Natural Language Processing (7)

Parsing (1): Constituency Parsing

Daisuke Kawahara

Department of Communications and Computer Engineering, Waseda University

Lecture Plan

1. Overview of Natural Language Processing
2. Formal Language Theory
3. Word Senses and Embeddings
4. Topic Models
5. Collocations, Language Models, and Recurrent Neural Networks
6. Sequence Labeling and Morphological Analysis
7. Parsing (1)
8. Parsing (2)
9. Transfer Learning
10. Knowledge Acquisition
11. Information Retrieval, Question Answering, and Machine Translation
12. Guest Talk (1)
13. Guest Talk (2)
14. Project: Survey or Programming
15. Project Presentation

Table of Contents

CFG and Syntactic Structure

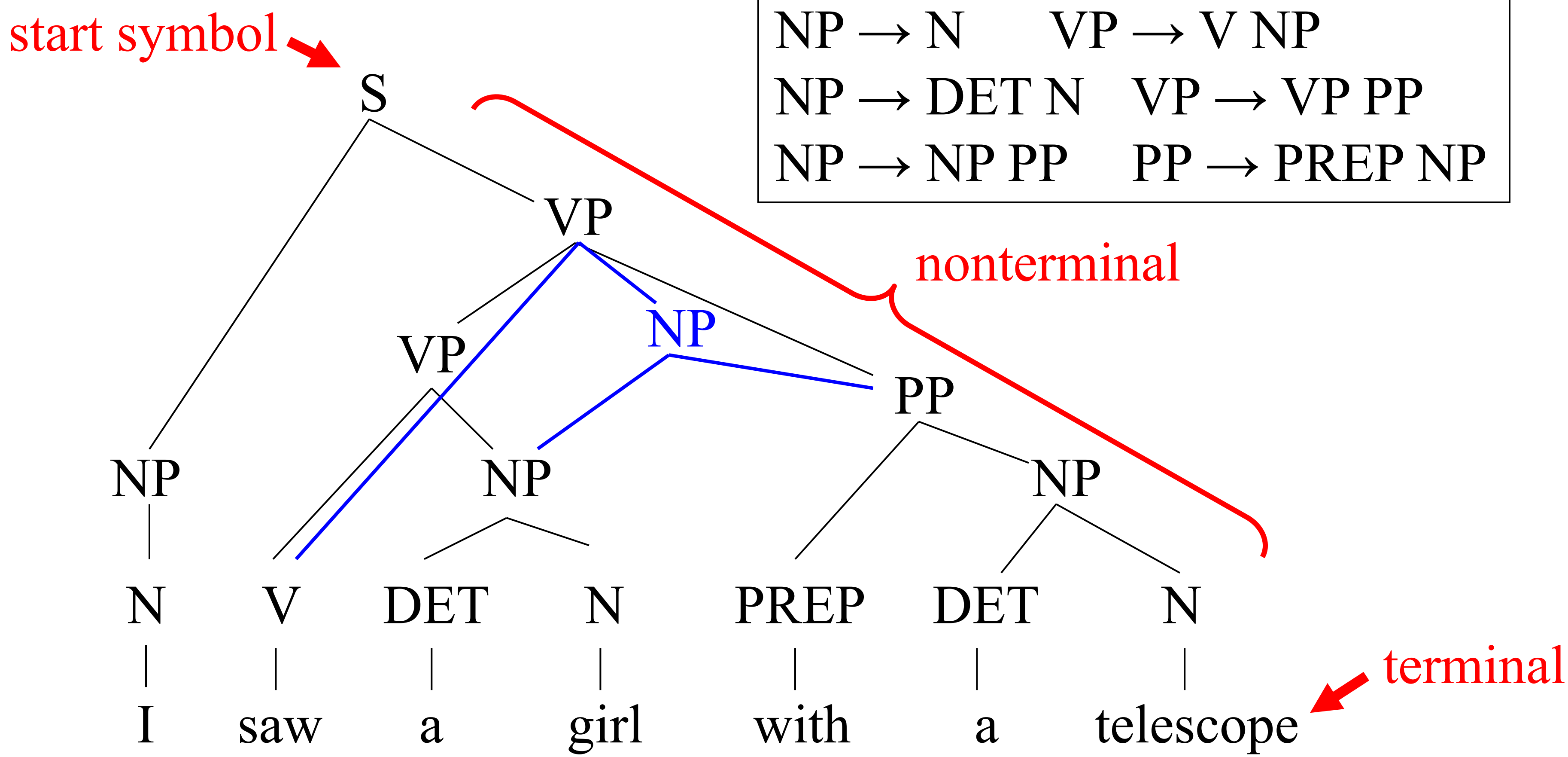
production rule

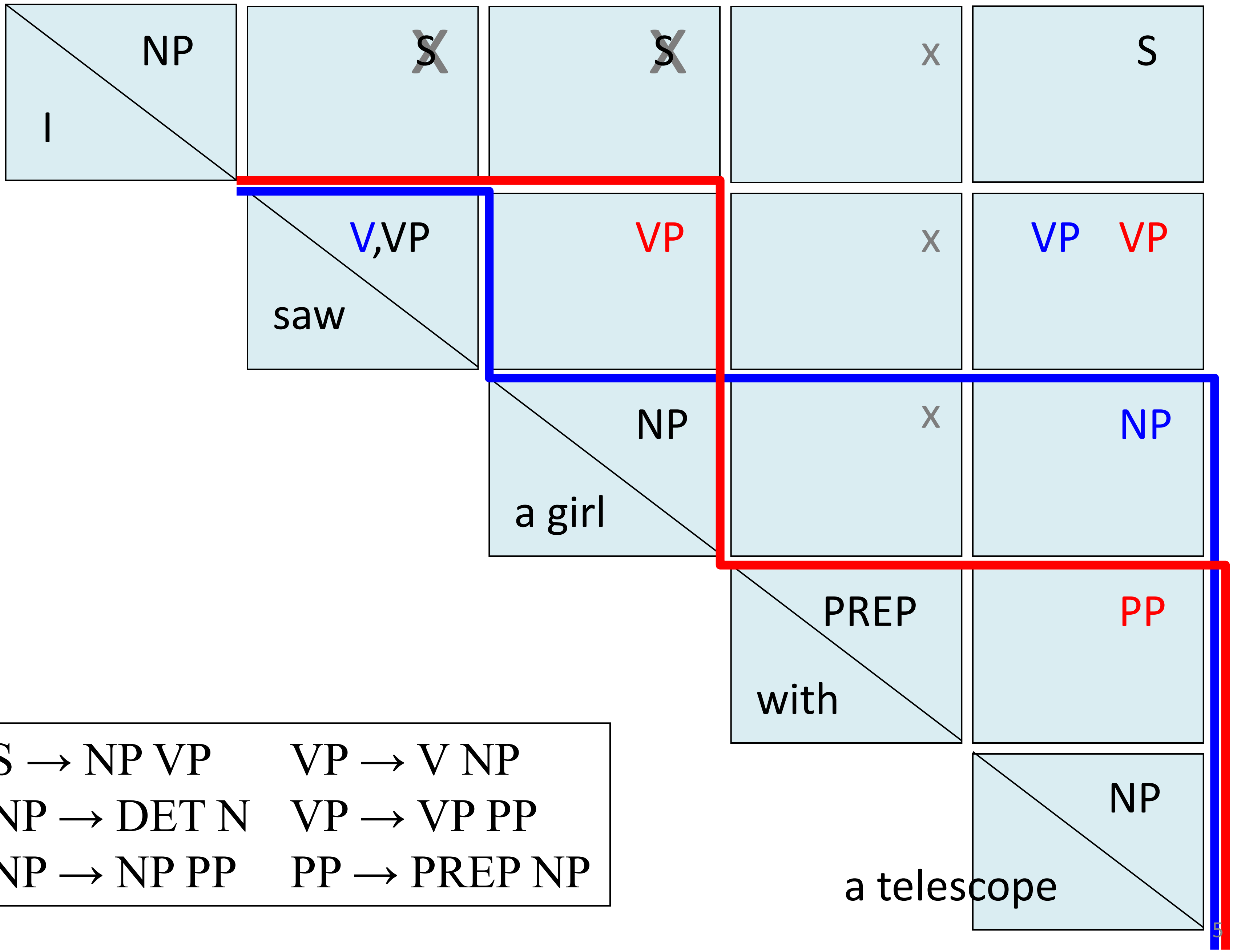
$S \rightarrow NP\ VP$ $\quad VP \rightarrow V$

$NP \rightarrow N$ $\quad VP \rightarrow V\ NP$

$NP \rightarrow DET\ N$ $\quad VP \rightarrow VP\ PP$

$NP \rightarrow NP\ PP$ $\quad PP \rightarrow PREP\ NP$





$S \rightarrow NP\ VP$ $\quad VP \rightarrow V\ NP$

$NP \rightarrow DET\ N$ $\quad VP \rightarrow VP\ PP$

$NP \rightarrow NP\ PP$ $\quad PP \rightarrow PREP\ NP$

Probabilistic Context Free Grammar (PCFG)

- A set of terminals $\{w^k\}, k = 1,\ldots,V$
- A set of nonterminals $\{N^i\}, i = 1,\ldots,n$
- A designated start symbol N^1
- A set of rules $\{N^i \rightarrow \zeta^j\}$
- A corresponding set of probabilities on rules

$$\forall i \sum_j P\left(N^i \rightarrow \zeta^j\right) = 1$$

A simple PCFG

S → NP VP	1.0
PP → P NP	1.0
VP → V NP	0.7
VP → VP PP	0.3
P → *with*	1.0
V → *saw*	1.0

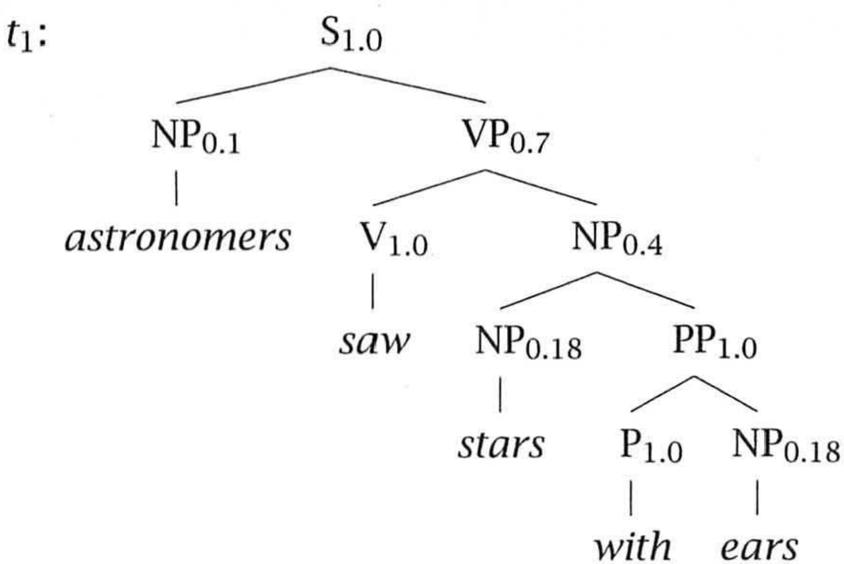
NP → NP PP	0.4
NP → *astronomers*	0.1
NP → *ears*	0.18
NP → *saw*	0.04
NP → *stars*	0.18
NP → *telescopes*	0.1

