CS395T: Structured Models for NLP
Lecture 19: Advanced NNs I

Greg Durrett

Administrivia

- Kyunghyun Cho (NYU) talk Friday 11am GDC 6.302
- Project 3 due today!
- Final project out today!
  - Proposal due in 1 week
  - Project presentations December 5/7 (timeslots to be assigned when proposals are turned in)
  - Final project due December 15 (no slip days!)

Project Proposals

- ~1 page
  - Define a problem, give context of related work (at least 3-4 relevant papers)
  - Propose a direction that you think is feasible and outline steps to get there, including what dataset you'll use
- Okay to change directions after the proposal is submitted, but run it by me if it's a big change

This Lecture

- Neural CRFs

- Tagging / NER

- Parsing

Neural CRF Basics

NER Revisited

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

PERSON LOC ORG

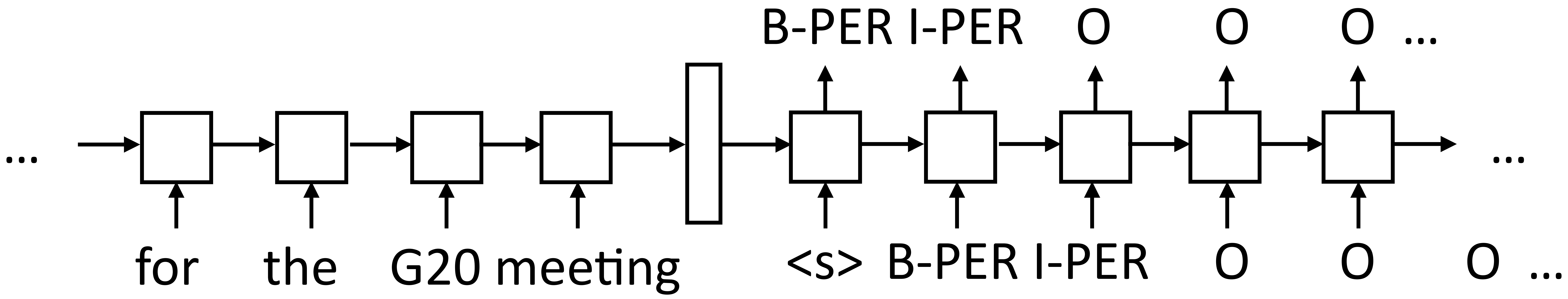
- Features in CRFs: I[tag=B-LOC & curr_word=*Hangzhou*], I[tag=B-LOC & prev_word=*to*], I[tag=B-LOC & curr_prefix=*Han*]
- Linear model over features
- Downsides:
  - Lexical features mean that words need to be seen in the training data
  - Can only use limited context windows
  - Linear model can't capture feature conjunctions effectively

LSTMs for NER

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

PERSON LOC ORG



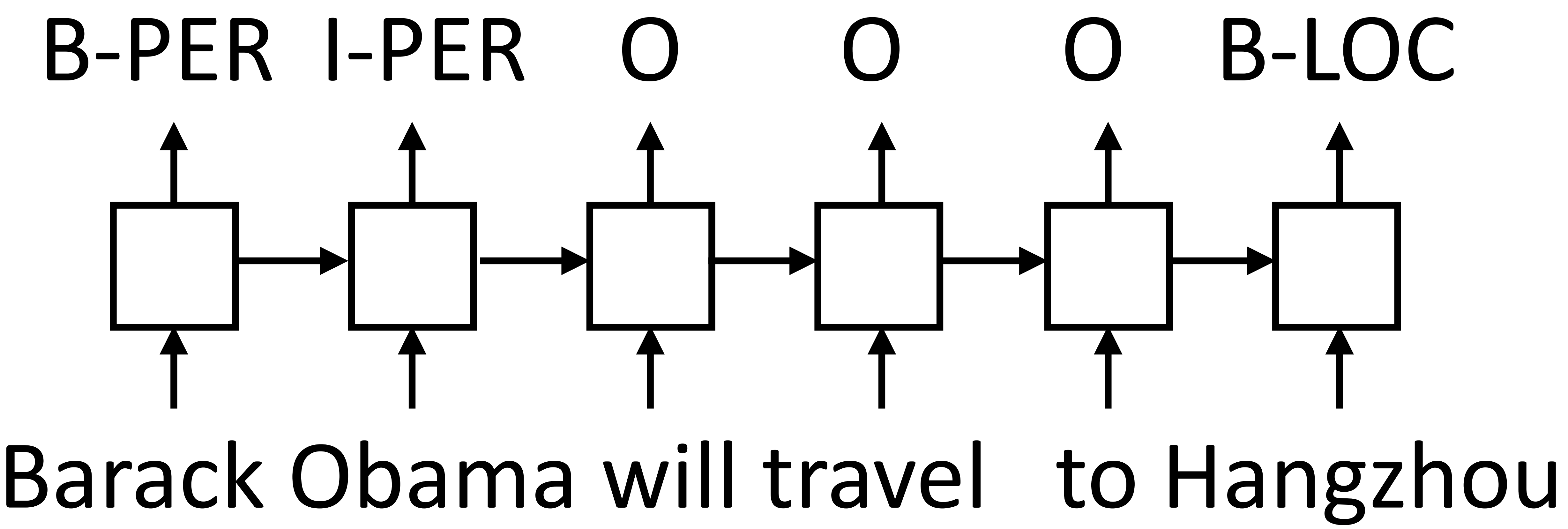
- Encoder-decoder (MT-like model)
- What are the strengths and weaknesses of this model compared to CRFs?

