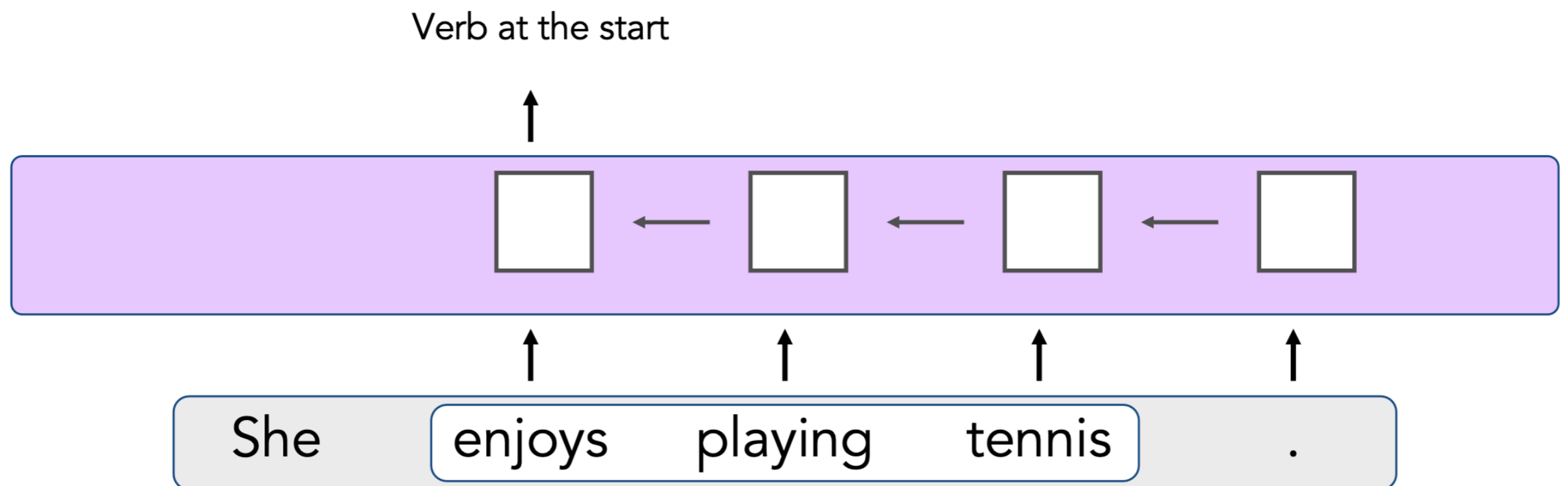


Span-Based Parsing

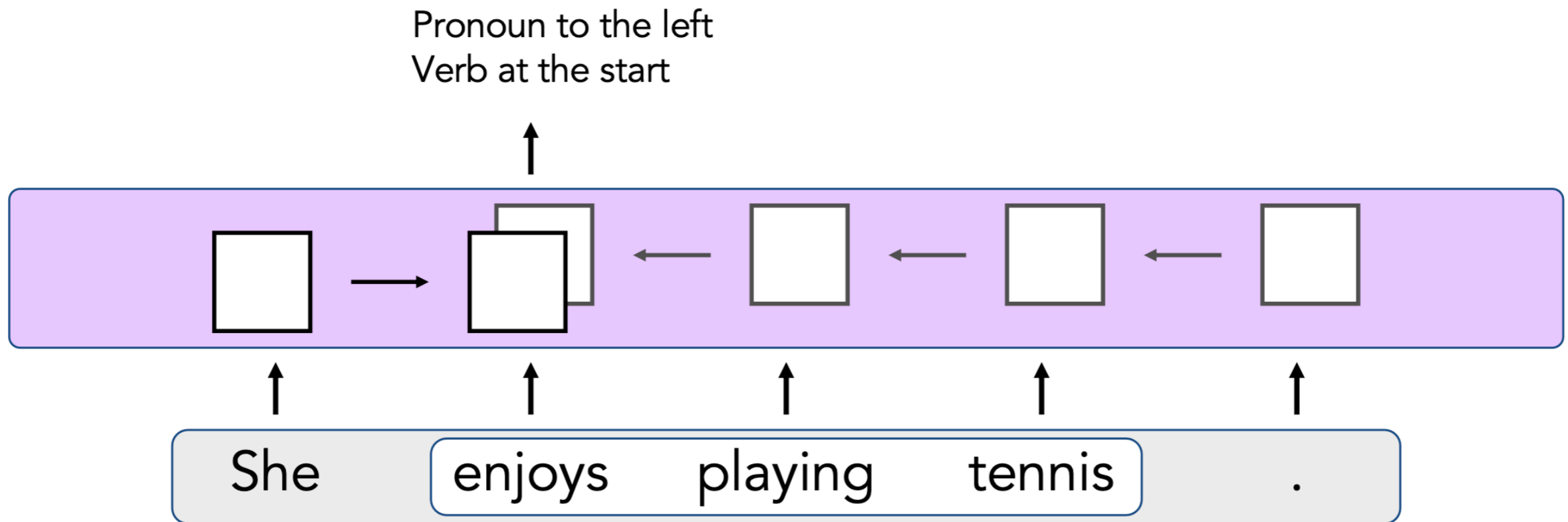
Pronoun to the left



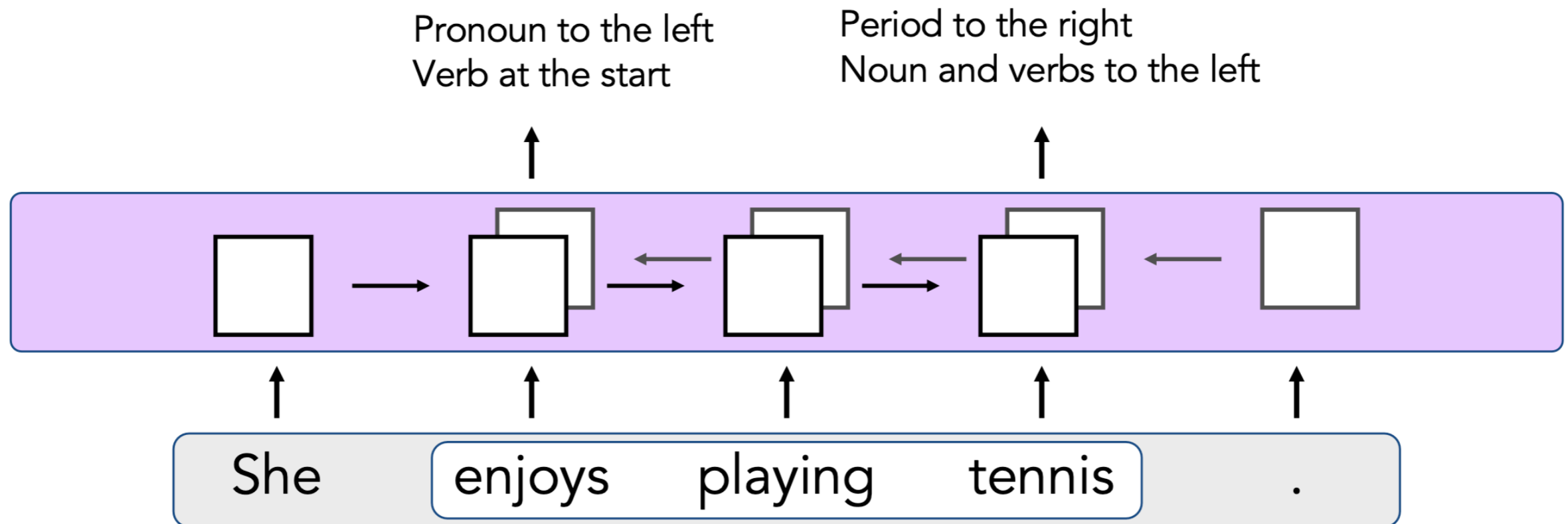
Span-Based Parsing