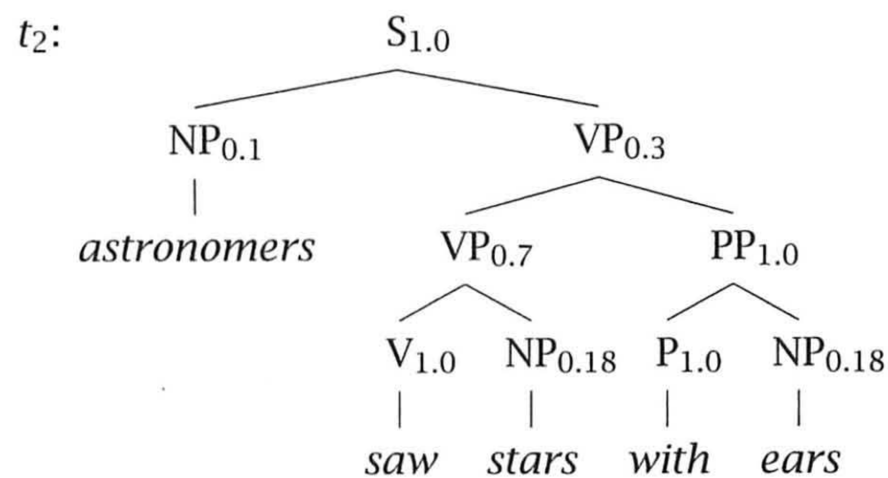
Chomsky Normal Form
- $A \rightarrow BC$
- $A \rightarrow \alpha$

$\Sigma\, P(\text{NP} \rightarrow *) = 1$

Parse Trees

astronomers saw stars with ears





$P(t_1) = 0.0009072$

$P(t_2) = 0.0006804$

$$P(S) = P(t_1) + P(t_2) = 0.0015876$$

Probabilistic Context Free Grammar (PCFG)

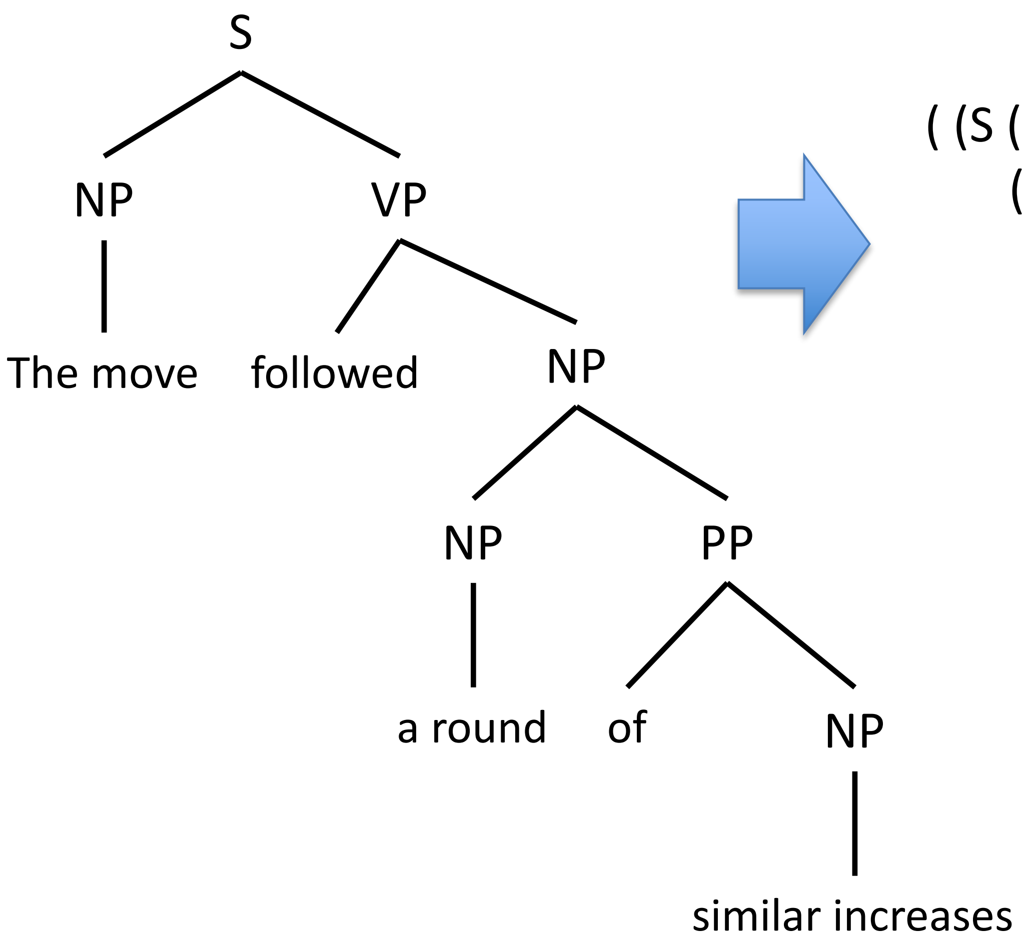
- A set of terminals $\{w^k\}, k = 1,\ldots,V$
- A set of nonterminals $\{N^i\}, i = 1,\ldots,n$
- A designated start symbol N^1
- A set of rules $\{N^i \to \zeta^j\}$
- A corresponding set of probabilities on rules

$$\forall i \sum_j P\left(N^i \to \zeta^j\right) = 1$$

Penn Treebank [Marcus+ 1993]

- An annotated corpus for English
  - Wall Street Journal (1 million words)
  - Brown Corpus
  - Switchboard Corpus (telephone conversation)
  - ATIS (Air Travel Information System) Corpus
- Released by LDC
- The de facto data for English parsing (training and evaluation)

A Penn Treebank Tree




```
( (S (NP The move)
     (VP followed
         (NP (NP a round)
             (PP of
                 (NP similar increases))))))
```

Phrasal Categories of Penn Treebank

S	Simple clause (sentence)
SBAR	S’ clause with complementizer
SBARQ	Wh-question S’ clause
SQ	Inverted Yes/No question S’ clause
SINV	Declarative inverted S’ clause
ADJP	Adjective phrase
ADVP	Adverbial phrase
NP	Noun phrase
PP	Prepositional phrase
QP	Quantifier phrase (inside NP)
VP	Verb phrase
WHNP	Wh-noun phrase
WHPP	Wh-prepositional phrase

CONJP	Multiword conjunction phrase
FRAG	Fragment
INTJ	Interjection
LST	List marker
NAC	Not a consistent grouping
NX	Nominal constituent inside NP
PRN	Parenthetical
PRT	Particle
RRC	Reduced relative clause
UCP	Unlike coordinated phrase
X	Unknown or uncertain
WHADJP	Wh-adjective phrase
WHADVP	Wh-adverbial phrase

Exercise

```
( (S (NP (NP (NNP Pierre) (NNP Vinken))
          (, ,)
          (NP (NP (CD 61) (NNS years))
               (ADJP (JJ old)))
          (, ,))
     (MD will)
     (VP (VB join)
          (NP (DT the) (NN board))
          (PP (IN as)
               (NP (DT a) (JJ nonexecutive) (NN director)))
          (ADVP (NP (NNP Nov.) (CD 29)))))
     (. .))
```

Probabilistic Context Free Grammar (PCFG)

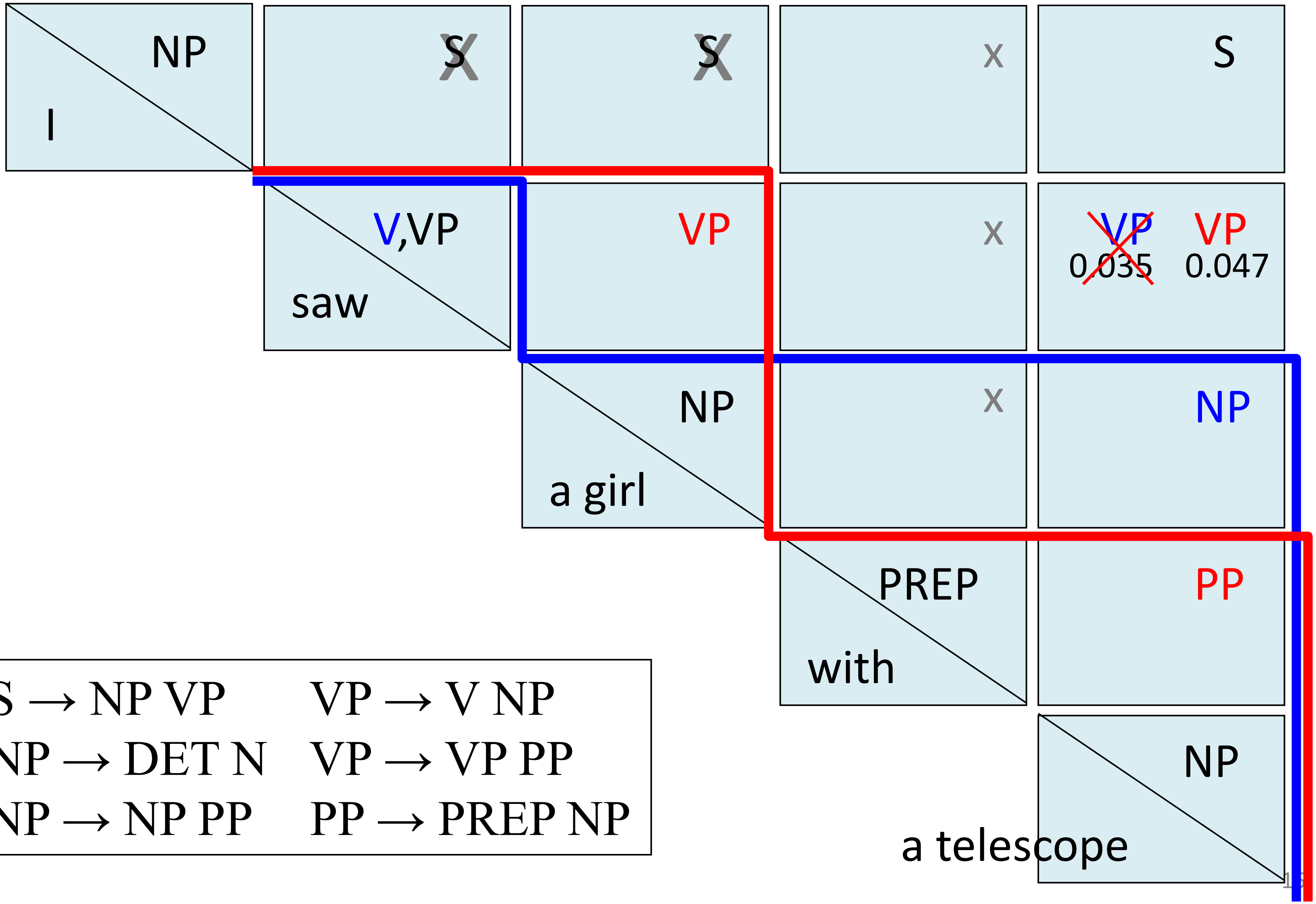
- A set of terminals $\{w^k\}, k = 1,\ldots,V$
- A set of nonterminals $\{N^i\}, i = 1,\ldots,n$
- A designated start symbol N^1
- A set of rules $\{N^i \rightarrow \zeta^j\}$
- A corresponding set of probabilities on rules

$$\forall i \sum_j P\left(N^i \rightarrow \zeta^j\right) = 1$$

Maximum Likelihood Estimation using a treebank

$$\hat{P}\left(N^i \rightarrow \zeta^j\right) = \frac{C\left(N^i \rightarrow \zeta^j\right)}{C\left(N^i\right)}$$