LSTMs for NER

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

PERSON LOC ORG



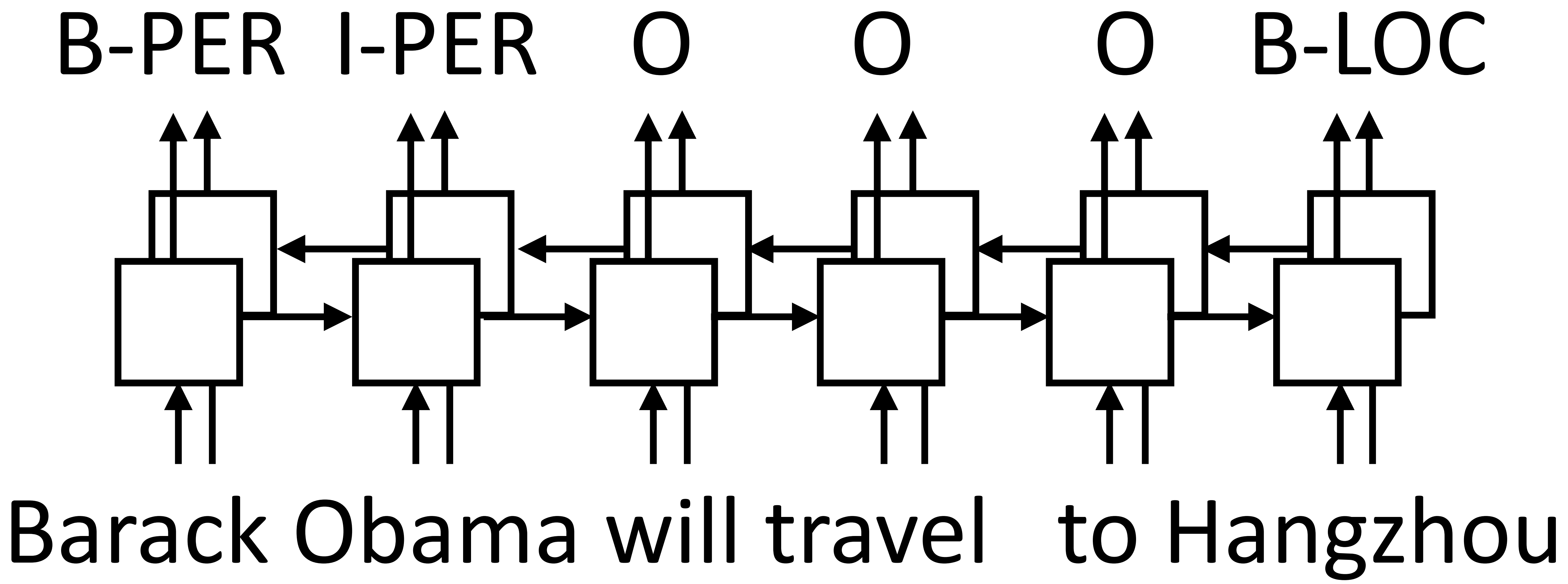
- Transducer (LM-like model)
- What are the strengths and weaknesses of this model compared to CRFs?