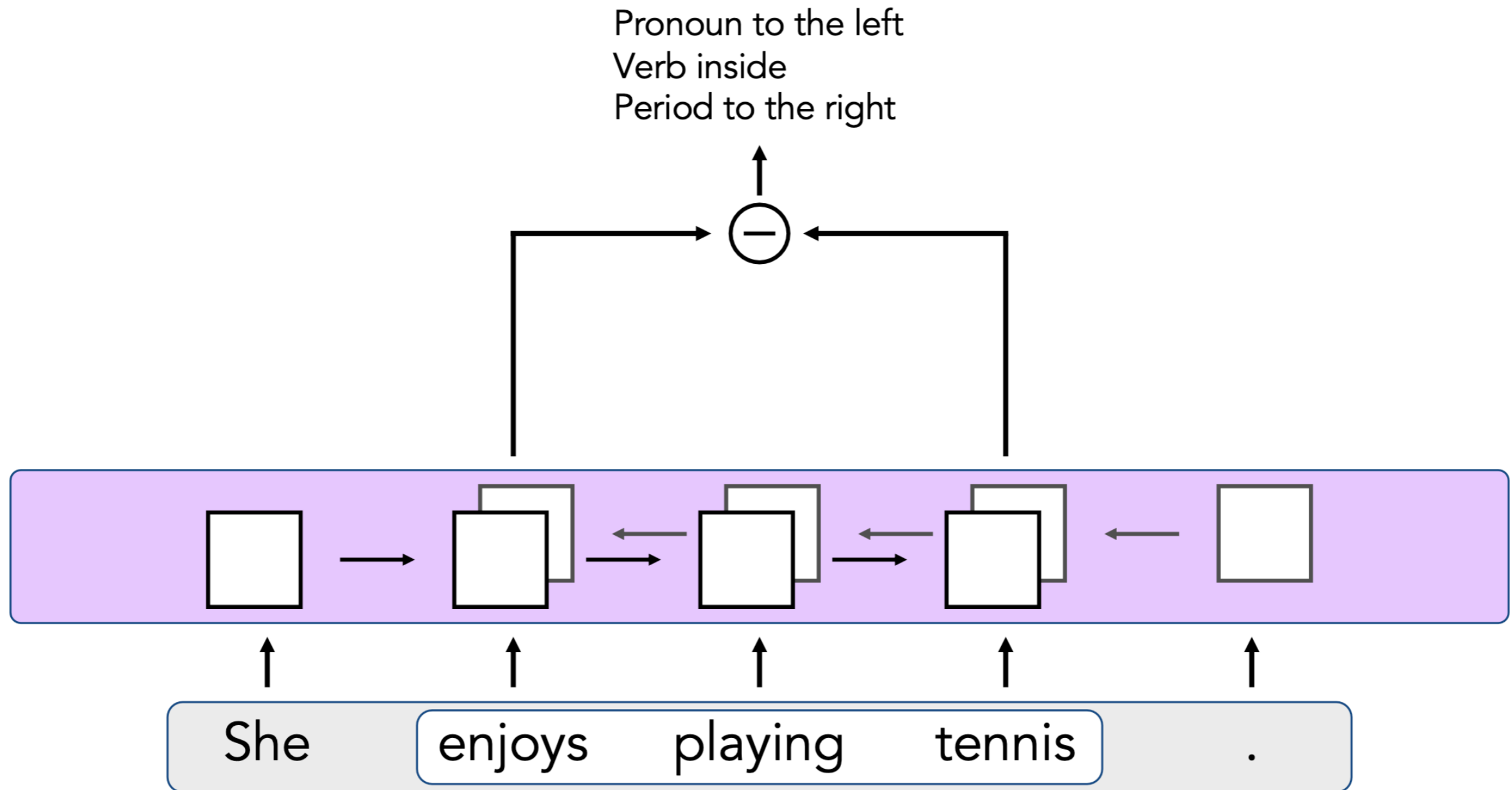
Span-Based Parsing



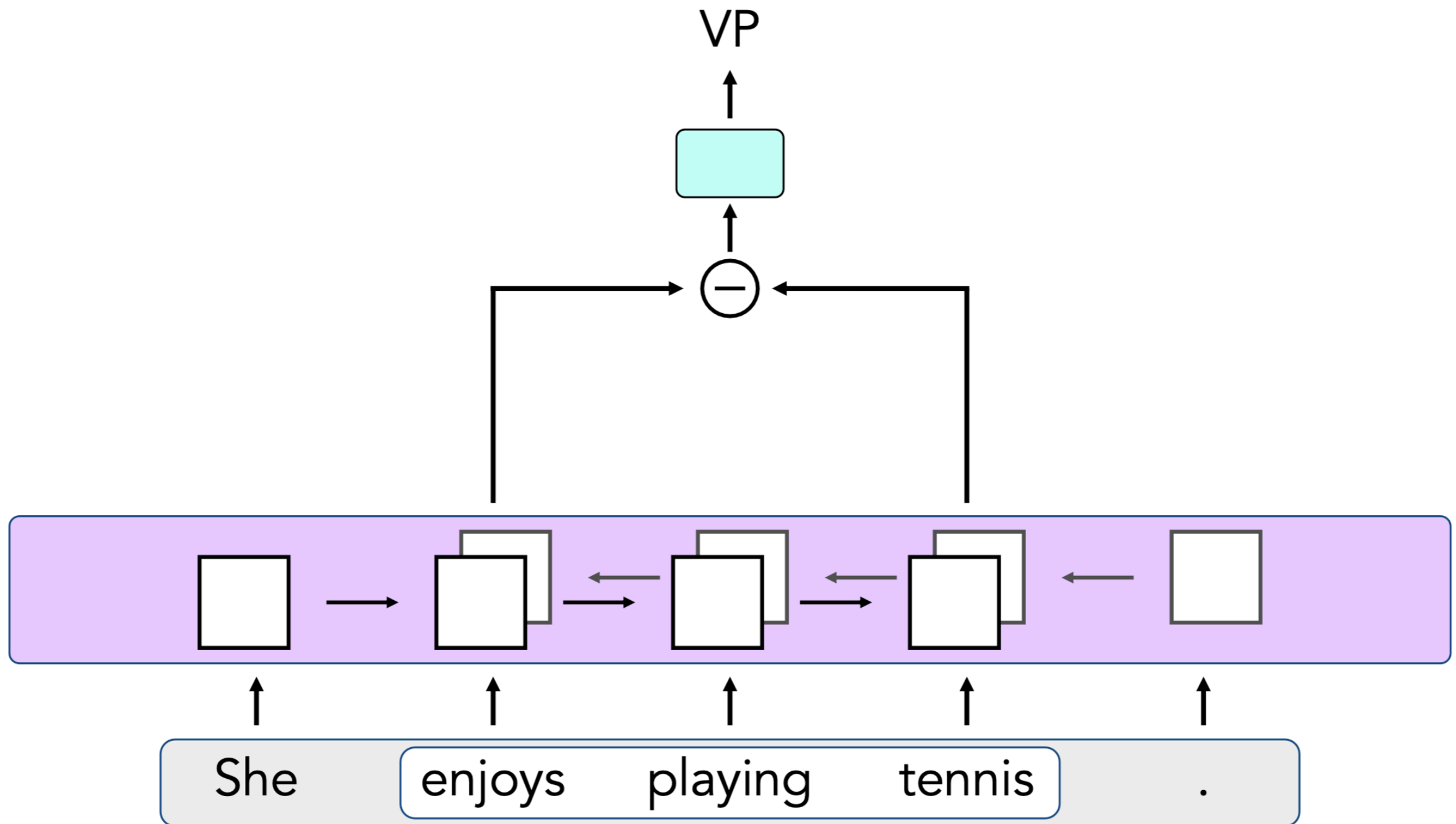
Span-Based Parsing



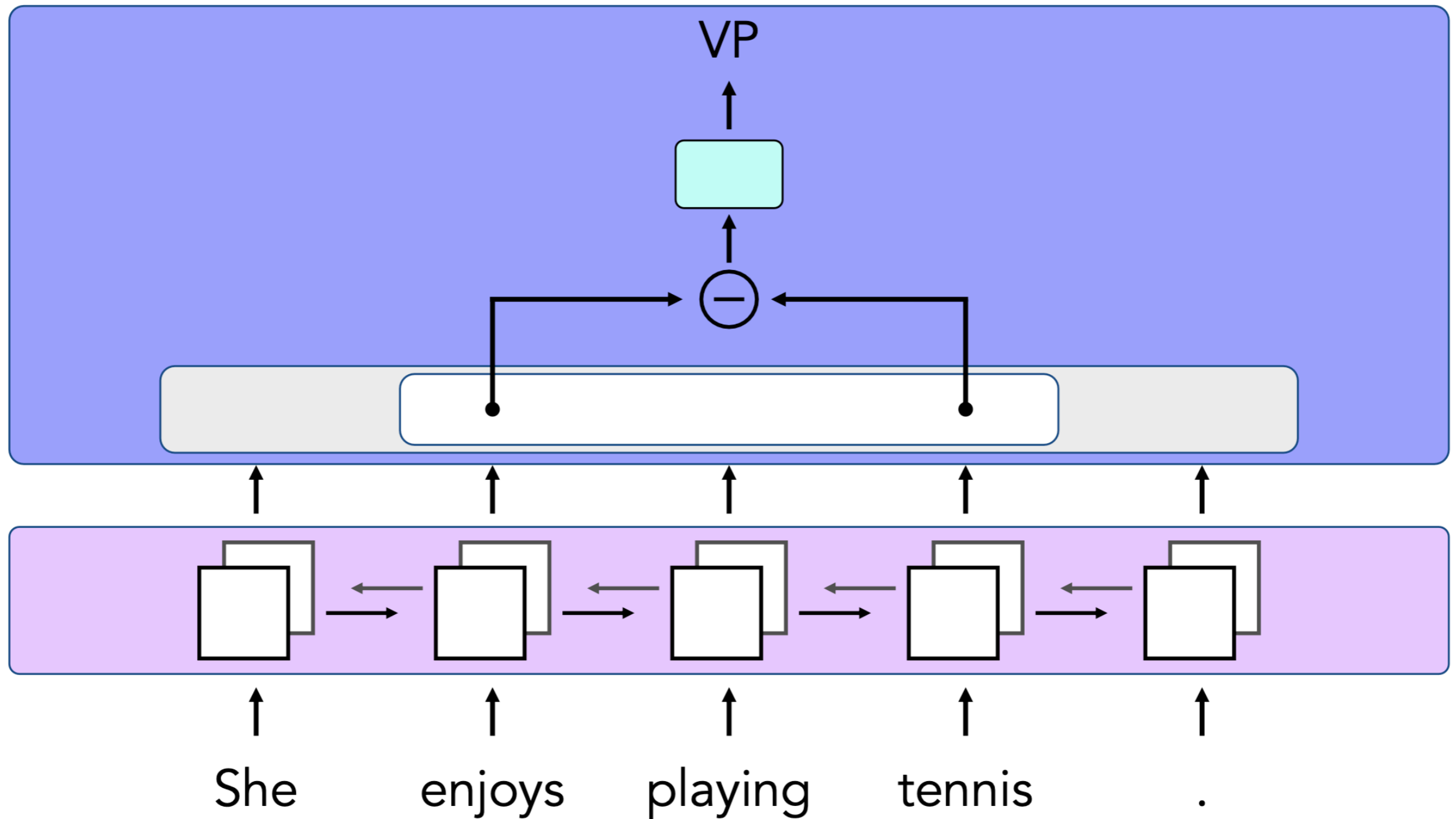
Span-Based Parsing