A Dynamic Programming Algorithm



$S \rightarrow NP\ VP$ $\quad VP \rightarrow V\ NP$

$NP \rightarrow DET\ N$ $\quad VP \rightarrow VP\ PP$

$NP \rightarrow NP\ PP$ $\quad PP \rightarrow PREP\ NP$

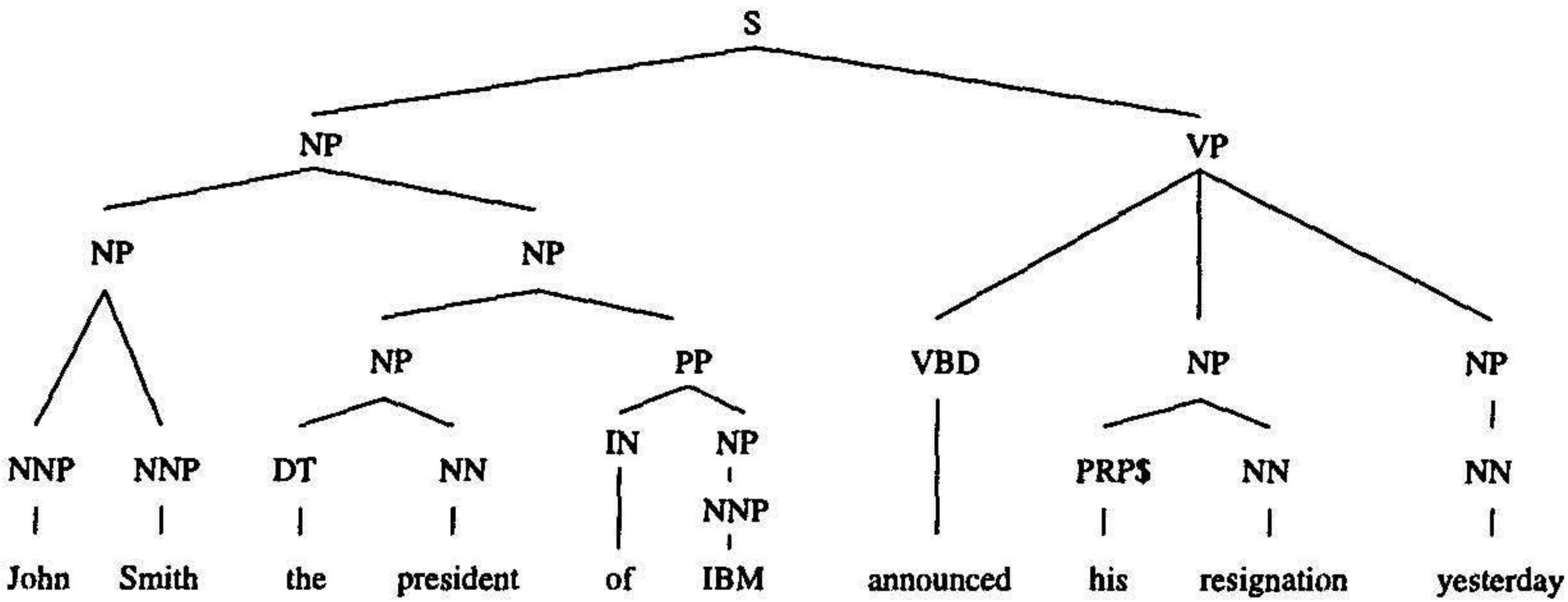
Extensions

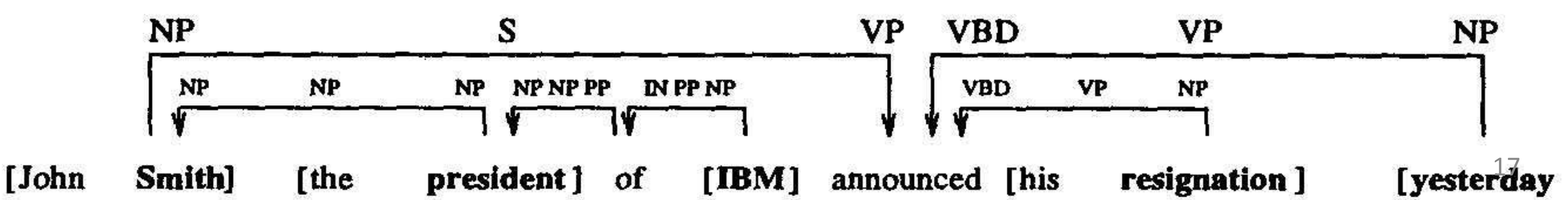
- Lexicalization
  → Use lexical information
- History
  – PCFG: too strong independence assumption
  → Use history
- Nonterminal Classification

Lexicalization

A New Statistical Parser Based on Bigram Lexical Dependencies

[Collins 1996]





Lexicalization

A New Statistical Parser Based on Bigram Lexical Dependencies

[Collins 1996]

$$T_{best} = \arg\max_{T} P(T \mid S) = \arg\max_{T} P(B \mid S) \times P(D \mid S, B)$$

Base NP Model Dependency Model

B={ [John Smith] , [the president], [IBM], [his resignation], [yesterday] }

NP S VP NP NP NP NP NP PP IN PP NP VBD VP NP

D={ Smith announced , Smith president , president of, of IBM , announced resignation

VBD VP NP

announced yesterday }

Base NP Model

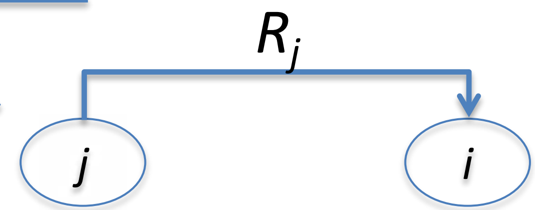
$$P(B \mid S) = \prod_{i=2 \ldots n} \hat{P}(G_i \mid w_{i-1}, t_{i-1}, w_i, t_i, c_i)$$

S(tart)
C(ontinue)
E(nd)
B(etween)
N(ull)

<u>S</u> John <u>C</u> Smith <u>B</u> the <u>C</u> president <u>E</u> of <u>S</u> IBM <u>E</u> announced …

Dependency Model

$$P(D \mid S,B) = \prod_{j=1}^{m} P(AF(j) = (i,R_j) \mid S,B)$$



e.g. *AF*(1) = (5, <NP, S, VP>)

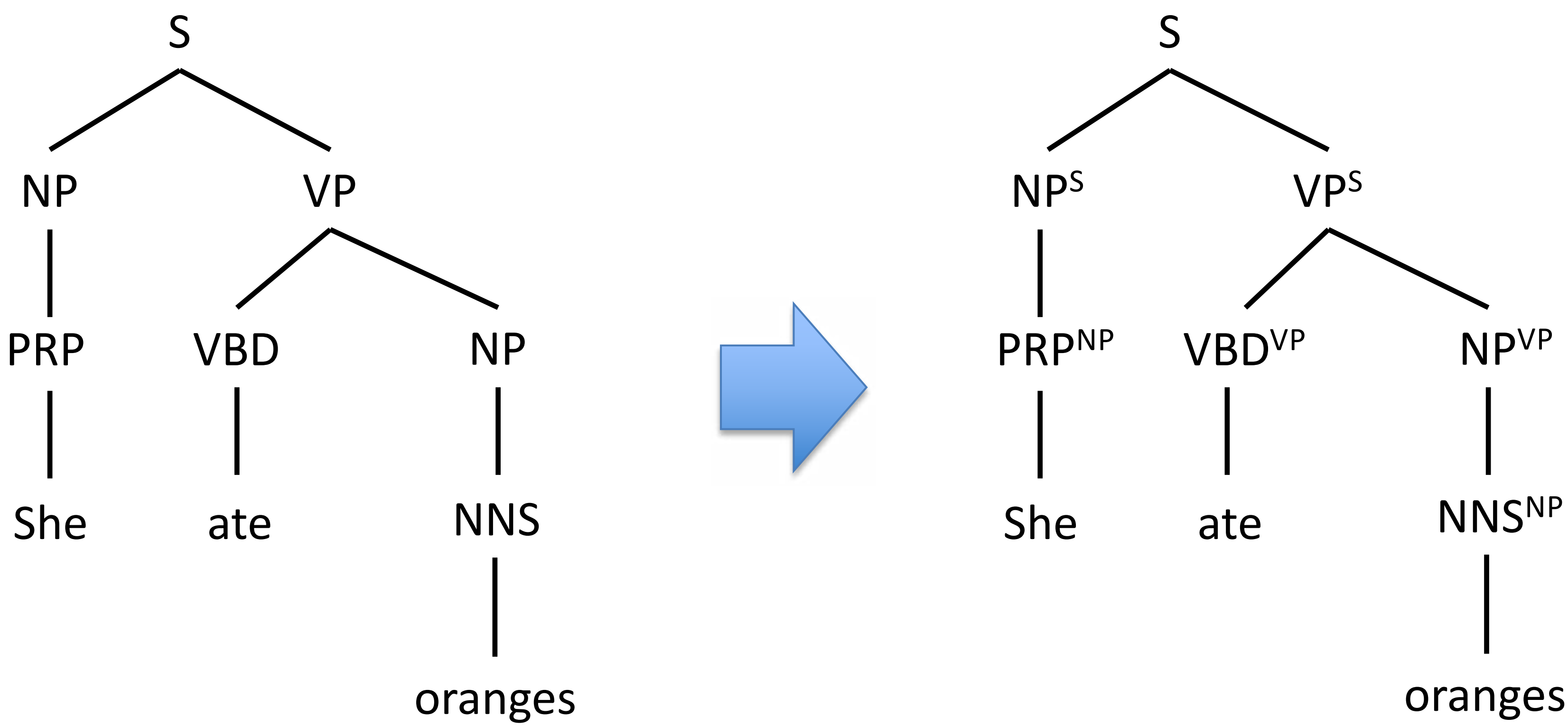
$$= \prod_{j=1}^{m} \frac{F(R_j \mid < w_j, t_j >, < w_i, t_i >)}{\sum_{k=1 \ldots m, k \neq j, p \in P} F(p \mid < w_j, t_j >, < w_k, t_k >)}$$