LSTMs for NER

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

PERSON LOC ORG



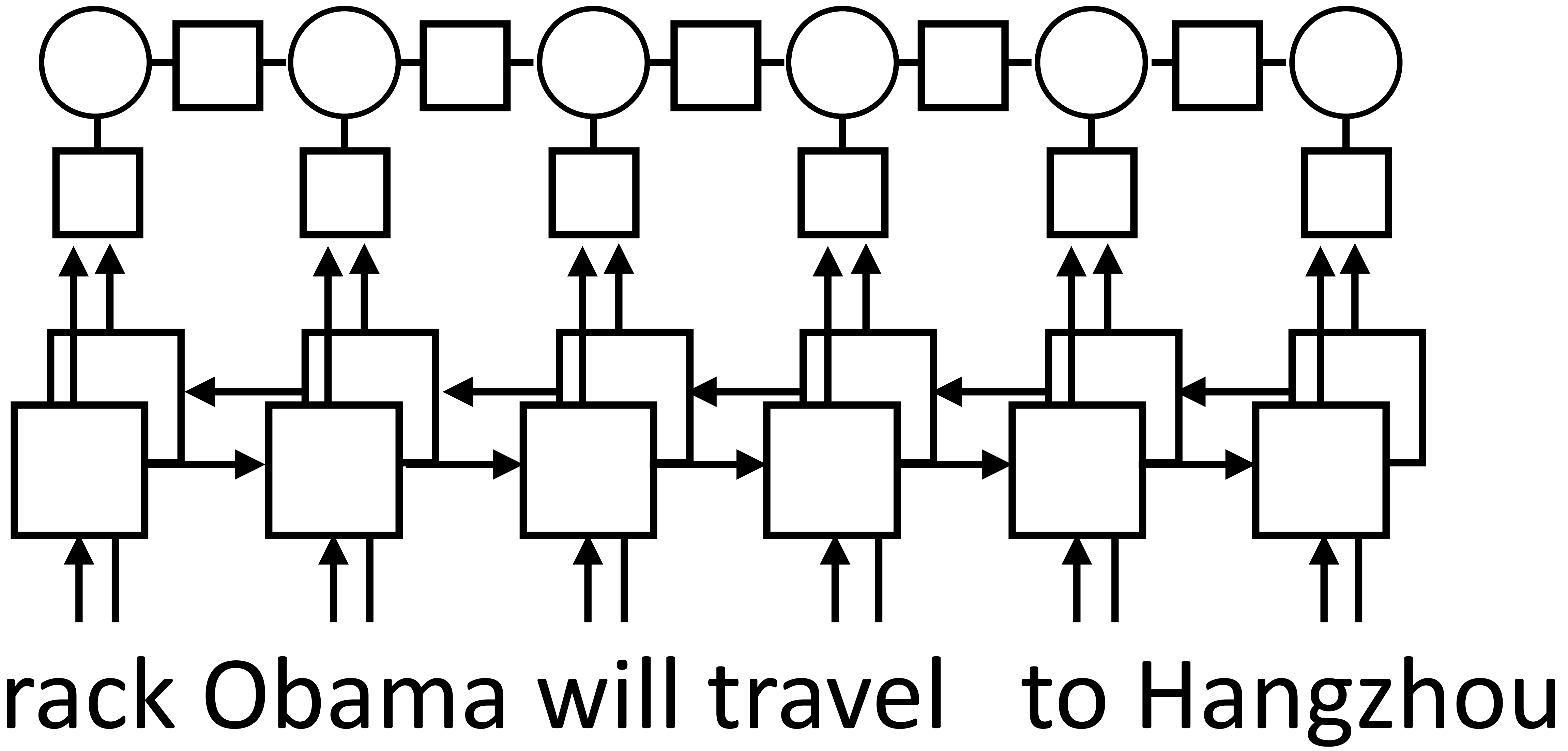
- Bidirectional transducer model
- What are the strengths and weaknesses of this model compared to CRFs?

Neural CRFs

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

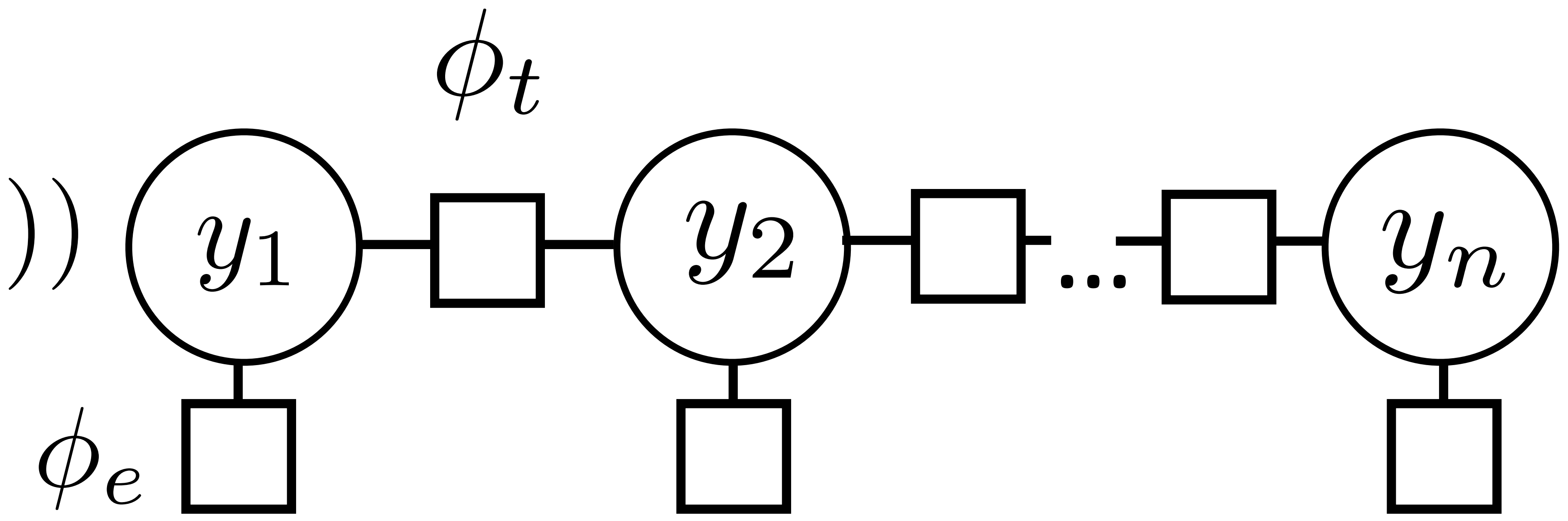
PERSON LOC ORG



- Neural CRFs: bidirectional LSTMs (or some NN) compute emission potentials, capture structural constraints in transition potentials