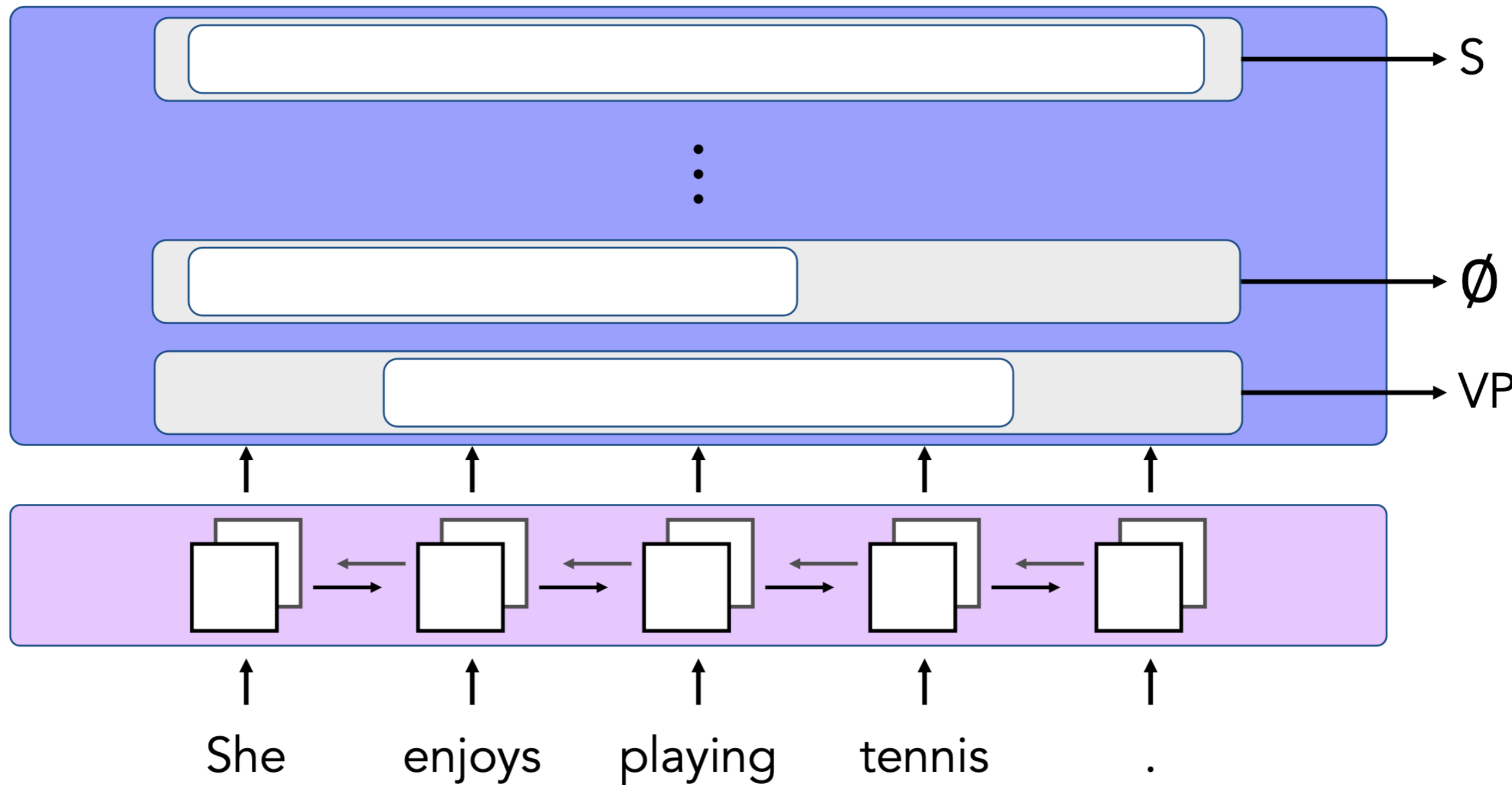
Span-Based Parsing



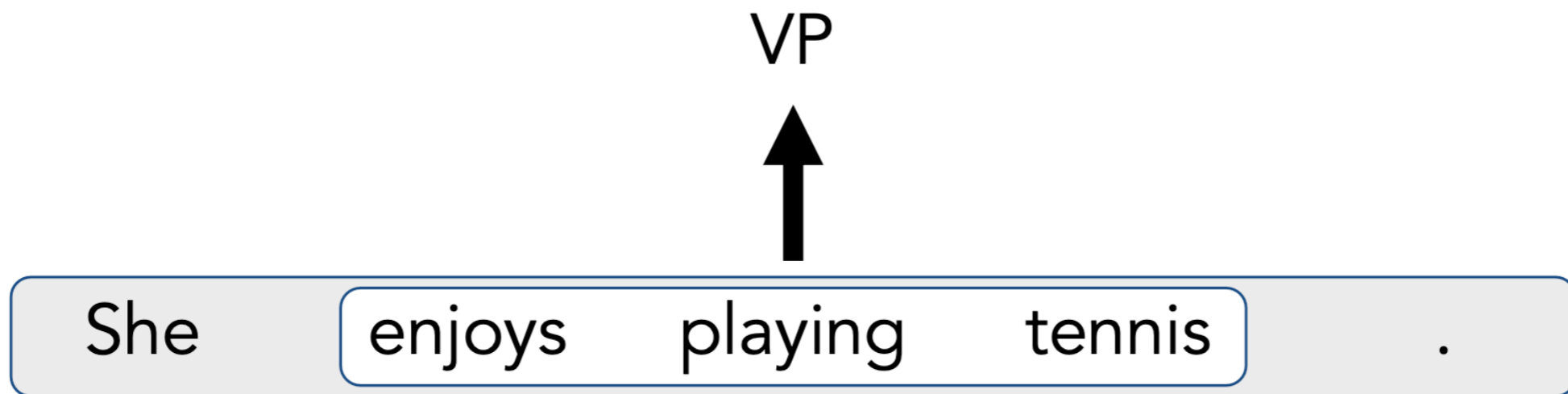
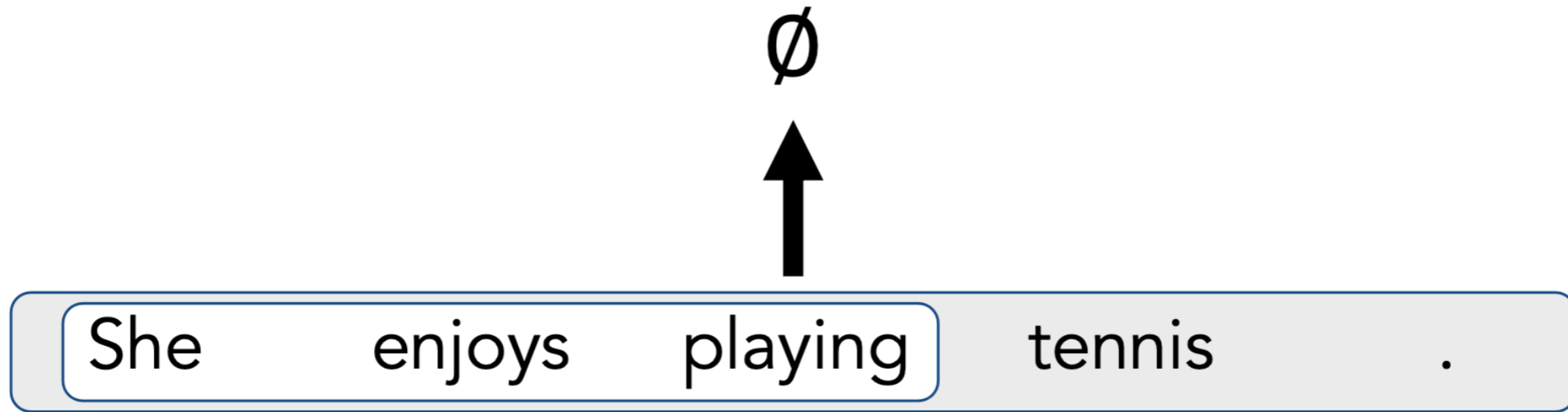
Span-Based Parsing



Span-Based Parsing



Span-Based Parsing