F(*R*|<a,b>, <c,d>) = *C*(*R*, <a,b>, <c,d>) / *C*(<a,b>, <c,d>)

e.g. $\dfrac{C(\text{<NP,S,VP>, <Smith, NNP>, <announced, VBD>})}{C(\text{<Smith, NNP>, <announced, VBD>})}$

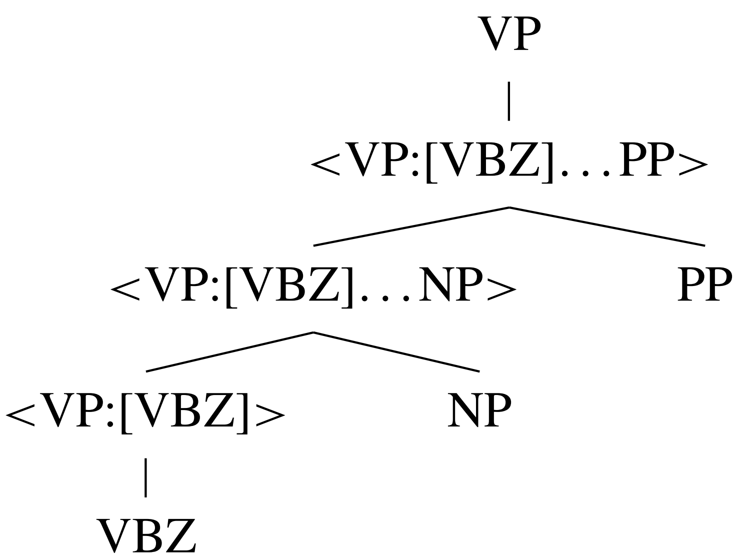
History: Parent Annotation

[Johnson 1998]



Vertical/Horizontal Markovization

[Klein & Manning 2003]



v=1, h=1 markovization of “VP → VBZ NP PP”

F1 and grammar size:

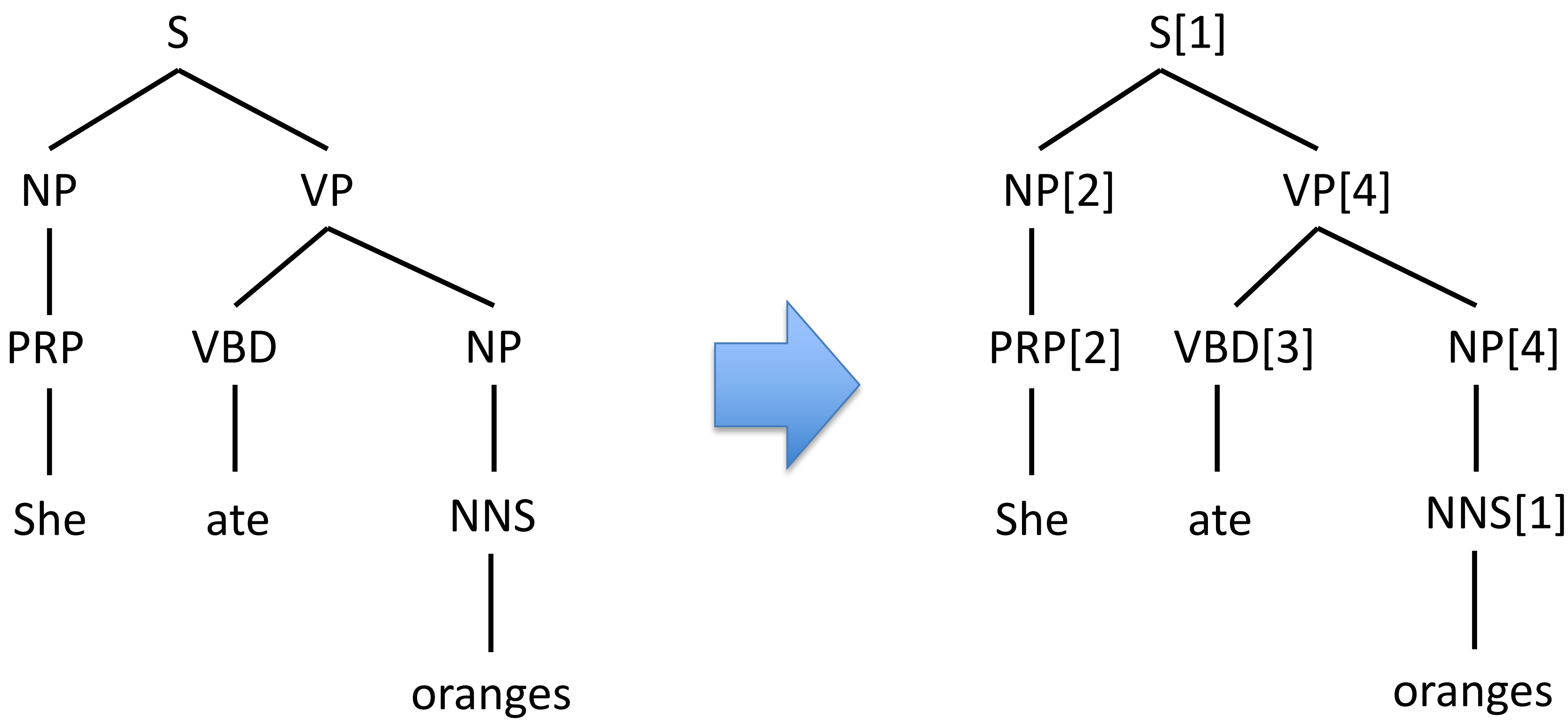
PCFG

Vertical Order		Horizontal Markov Order $h = 0$	$h = 1$	$h \leq 2$	$h = 2$	$h = \infty$
$v = 1$	No annotation	71.27	72.5	73.46	72.96	72.62
		(854)	(3119)	(3863)	(6207)	(9657)
$v \leq 2$	Sel. Parents	74.75	77.42	77.77	77.50	76.91
		(2285)	(6564)	(7619)	(11398)	(14247)
$v = 2$	All Parents	74.68	77.42	77.81	77.50	76.81
		(2984)	(7312)	(8367)	(12132)	(14666)
$v \leq 3$	Sel. GParents	76.50	78.59	79.07	78.97	78.54
		(4943)	(12374)	(13627)	(19545)	(20123)
$v = 3$	All GParents	76.74	79.18	79.74	79.07	78.72
		(7797)	(15740)	(16994)	(22886)	(22002)

Parent Annotation

Nonterminal Classification

[Matsuzaki+ 2005] [Petrov+ 2006]



VBZ

VBZ-0	gives	sells	takes
VBZ-1	comes	goes	works
VBZ-2	includes	owns	is
VBZ-3	puts	provides	takes
VBZ-4	says	adds	Says
VBZ-5	believes	means	thinks
VBZ-6	expects	makes	calls
VBZ-7	plans	expects	wants
VBZ-8	is	's	gets
VBZ-9	's	is	remains
VBZ-10	has	's	is
VBZ-11	does	Is	Does

NNP

NNP-0	Jr.	Goldman	INC.
NNP-1	Bush	Noriega	Peters
NNP-2	J.	E.	L.
NNP-3	York	Francisco	Street
NNP-4	Inc	Exchange	Co
NNP-5	Inc.	Corp.	Co.
NNP-6	Stock	Exchange	York
NNP-7	Corp.	Inc.	Group
NNP-8	Congress	Japan	IBM
NNP-9	Friday	September	August
NNP-10	Shearson	D.	Ford
NNP-11	U.S.	Treasury	Senate
NNP-12	John	Robert	James
NNP-13	Mr.	Ms.	President
NNP-14	Oct.	Nov.	Sept.
NNP-15	New	San	Wall

JJS

JJS-0	largest	latest	biggest
JJS-1	least	best	worst
JJS-2	most	Most	least

DT

DT-0	the	The	a
DT-1	A	An	Another
DT-2	The	No	This
DT-3	The	Some	These
DT-4	all	those	some
DT-5	some	these	both
DT-6	That	This	each
DT-7	this	that	each
DT-8	the	The	a
DT-9	no	any	some
DT-10	an	a	the
DT-11	a	this	the

CD

CD-0	1	50	100
CD-1	8.50	15	1.2
CD-2	8	10	20
CD-3	1	30	31
CD-4	1989	1990	1988
CD-5	1988	1987	1990
CD-6	two	three	five
CD-7	one	One	Three
CD-8	12	34	14
CD-9	78	58	34
CD-10	one	two	three
CD-11	million	billion	trillion

PRP

PRP-0	It	He	I
PRP-1	it	he	they
PRP-2	it	them	him

RBR

RBR-0	further	lower	higher
RBR-1	more	less	More
RBR-2	earlier	Earlier	later

IN

IN-0	In	With	After
IN-1	In	For	At
IN-2	in	for	on
IN-3	of	for	on
IN-4	from	on	with
IN-5	at	for	by
IN-6	by	in	with
IN-7	for	with	on
IN-8	If	While	As
IN-9	because	if	while
IN-10	whether	if	That
IN-11	that	like	whether
IN-12	about	over	between
IN-13	as	de	Up
IN-14	than	ago	until
IN-15	out	up	down