Neural CRFs

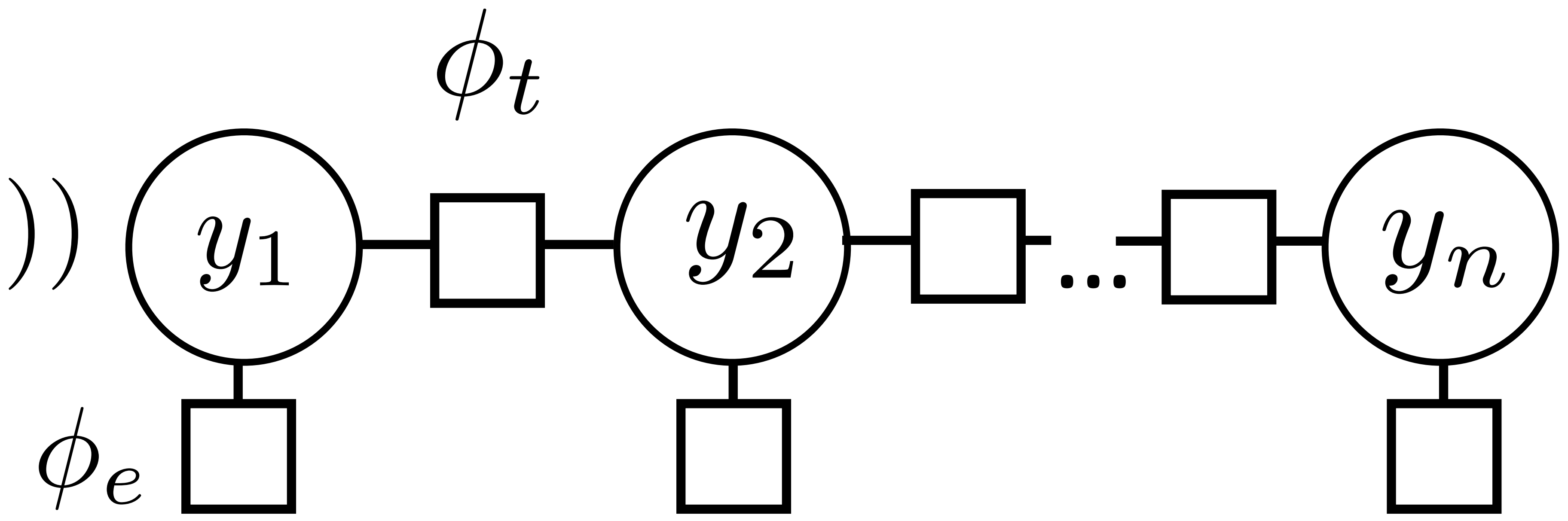
$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\prod_{i=2}^{n}\exp(\phi_t(y_{i-1}, y_i))\prod_{i=1}^{n}\exp(\phi_e(y_i, i, \mathbf{x}))$$



- Conventional: $\phi_e(y_i, i, \mathbf{x}) = w^\top f_e(y_i, i, \mathbf{x})$
- Neural: $\phi_e(\mathbf{y}, i, \mathbf{x}) = W f(i, \mathbf{x})$ f/phi are vectors, len(phi) = num labels
- *f*(*i*, ***x***) could be the output of a feedforward neural network looking at the words around position *i*, or the *i*th output of an LSTM, …
- Neural network computes unnormalized potentials that are consumed and “normalized” by a structured model
- Inference: compute *f*, use Viterbi (or beam)

Computing Gradients

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\prod_{i=2}^{n}\exp(\phi_t(y_{i-1}, y_i))\prod_{i=1}^{n}\exp(\phi_e(y_i, i, \mathbf{x}))$$



- Conventional: $\phi_e(y_i, i, \mathbf{x}) = w^\top f_e(y_i, i, \mathbf{x})$
- Neural: $\phi_e(\mathbf{y}, i, \mathbf{x}) = W f(i, \mathbf{x})$

$$\frac{\partial \mathcal{L}}{\partial \phi_{e,i}} = -P(y_i = s|\mathbf{x}) + I[s \text{ is gold}]$$

"error signal", compute with F-B

- For linear model: $\dfrac{\partial \phi_{e,i}}{w_i} = f_{e,i}(y_i, i, \mathbf{x})$