Training: Margin Loss

- Find the best tree using the current model

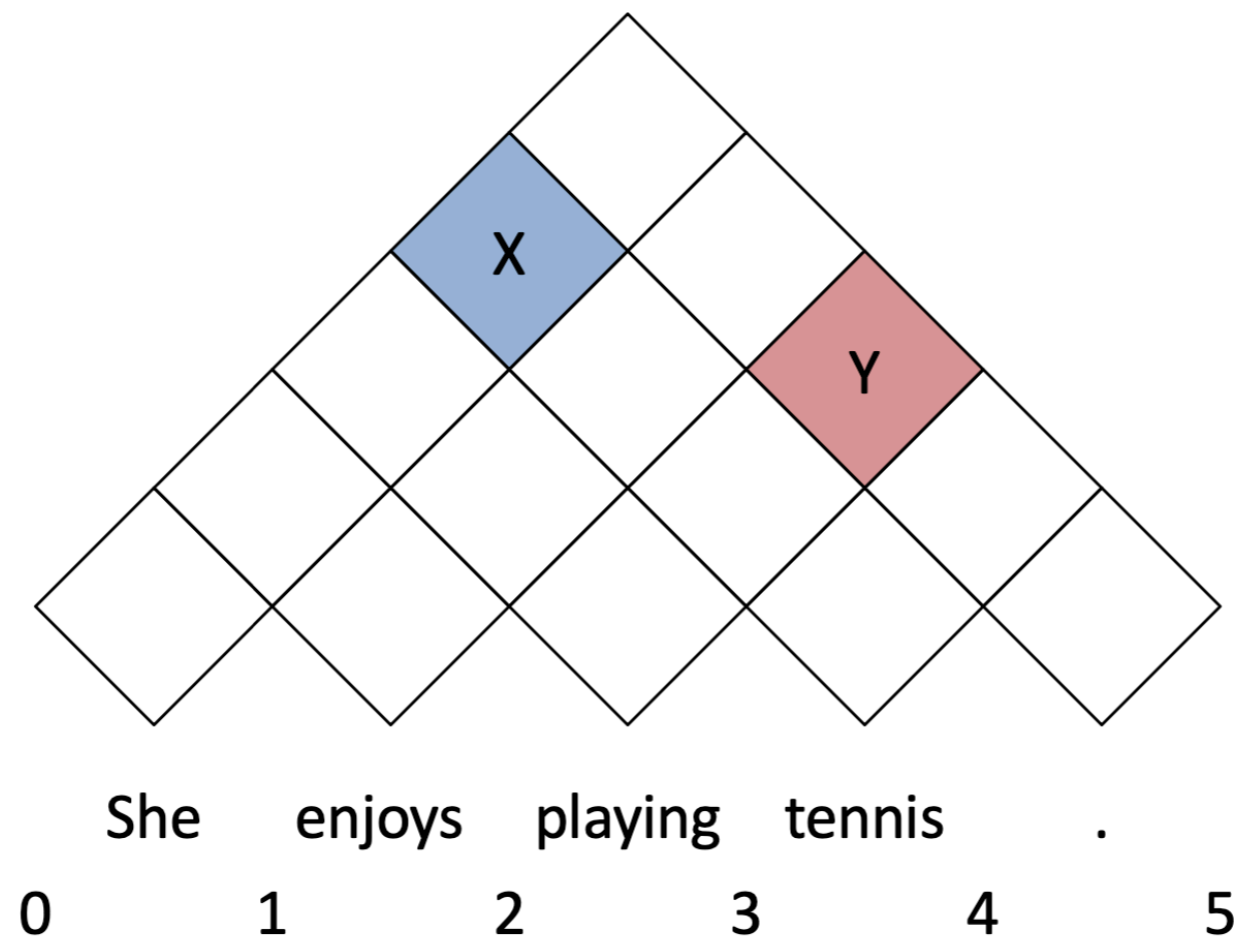
$$\hat{T} = \operatorname{argmax}_T [s_{\text{tree}}(T)].$$

- Margin loss:

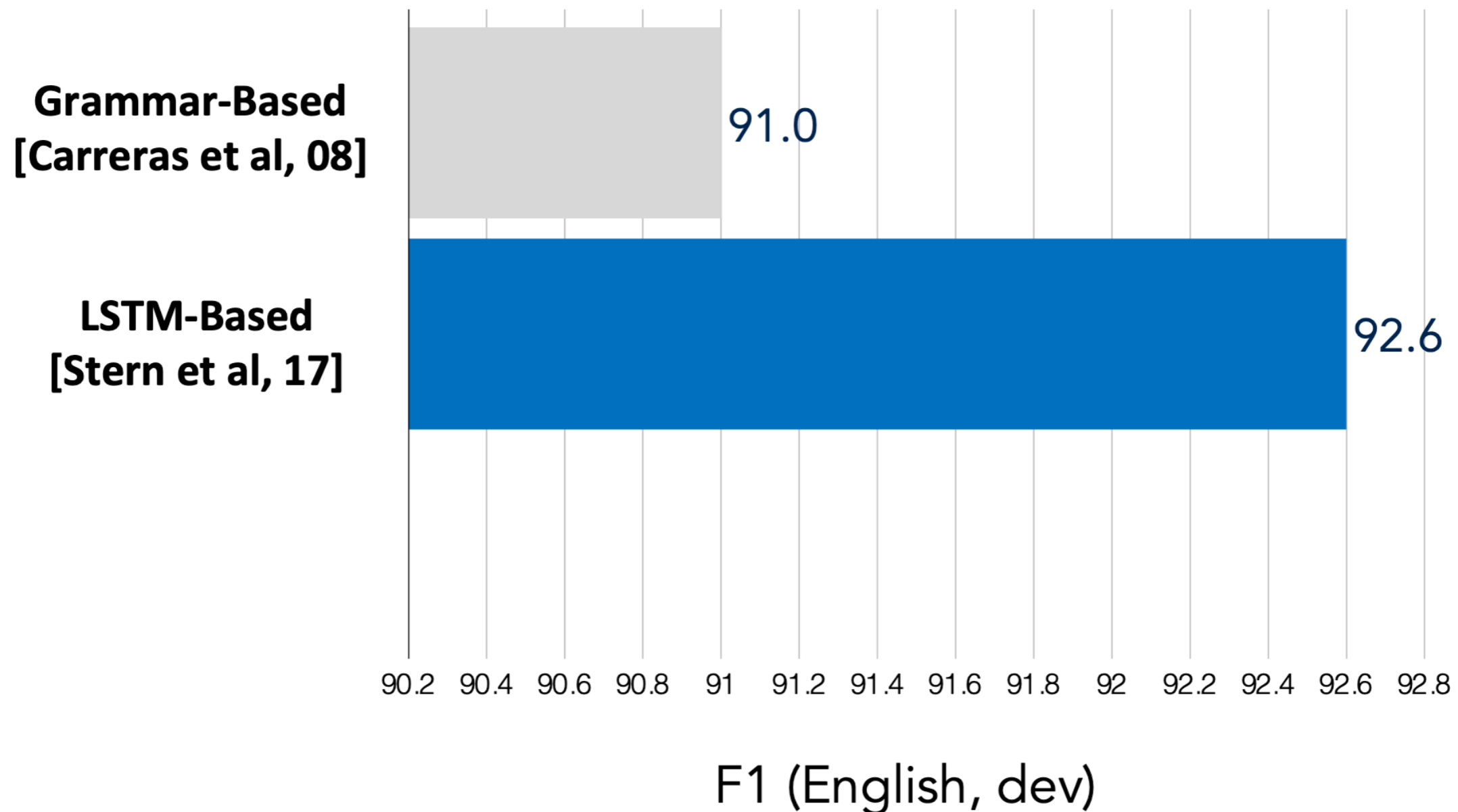
$$\max \left(0, 1 - s_{\text{tree}}(T^*) + s_{\text{tree}}(\hat{T}) \right)$$

Decoding: CYK

- Same as counting-based PCFG
- Use the learned scores for possible spans in the following chart