RB

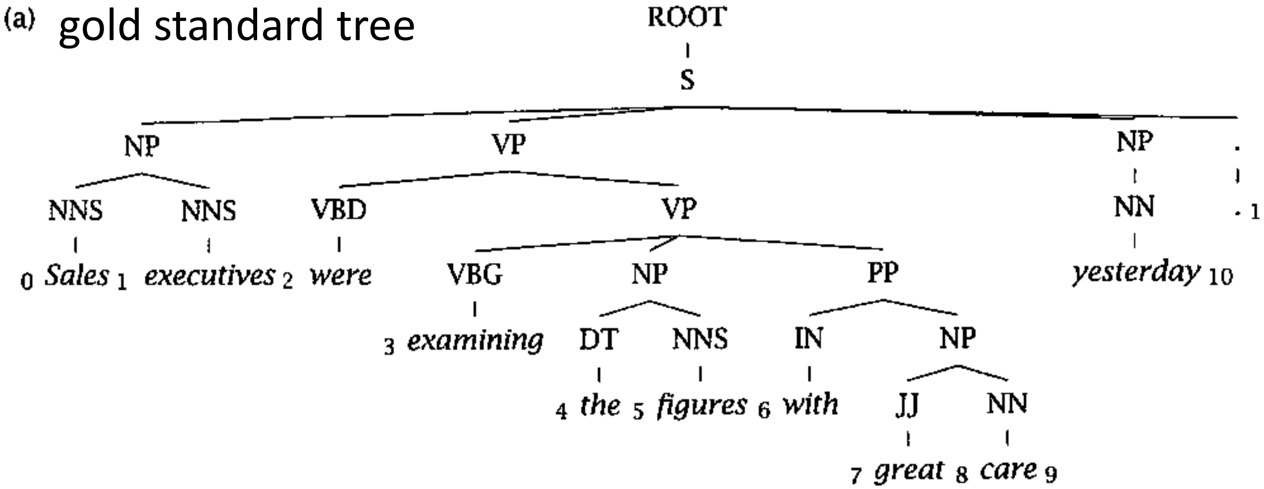
RB-0	recently	previously	still
RB-1	here	back	now
RB-2	very	highly	relatively
RB-3	so	too	as
RB-4	also	now	still
RB-5	however	Now	However
RB-6	much	far	enough
RB-7	even	well	then
RB-8	as	about	nearly
RB-9	only	just	almost
RB-10	ago	earlier	later
RB-11	rather	instead	because
RB-12	back	close	ahead
RB-13	up	down	off
RB-14	not	Not	maybe
RB-15	n't	not	also

[Petrov+ 2006]

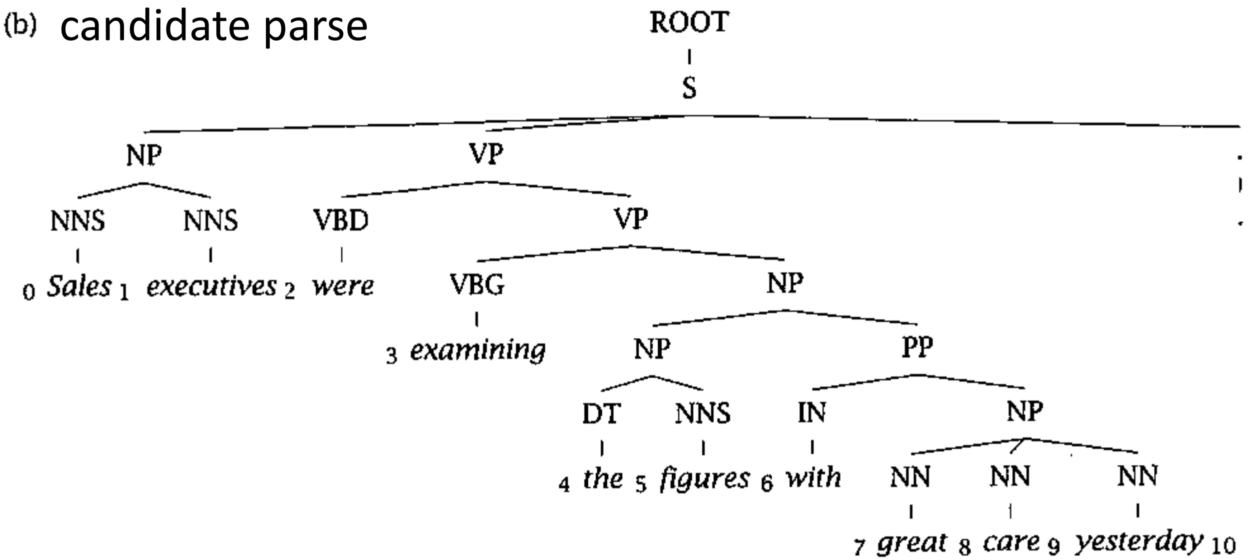
Evaluation

- Precision
- Recall
- Labeled Precision
- Labeled Recall

(a) gold standard tree



(b) candidate parse



(c) Brackets in gold standard tree (a.):
S-(0:11), NP-(0:2), VP-(2:9), VP-(3:9), **NP-(4:6)**, PP-(6-9), NP-(7,9), *NP-(9:10)

(d) Brackets in candidate parse (b.):
S-(0:11), NP-(0:2), VP-(2:10), VP-(3:10), NP-(4:10), **NP-(4:6)**, PP-(6-10), NP-(7,10)

(e)

Precision:	3/8 = 37.5%	Crossing Brackets:	0
Recall:	3/8 = 37.5%	Crossing Accuracy:	100%
Labeled Precision:	3/8 = 37.5%	Tagging Accuracy:	10/11 = 90.9%
Labeled Recall:	3/8 = 37.5%		

Performance on English (supervised)

	LP	LR	F1
[Magerman 1995]	84.0	84.3	84.2
[Charniak 1997]	86.7	86.6	86.7
[Collins 1997]	87.5	88.1	87.8
[Charniak 2000]	89.6	89.5	89.6
[Petrov&Klein 2007]	90.2	89.9	90.1
[Carreras+ 2008]	90.7	91.4	91.1
[Shindo+ 2012]			91.1
[Zhu+ 2013]	90.2	90.7	90.4

	LP	LR	F1
[Socher+ 2013]			90.4
[Watanabe+ 2015]			90.7
[Mi&Huang 2015]	90.7	90.9	90.8
[Cross&Huang 2016]	90.5	92.1	91.3
[Dyer+ 2016]			91.7
[Stern+ 2017]	90.6	93.0	91.8
[Stern+ 2017]	92.6	92.6	92.6
[Gaddy+ 2018]			92.1
[Kitaev&Klein 2018]	93.2	93.9	93.6

Performance on English (semi-supervised, reranking, etc.)

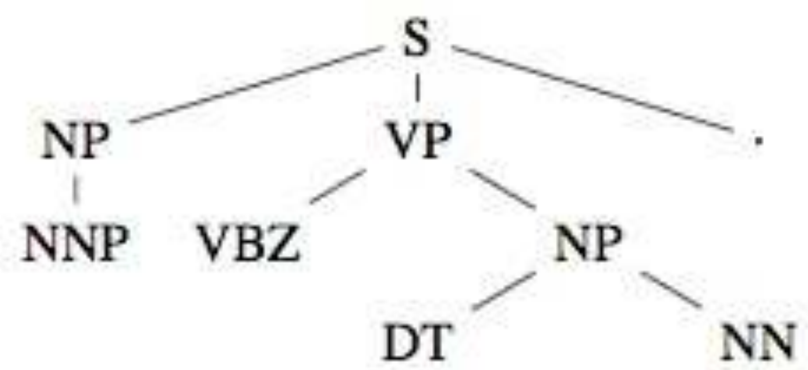
	F1	Note
[Charniak&Johnson 2005]	91.0	reranking
[McClosky+ 2006]	92.1	self-training
[Shindo+ 2012]	92.4	ensemble
[Vinyals+ 2015]	92.8	tri-training
[Dyer+ 2016]	92.4	reranking
[Choe&Charniak 2016]	93.8	tri-training
[Kuncoro+ 2017]	93.6	reranking
[Liu&Zhang 2017]	94.2	reranking/tri-training
[Fried+ 2017]	94.7	ensemble/reranking

Performance on English (pre-trained)

	F1	Note
[Kitaev&Klein 2018]	95.1	pre-train (ELMo)
[Kitaev+ 2019]	95.6	pre-train (BERT)
[Zhou&Zhao 2019]	96.3	pre-train (XLNet)
[Yang&Deng 2020]	96.3	pre-train (XLNet)
[Mrini+ 2020]	96.4	pre-train (XLNet)

Serialization [Vinyals+ 2015]

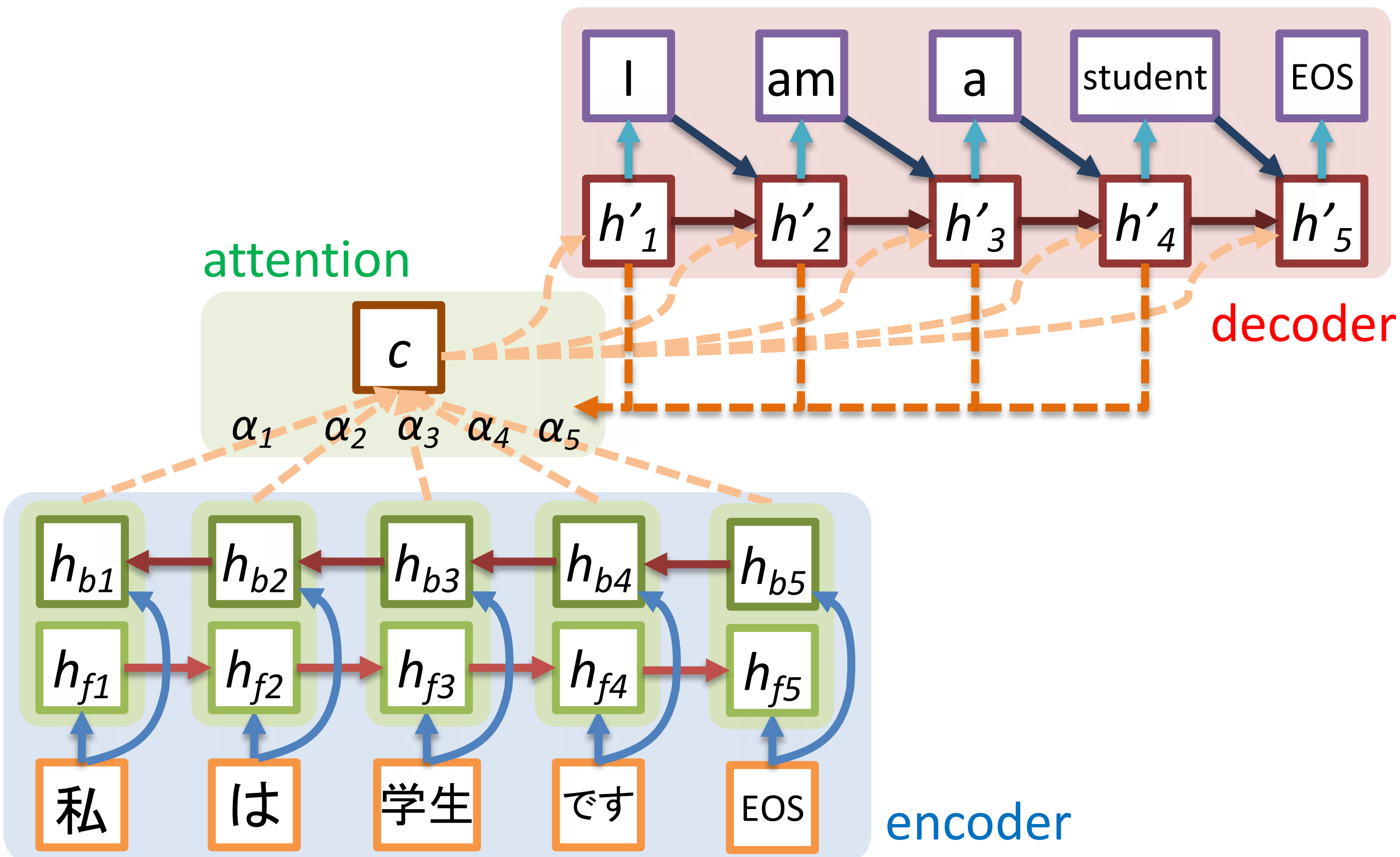
John has a dog . →



John has a dog . → (S (NP NNP $)_{NP}$ (VP VBZ (NP DT NN $)_{NP}$ $)_{VP}$. $)_{S}$