chain rule say to multiply together, gives our update

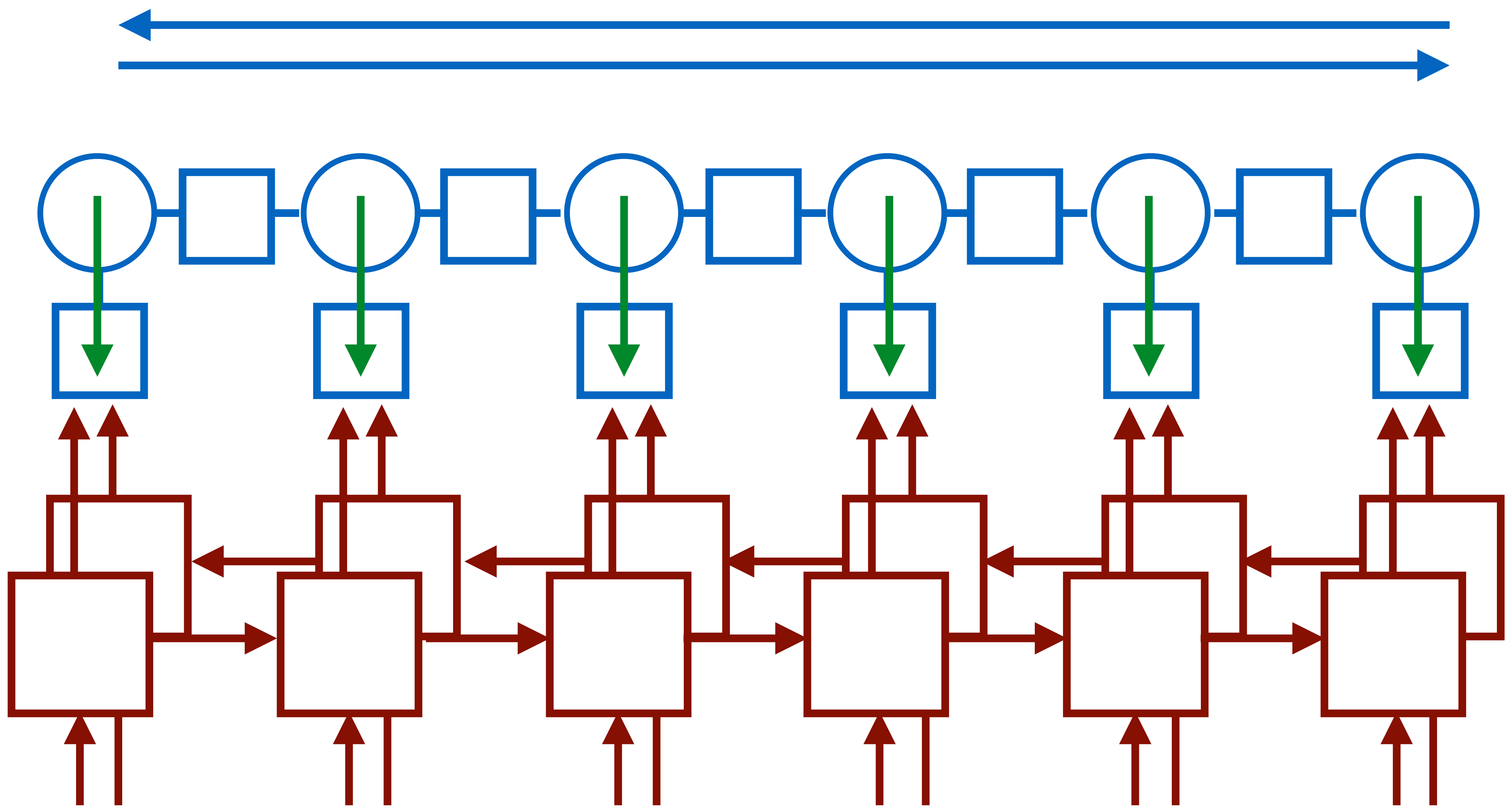
- For neural model: compute gradient of phi w.r.t. parameters of neural net

Neural CRFs

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

PERSON LOC ORG



2) Run forward-backward

3) Compute error signal

1) Compute f(**x**)

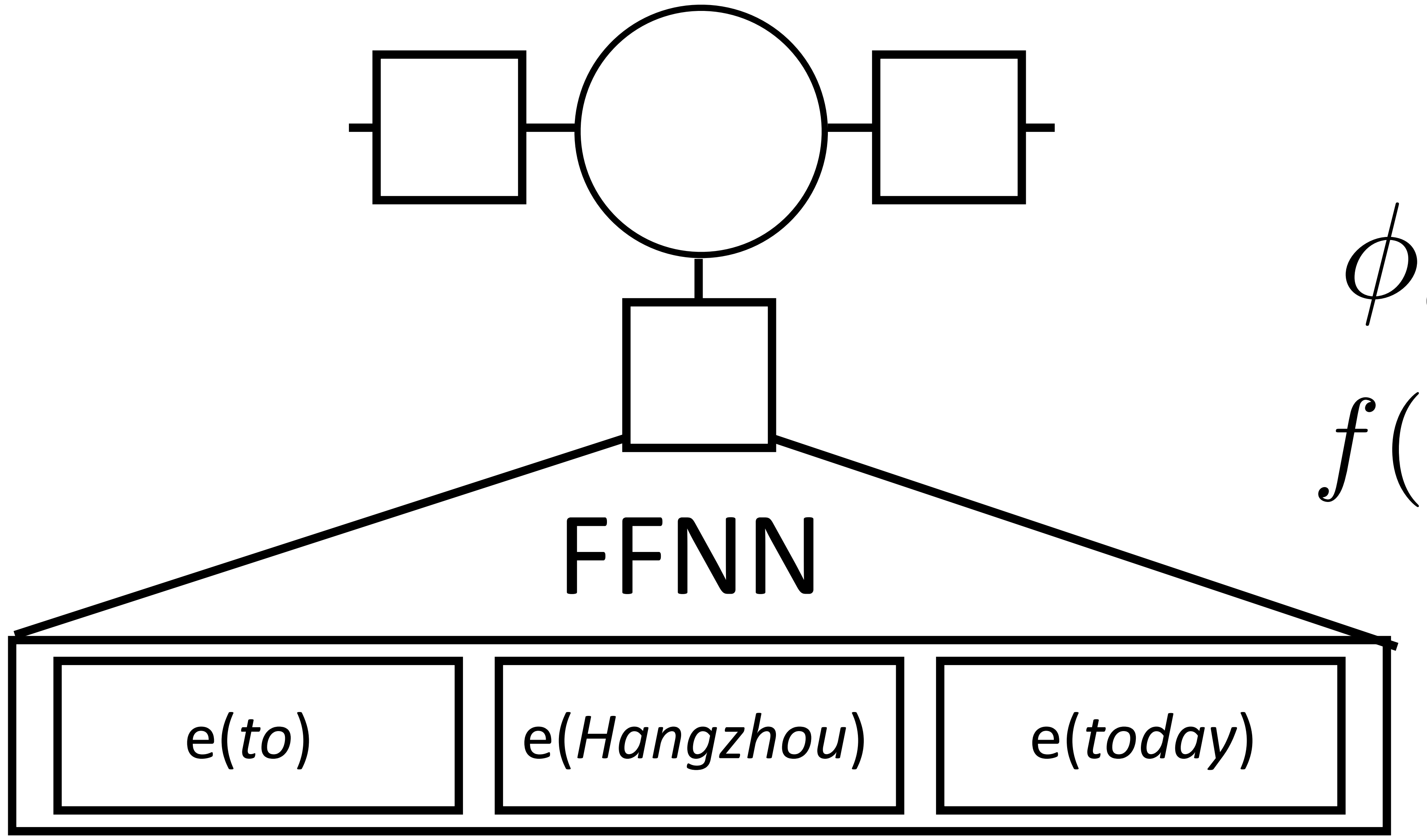
4) Backprop (no knowledge of sequential structure required)