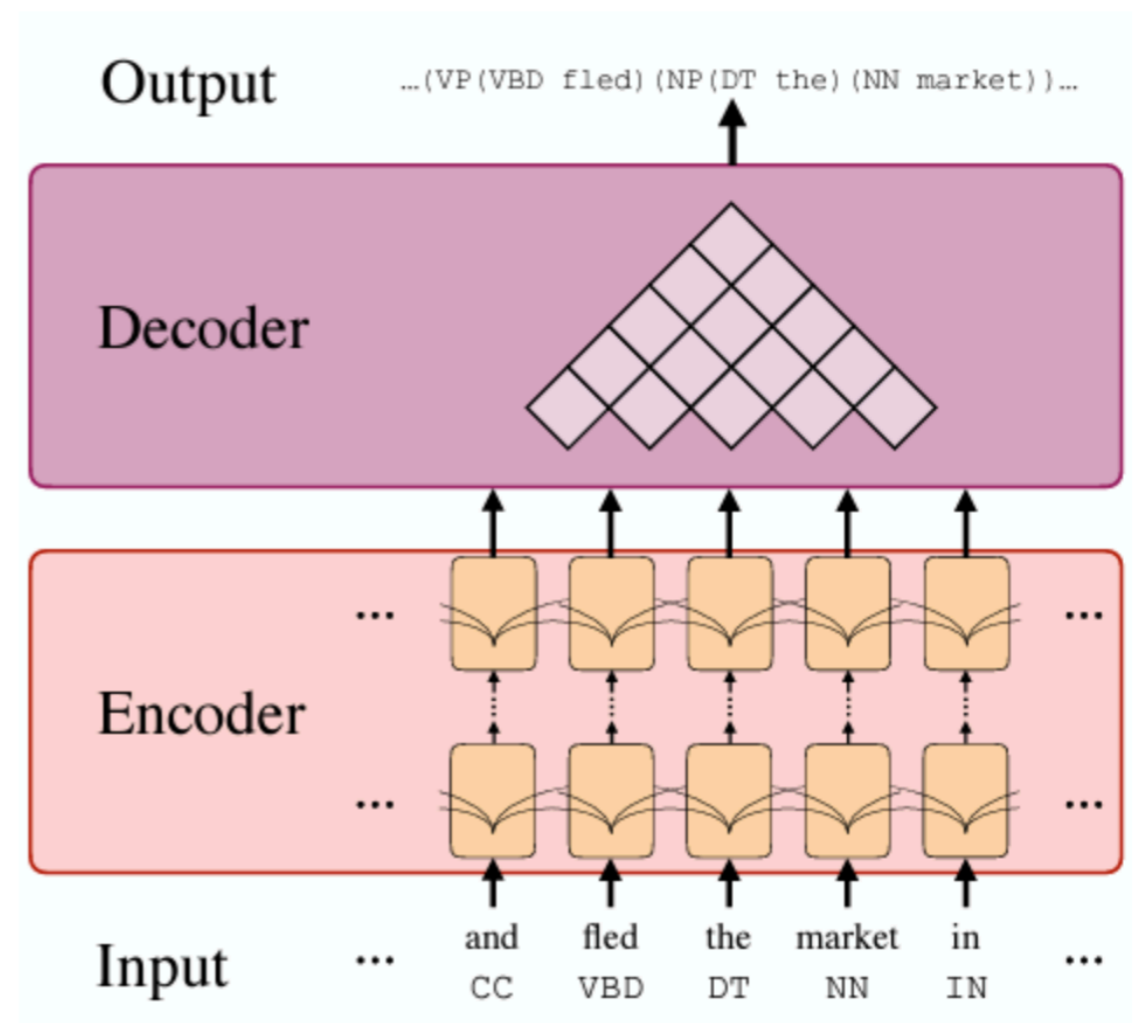


Improves over non-neural methods

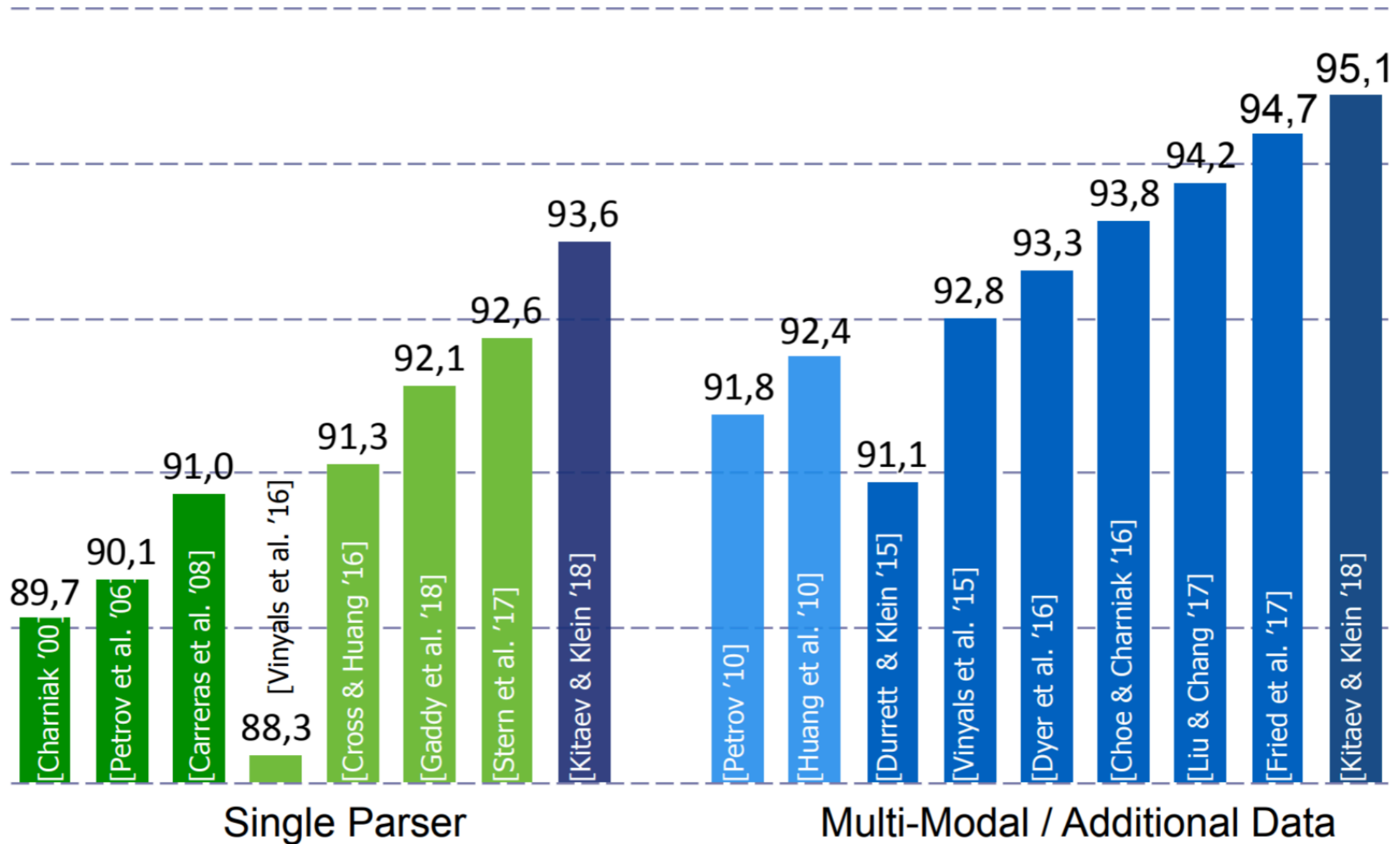


SoTA Self-attention-based Parser

- Use the Transformer encoder instead of Bi-LSTM
 - Split the word hidden vector from Transformer into two half vectors $h_i = [\vec{h}_i; \overleftarrow{h}_i]$
- Replace the forward and backward hidden vectors of Bi-LSTM by the new vectors



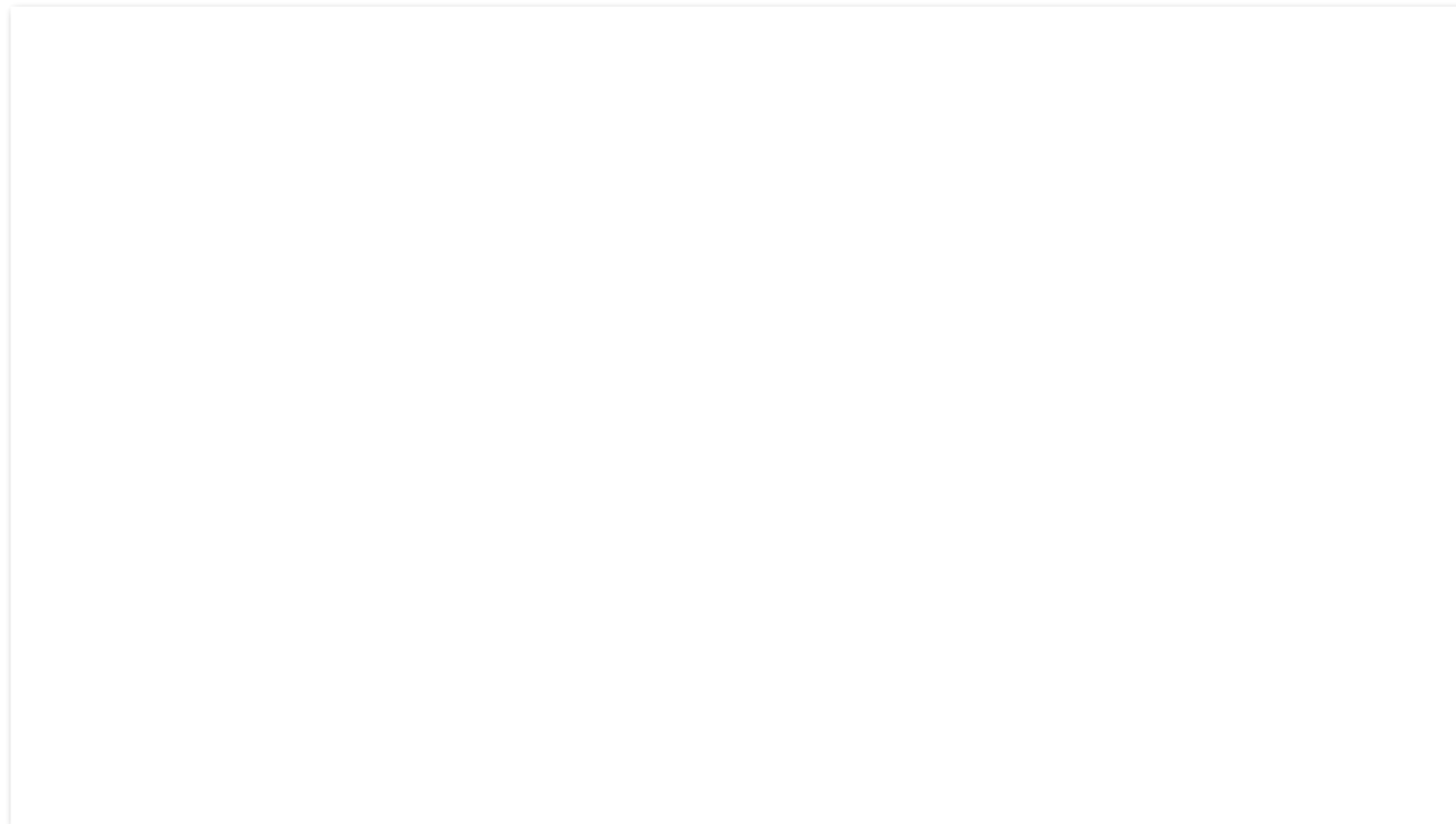
Historical Trends on Penn Treebank



Unsupervised Parsing (Grammar Induction)

Grammar Induction (Unsupervised Parsing)

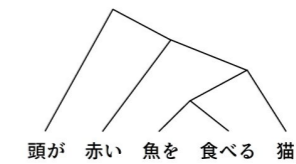
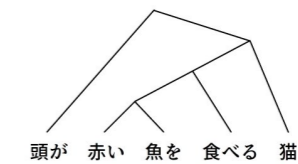
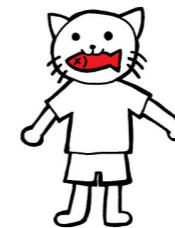
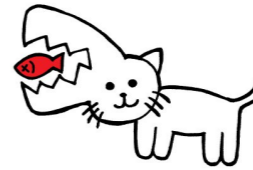
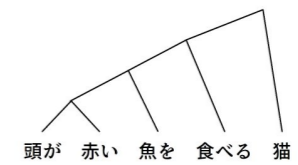
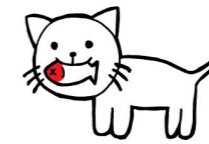
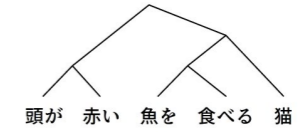
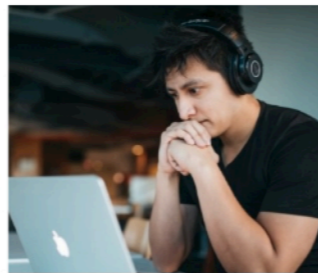
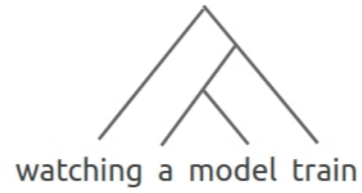
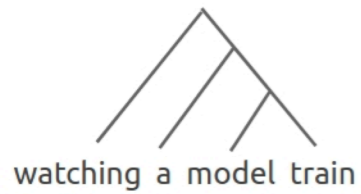
Learning a set of (probabilistic) grammar rules