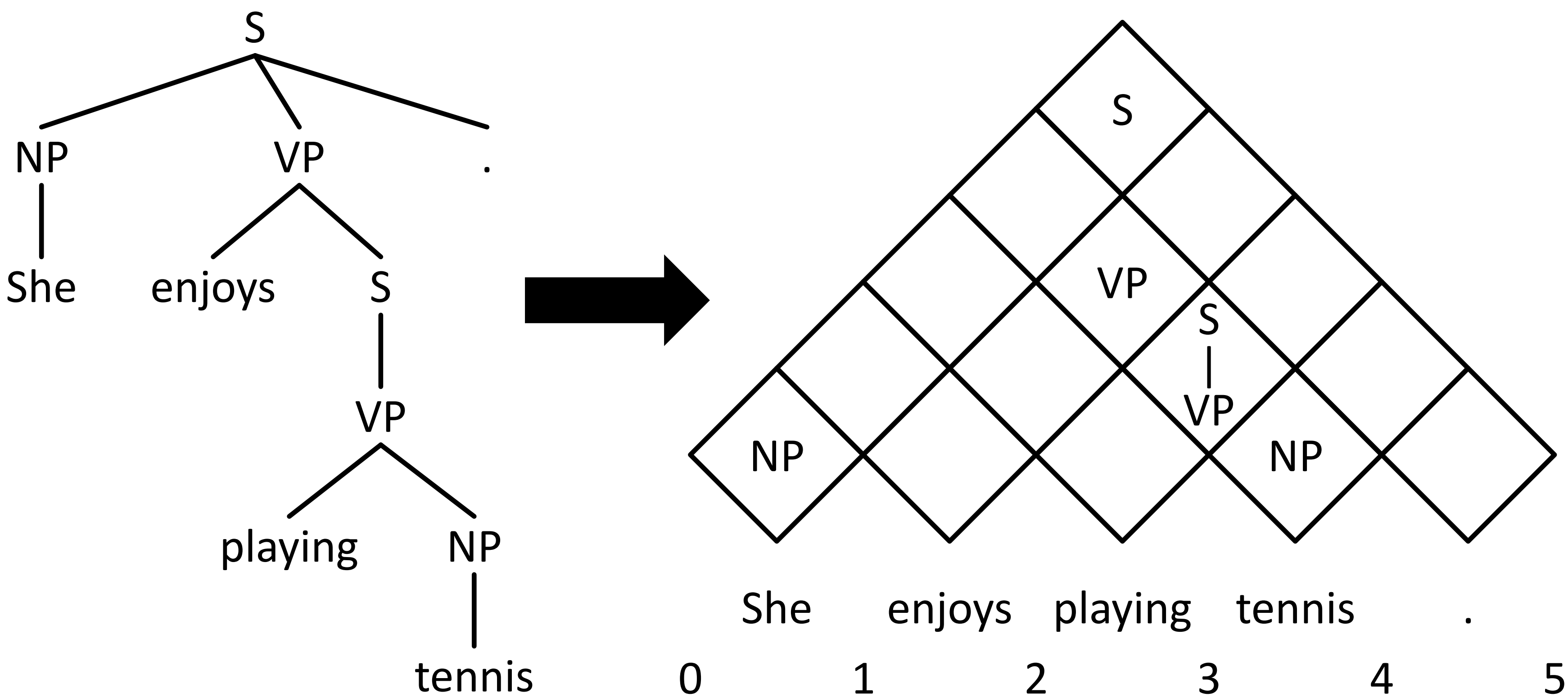
Attention-based Neural Machine Translation

[Bahdanau+ 2014]

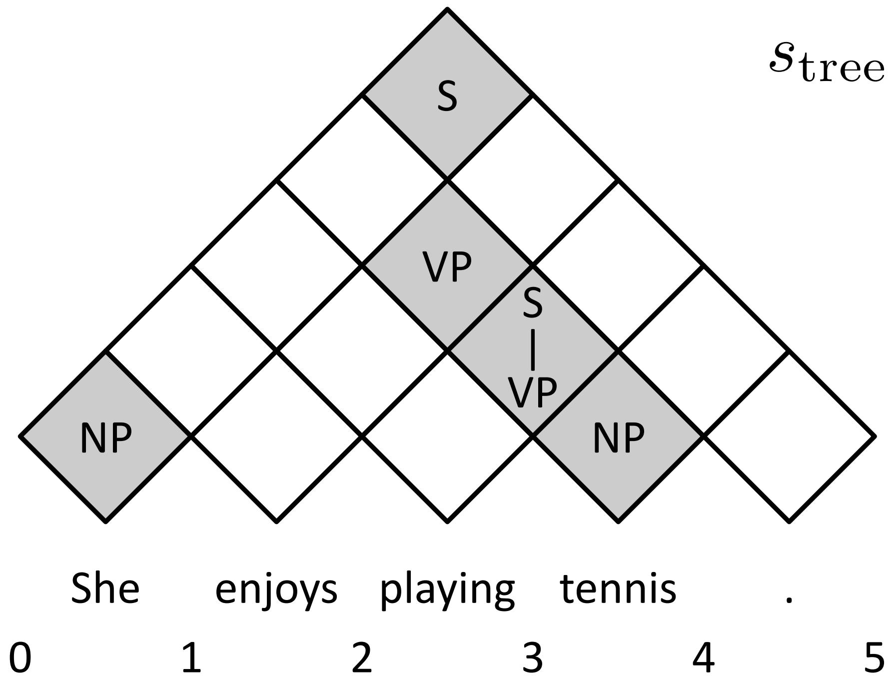


Neural Span Classification

[Stern+ 2017]



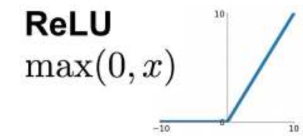
Neural Span Classification



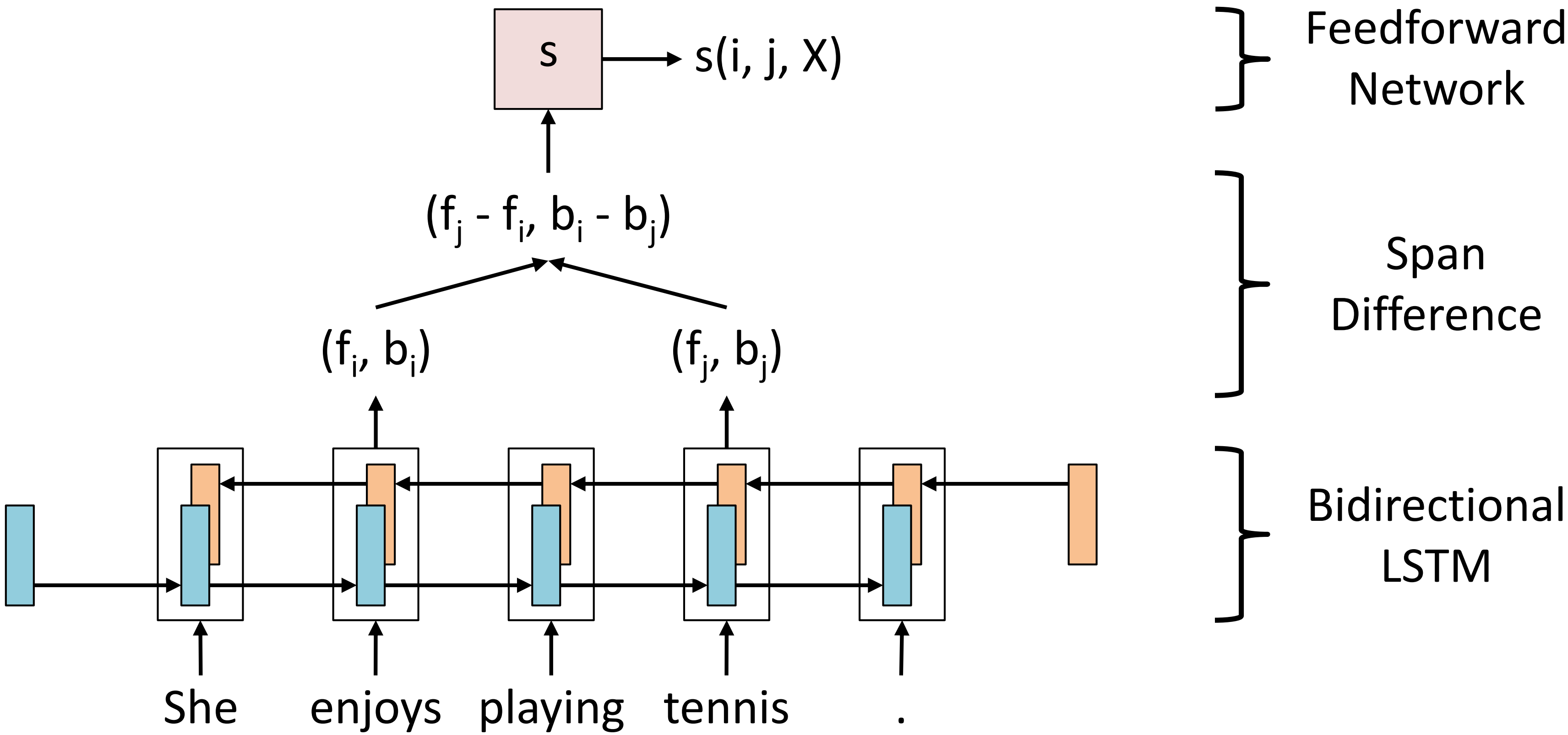
$$s_{\text{tree}}(T) = \sum_{(\ell,(i,j)) \in T} s(i,j,\ell)$$
$$= s(0,5,\text{S})$$
$$+ \; s(0,1,\text{NP})$$
$$+ \; s(1,4,\text{VP})$$
$$+ \; s(2,4,\text{S–VP})$$
$$+ \; s(3,4,\text{NP})$$

$$s_{\text{best}}(i,j) = \overbrace{\max_{\ell}[s(i,j,\ell)]}^{\text{Pick best label}} + \underbrace{\max_{k}[s_{\text{best}}(i,k) + s_{\text{best}}(k,j)]}_{\text{Pick best split point}}$$