FFNN Neural CRF for NER

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

PERSON LOC ORG



previous word curr word next word

to ***Hangzhou*** *today*

$$\phi_e = Wg(Vf(\mathbf{x}, i))$$

$$f(\mathbf{x}, i) = [\mathrm{emb}(\mathbf{x}_{i-1}), \mathrm{emb}(\mathbf{x}_i), \mathrm{emb}(\mathbf{x}_{i+1})]$$

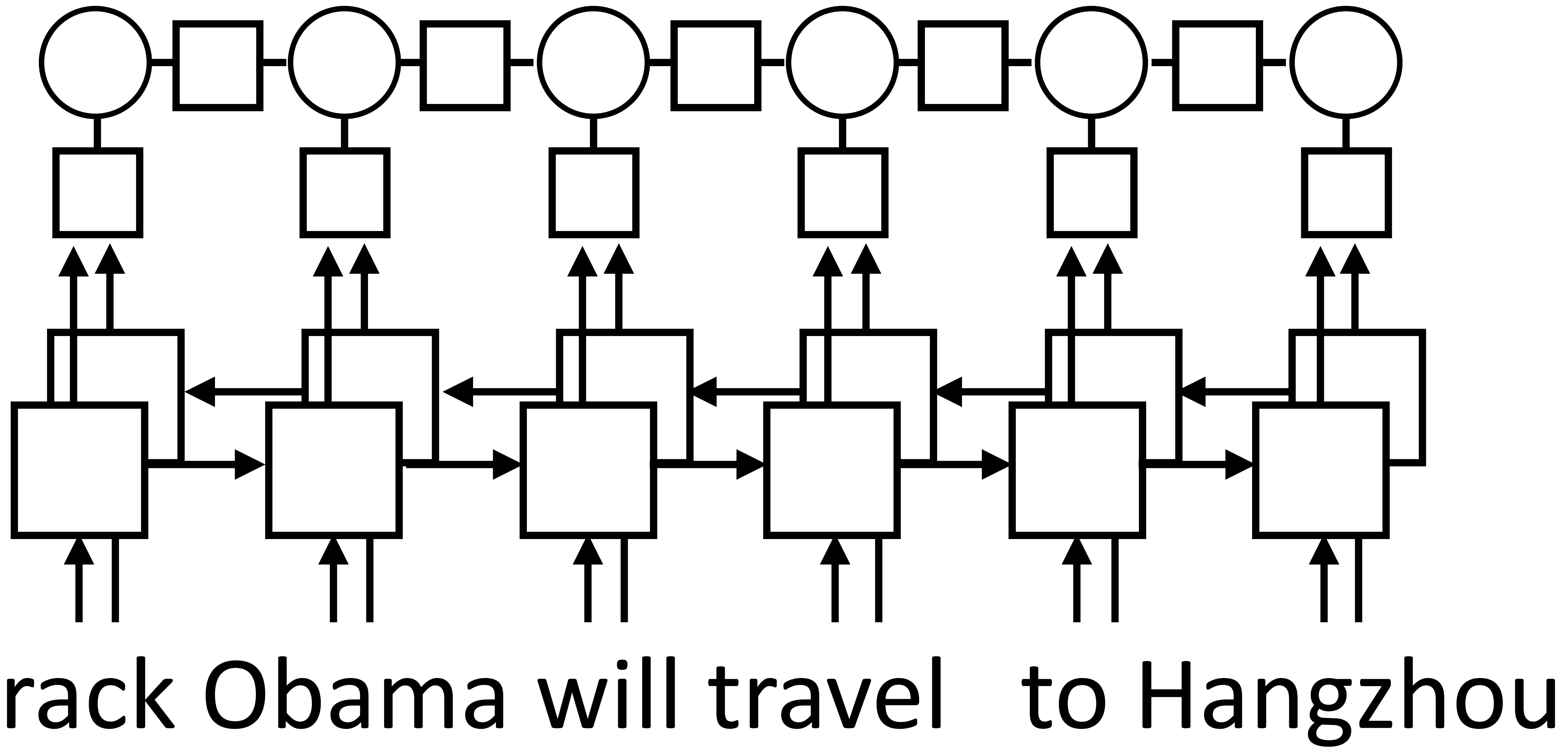
- Or *f*(**x**) looks at output of LSTM, or another model…