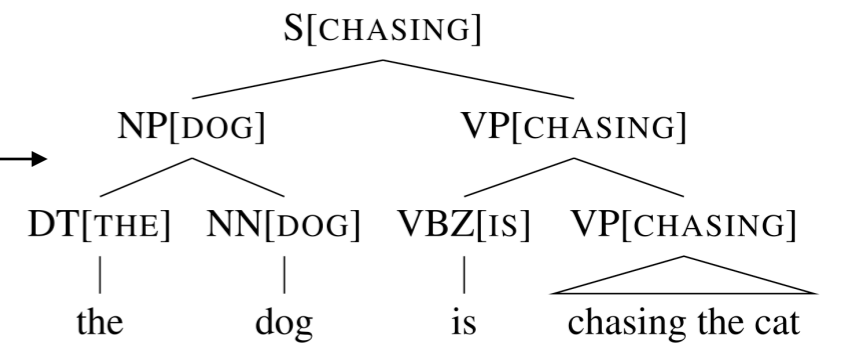
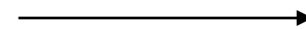
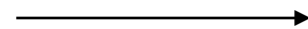


Typical grammar induction methods

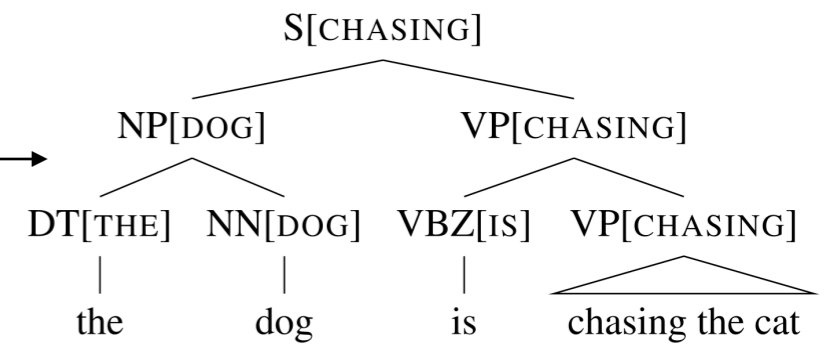
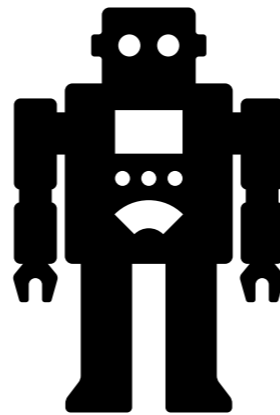
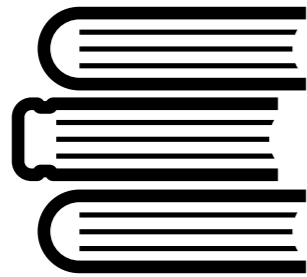
unsupervised constituency and dependency parsing

Explosion of ambiguity





Mystery: humans learn to parse without learning to parse (from labels)



Grammar Induction

Neural PCFGs

- Neural parameterization for PCFGs

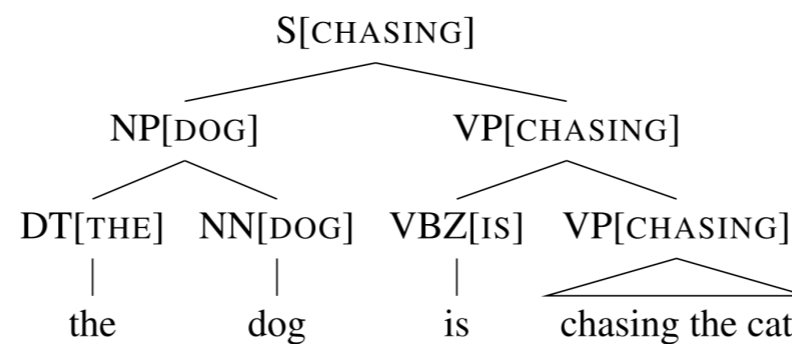
$$\pi_{T \rightarrow w} = \text{NEURALNET}(\mathbf{w}_T) = \frac{\exp(\mathbf{u}_w^\top f(\mathbf{w}_T))}{\sum_{w' \in \Sigma} \exp(\mathbf{u}_{w'}^\top f(\mathbf{w}_T))}$$

$$\pi_{T \rightarrow w} \propto \exp \left(\underbrace{\mathbf{u}_w^\top}_{\text{output emb.}} \overbrace{f(\mathbf{w}_T)}^{\text{shared neural net}} \underbrace{\mathbf{w}_T}_{\text{input emb.}} \right)$$

- same training method
- Where's the magic?

Neural L-PCFGs

- You can further improve Neural PCFGs by adding head annotations



$$S \rightarrow A, \quad A \in \mathcal{N}$$

$$A \rightarrow BC, \quad A \in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P}$$

$$T \rightarrow \alpha, \quad T \in \mathcal{P}$$

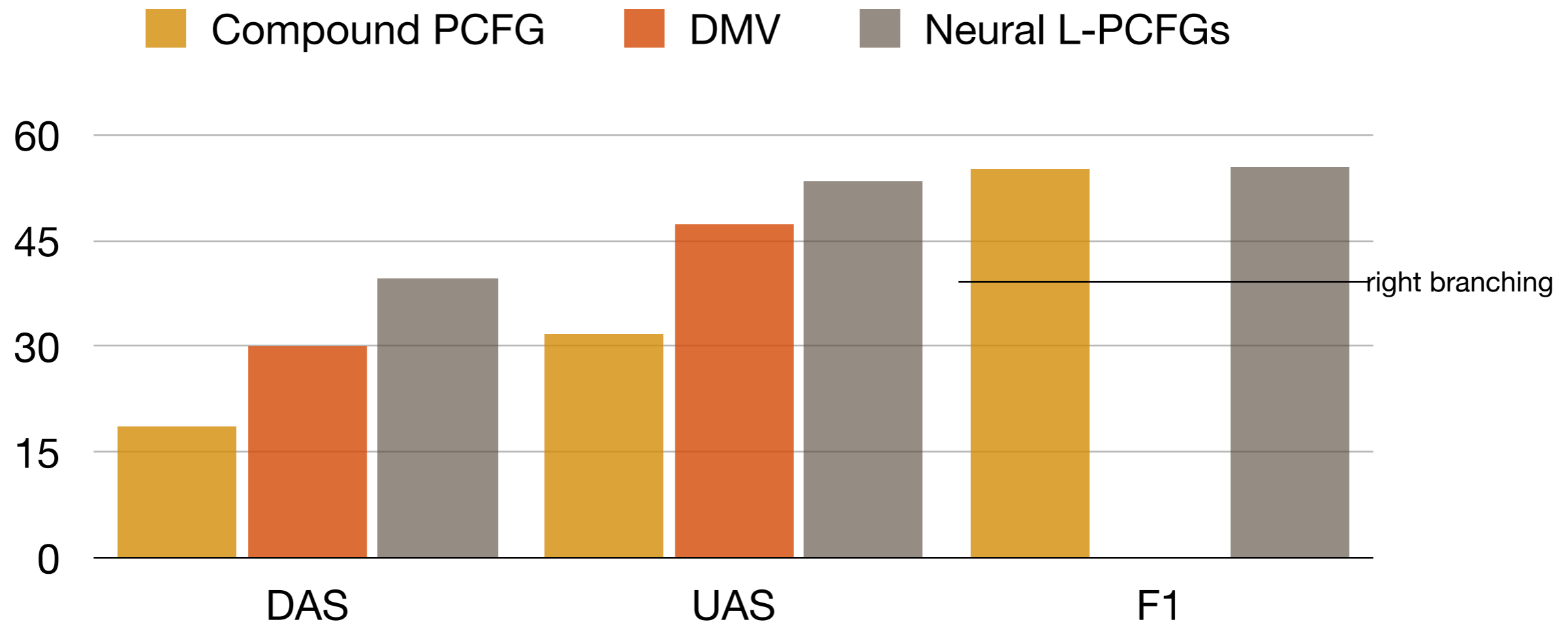
$$\textcircled{1} \quad S \rightarrow A[\alpha], \quad A \in \mathcal{N}$$