Neural Span Classification



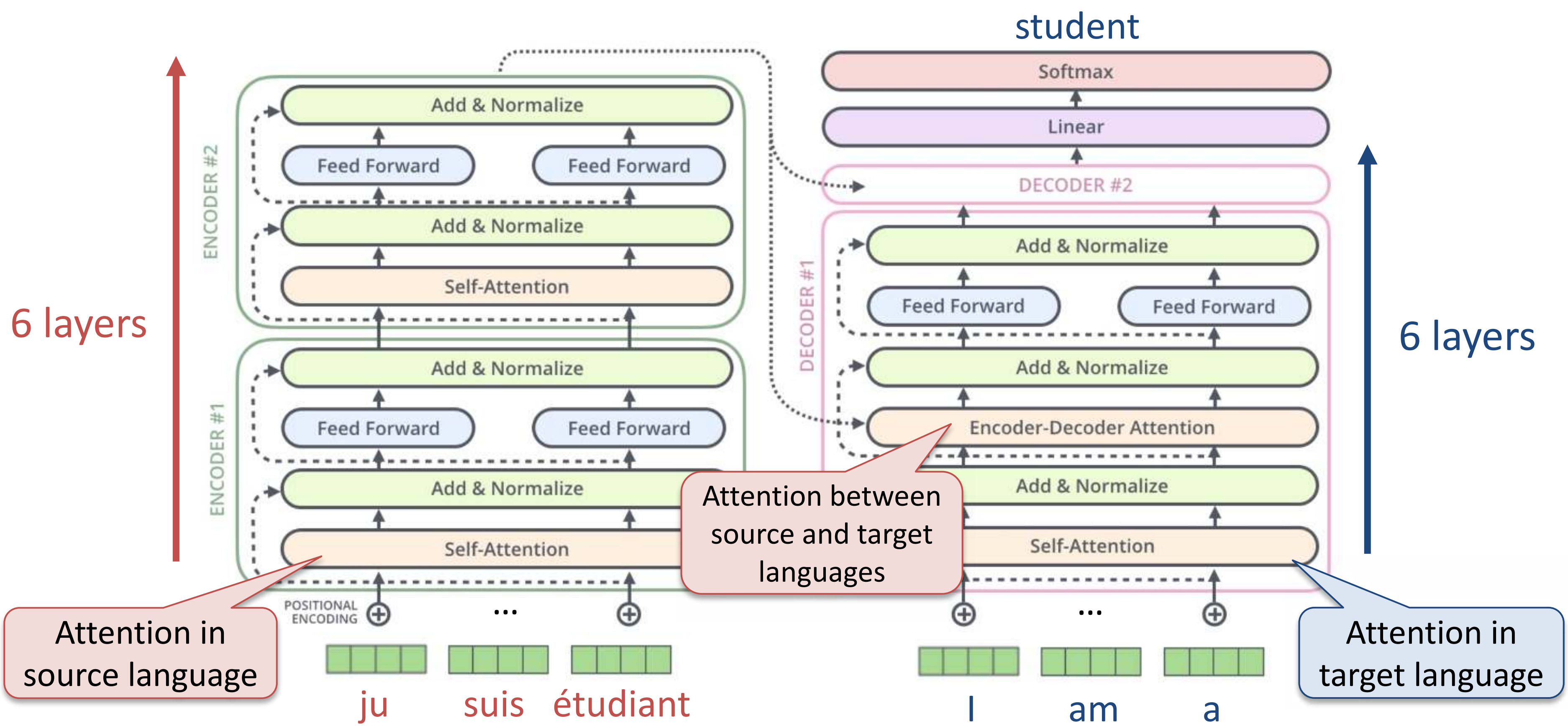
$$\mathrm{FeedForward}(x) = W_2 \mathrm{relu}(W_1 x + b_1) + b_2$$



Transformer: “Attention is All You Need”

[Vaswani+ 2017]

Output: next word in target language



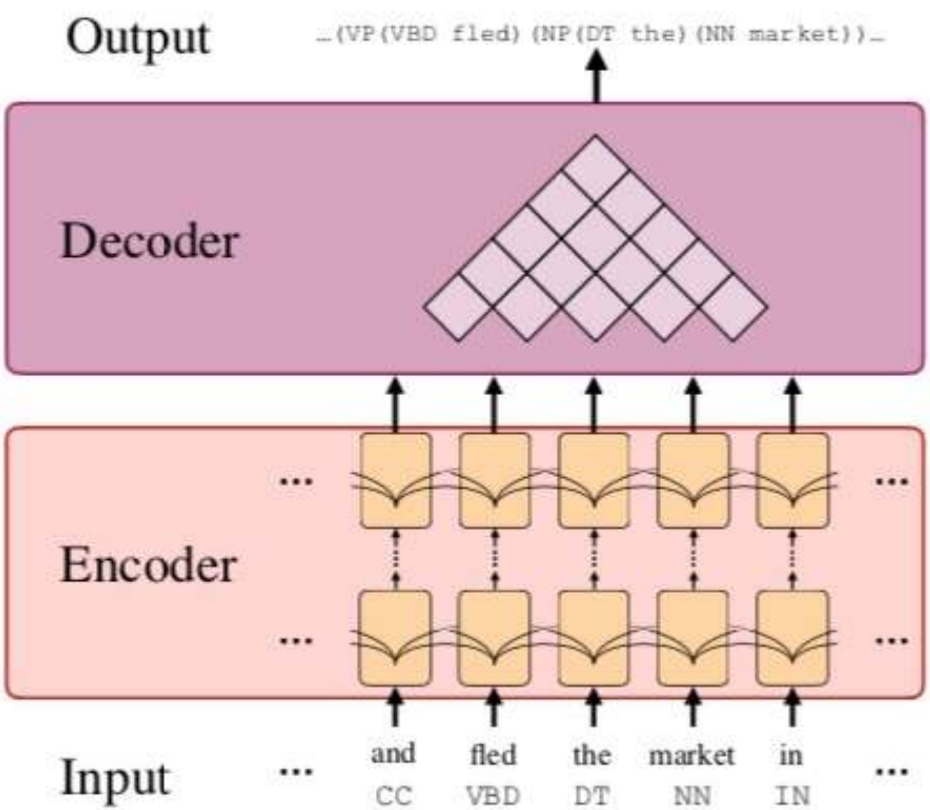
Input: sentence in source language

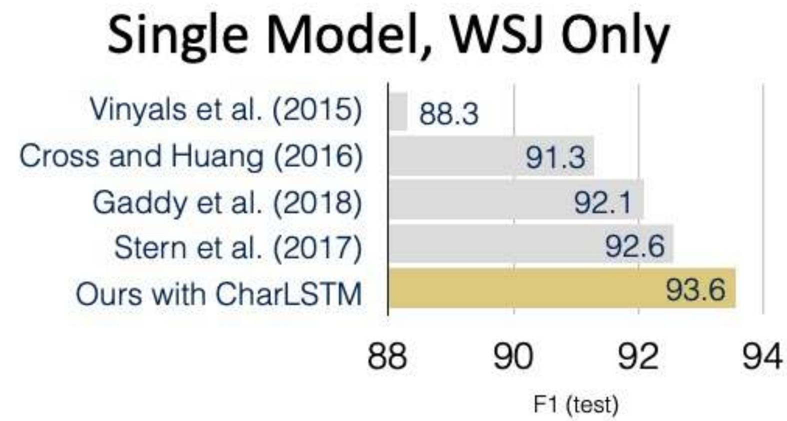
Words previously generated

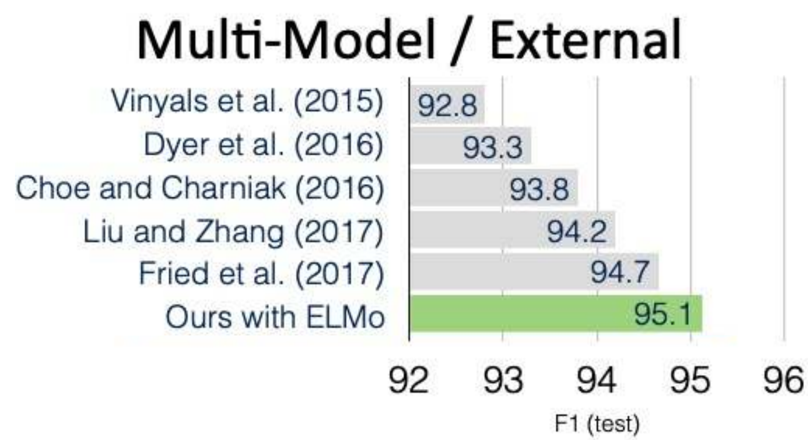
https://jalammar.github.io/illustrated-transformer/

Using Self Attention

[Kitaev&Klein 2018]