Neural CRFs

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

PERSON LOC ORG



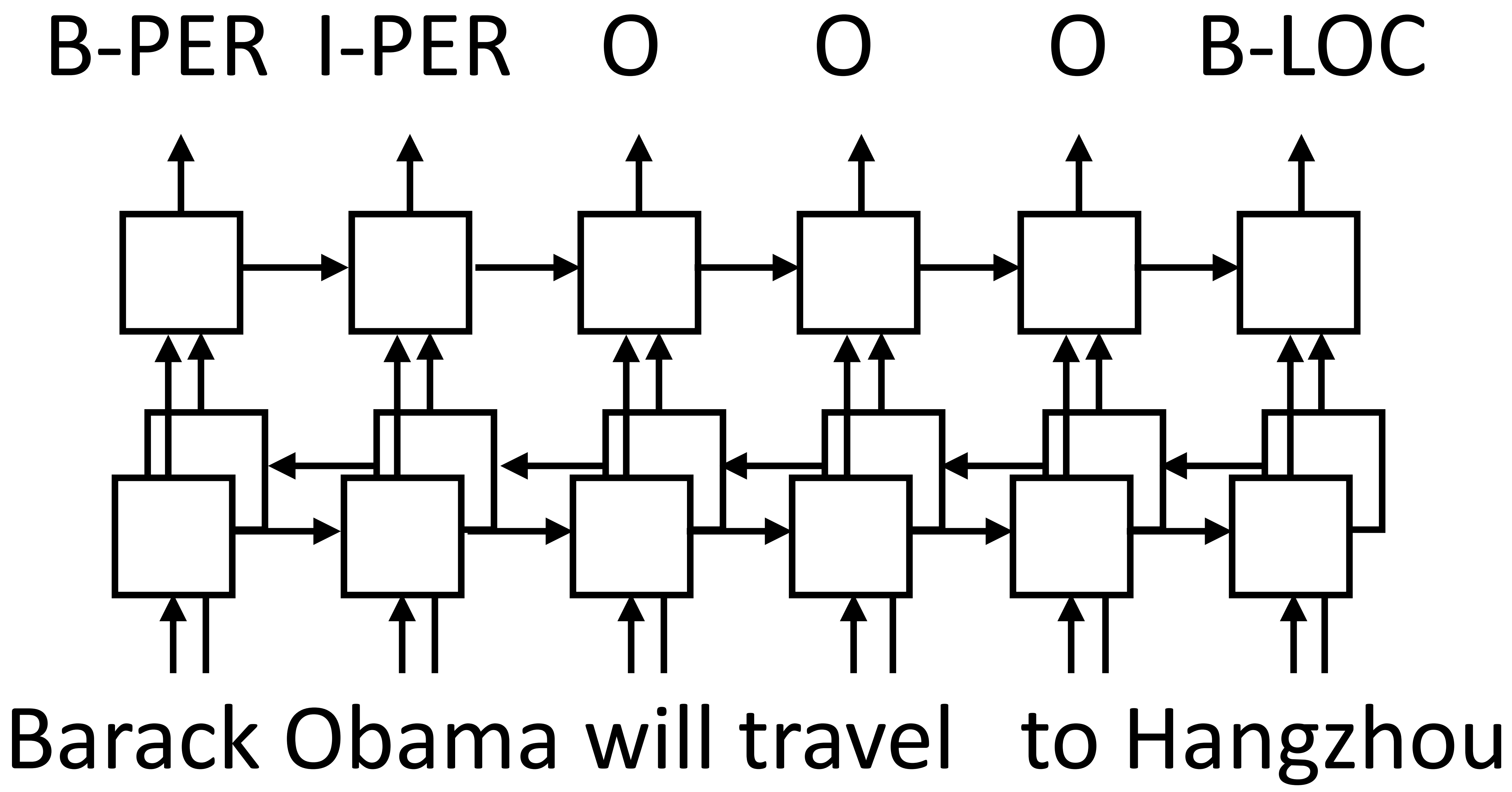
- Neural CRFs: bidirectional LSTMs compute emission potentials, also transition potentials (usually based on sparse features)