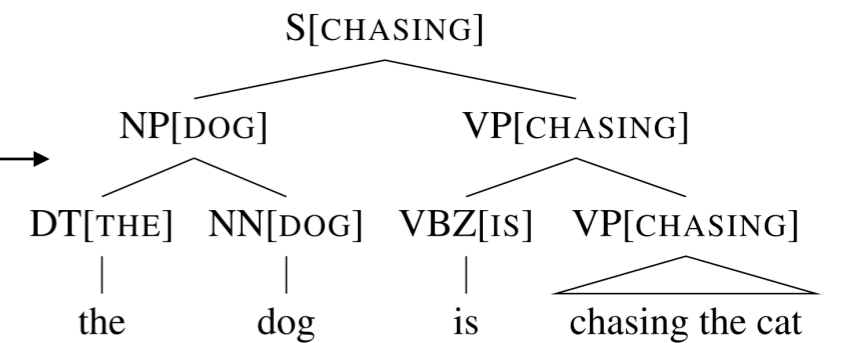
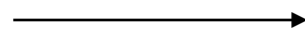
$$\textcircled{2l} \quad A[\alpha] \rightarrow B[\alpha]C[\beta], \quad A \in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P}$$

$$\textcircled{2r} \quad A[\alpha] \rightarrow B[\beta]C[\alpha], \quad A \in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P}$$

$$\textcircled{3} \quad T[\alpha] \rightarrow \alpha, \quad T \in \mathcal{P}$$

Results

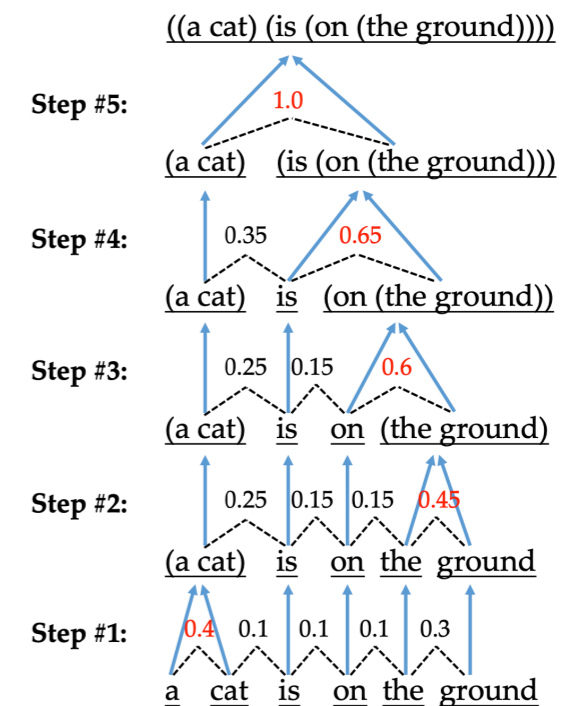
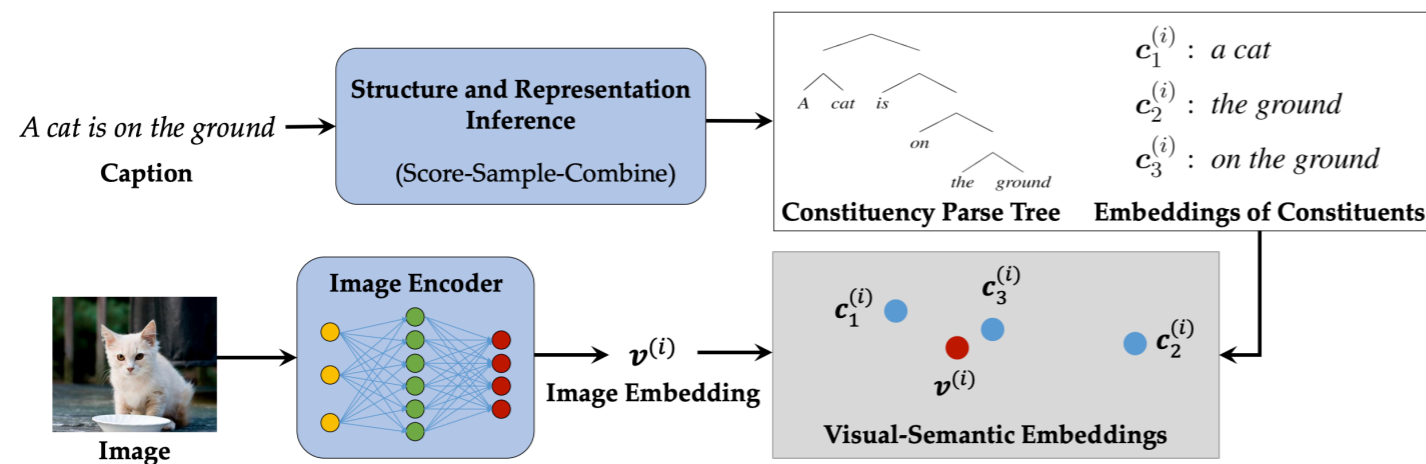




Key to the mystery: visual prior?

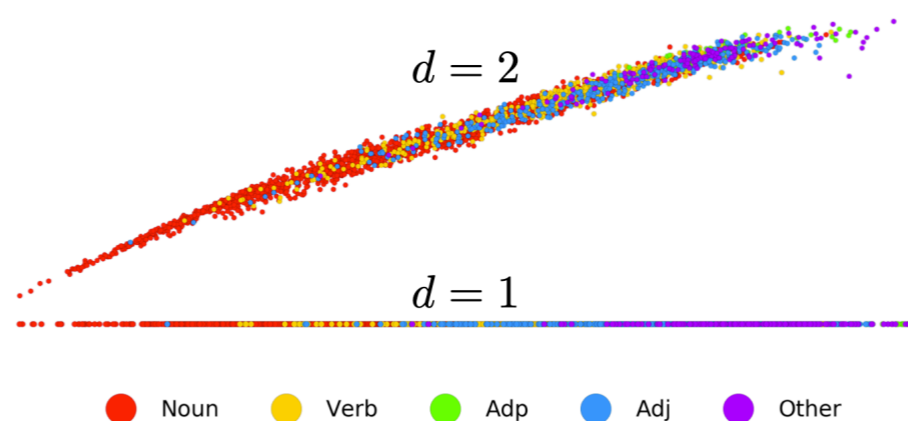
Visual Prior Grammar Induction

- Visual grounded neural syntax acquisition



Visual Prior Grammar Induction

- Visual grounded neural syntax acquisition
 - Similar results even if the dimension of embeddings get shrunk to 1 or 2.
 - embedding tags
 - concrete



Model	NP	VP	PP	ADJP	Avg. F_1
Shi2019	79.6	26.2	42.0	22.0	50.4 ± 0.3
Shi2019*	80.5	26.9	45.0	21.3	51.4 ± 1.1
1, SWS, CME	77.2	17.0	53.4	18.2	49.7 ± 5.9
2, SWS, CME	80.8	19.1	52.3	17.1	51.6 ± 0.6
+HI					
Shi2019	74.6	32.5	66.5	21.7	53.3 ± 0.2
Shi2019*	73.1	33.9	64.5	22.5	51.8 ± 0.3
1, SWS, CME	74.0	35.2	62.0	24.2	51.8 ± 0.4
2, SWS, CME	73.8	30.2	63.7	21.9	51.3 ± 0.1
+HI+FastText					
Shi2019	78.8	24.4	65.6	22.0	54.4 ± 0.3
Shi2019*	77.3	23.9	64.3	21.9	53.3 ± 0.1
1, SWS, CME	76.6	21.9	68.7	20.6	53.5 ± 1.4
2, SWS, CME	77.5	22.8	66.3	19.3	53.6 ± 0.2
+HI+FastText-IN					
Shi2019*	78.3	26.6	67.5	22.1	54.9 ± 0.1
1, SM, CMX	79.6	29.0	38.3	23.5	49.7 ± 0.2
1, SMHI, CMX	77.6	45.0	72.3	24.3	57.5 ± 0.1
1, SM, CME	80.0	26.9	62.2	23.2	54.3 ± 0.2
1, SMHI, CME	76.5	20.5	63.6	22.7	52.2 ± 0.3
1, SWS, CME	77.7	26.3	72.5	22.0	55.5 ± 0.1
2, SWS, CME	78.5	26.3	69.5	21.1	55.2 ± 0.1

Visual Prior Grammar Induction

- Recommend readings
 - *Visually Grounded Compound PCFGs.*
 - *Dependency Induction Through the Lens of Visual Perception*

Questions?