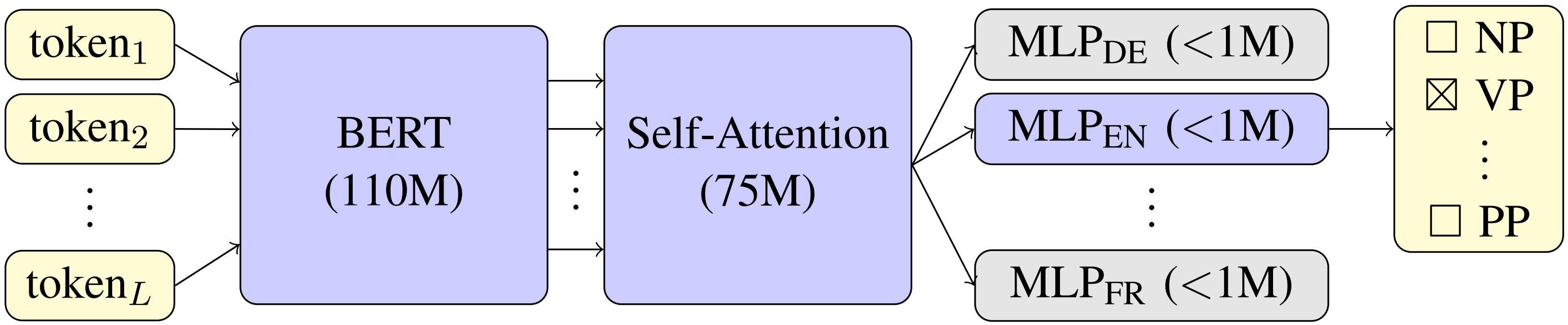


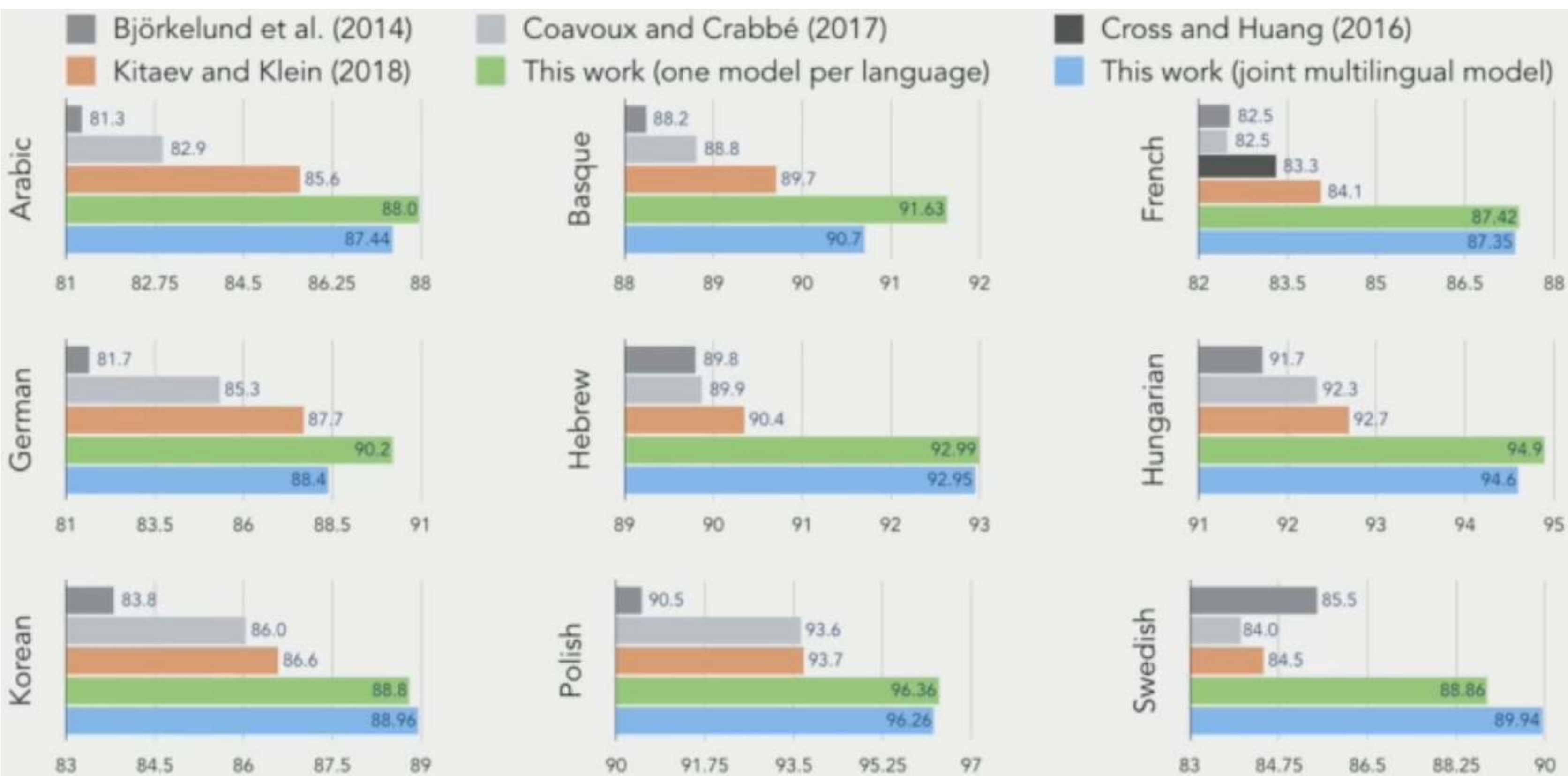




Joint Multilingual Span Classification

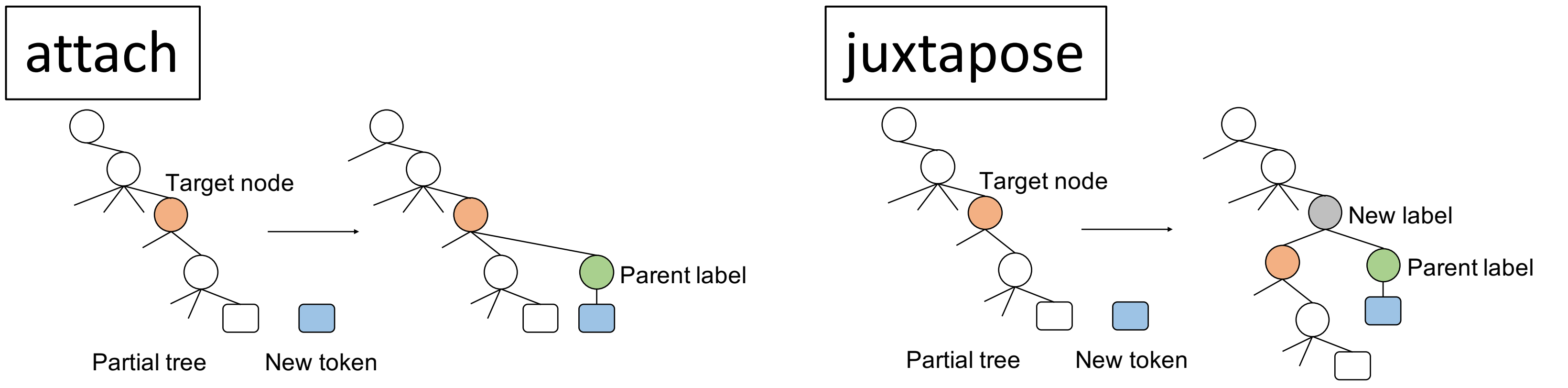
[Kitaev+ 2019]



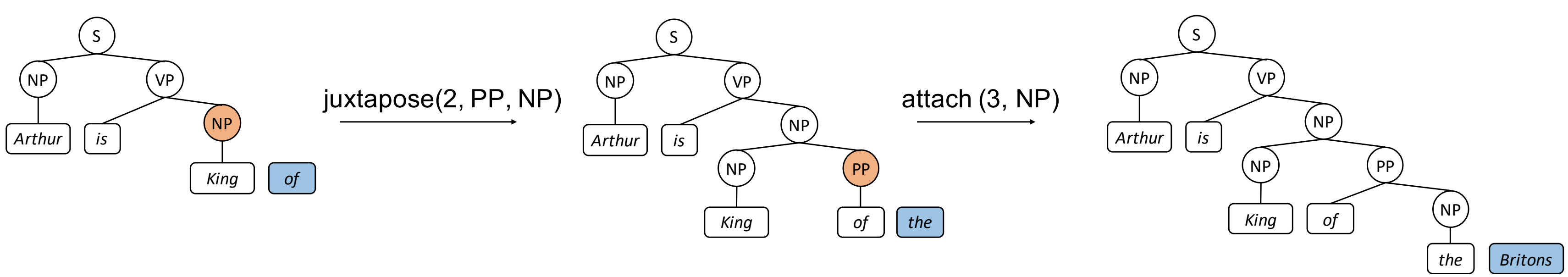


Transition-based Constituency Parsing

- An attach-juxtapose parser [Yang&Deng 2020]

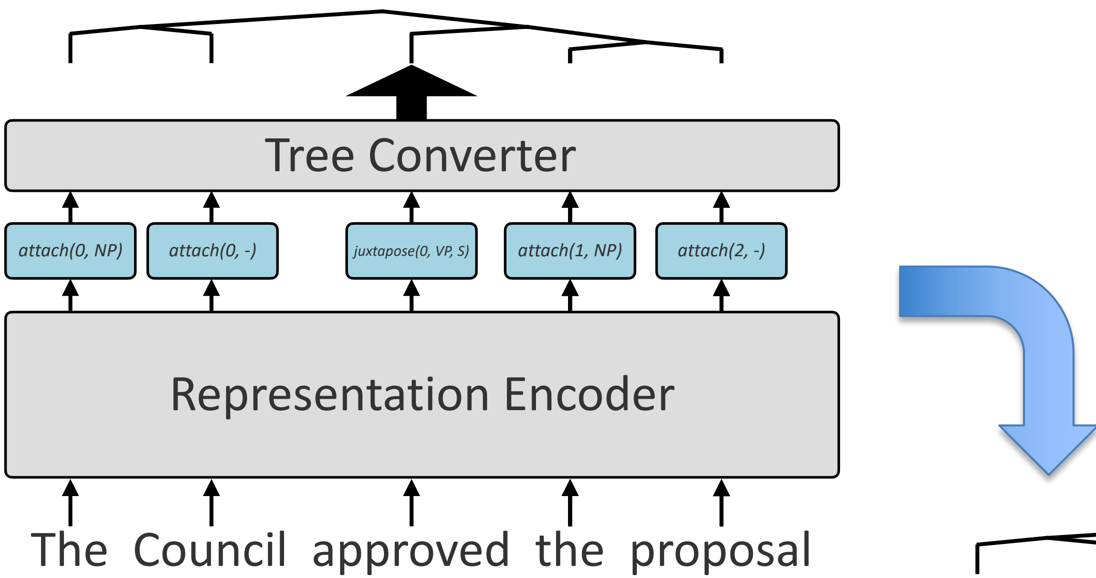


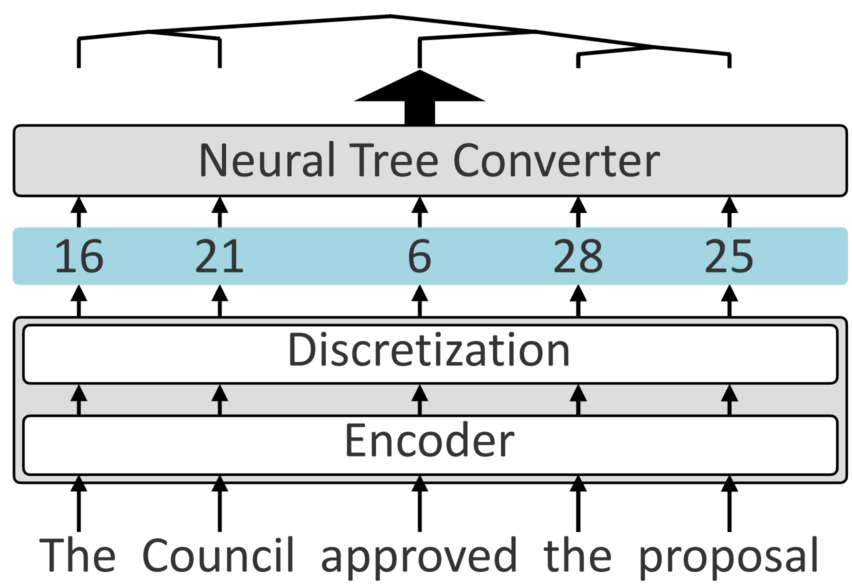
Example:



Learning Representations for Parsing

[Kitaev+ 2022]





	Encoder Type		
	Bi (↔)		Uni (→)
Representation	BERT	GPT-2	GPT-2
Span Classification (Kitaev et al., 2019)	95.59	**95.10**†	93.95†
Attach-Juxtapose (Yang and Deng, 2020)	**95.79**	94.53†	87.66†
Learned (This work)	95.55	–	**94.97**

Summary

- Review of CFG and CKY parsing
- Probabilistic CFG
- Treebanks
- Extensions
  - Lexicalization
  - History
  - Nonterminal Classification
- Evaluation criteria and SOTA of English parsing