LSTMs for NER

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

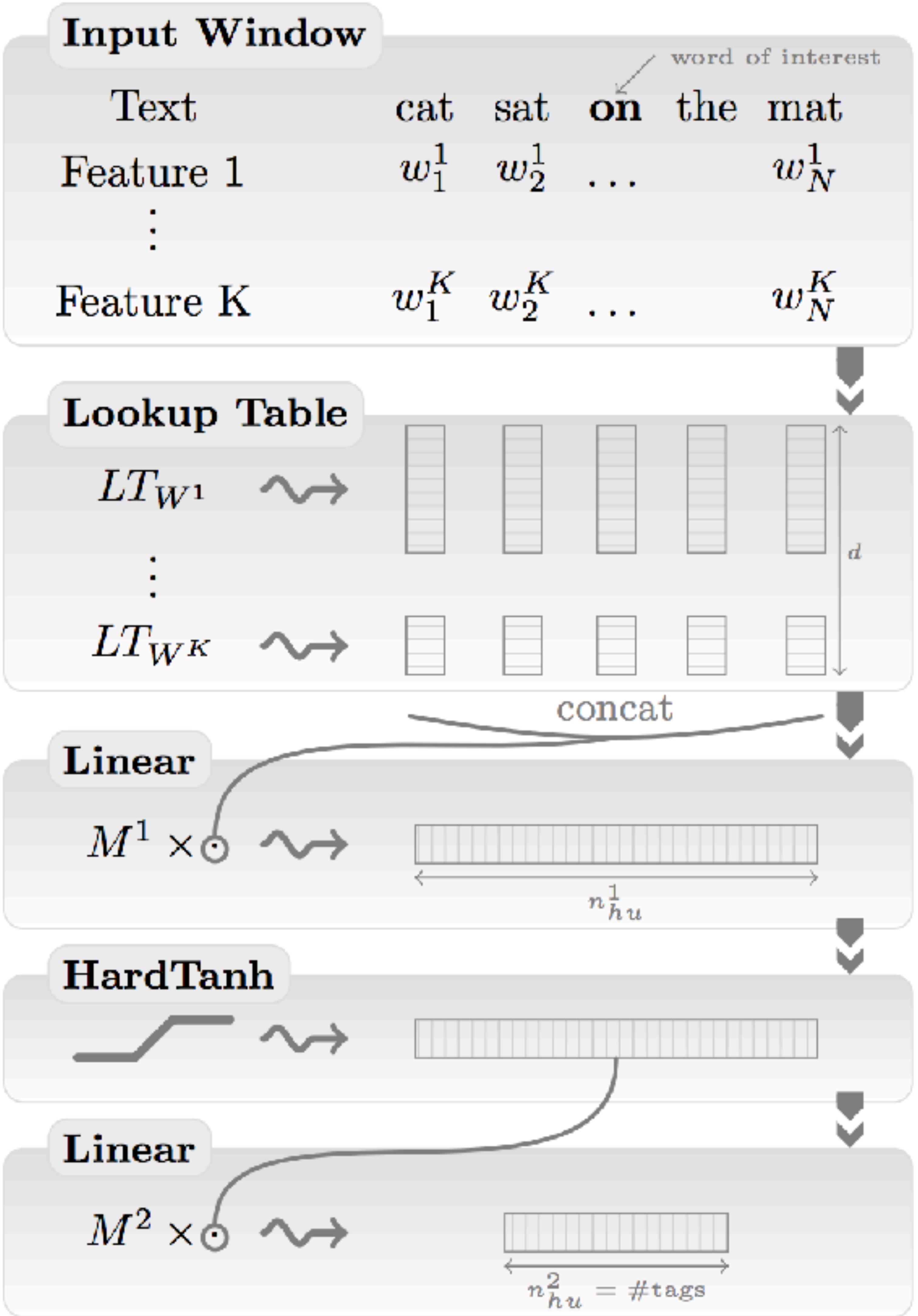
PERSON LOC ORG



- How does this compare to neural CRF?

"NLP (Almost) From Scratch"

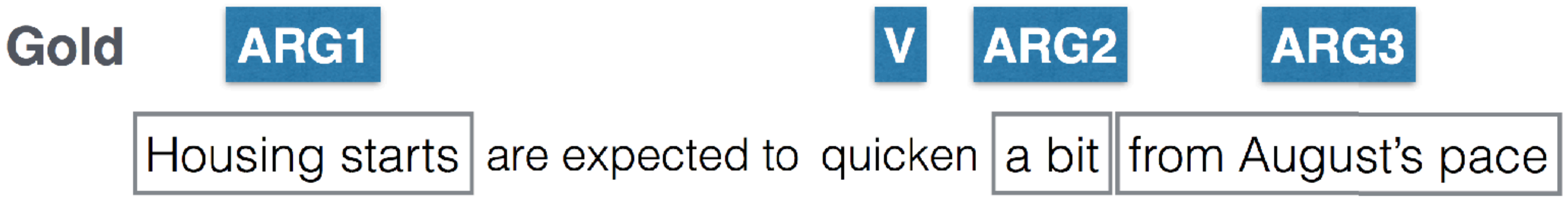
Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
NN+WLL	96.31	89.13	79.53	55.40
NN+SLL	96.37	90.33	81.47	70.99
NN+WLL+LM1	97.05	91.91	85.68	58.18
NN+SLL+LM1	97.10	93.65	87.58	73.84
NN+WLL+LM2	97.14	92.04	86.96	58.34
NN+SLL+LM2	97.20	93.63	88.67	74.15



- WLL: independent classification; SLL: neural CRF
- LM1/LM2: pretrained word embeddings from a language model over large corpora

Collobert, Weston, et al. 2008, 2011

How do we use a tagger for SRL?

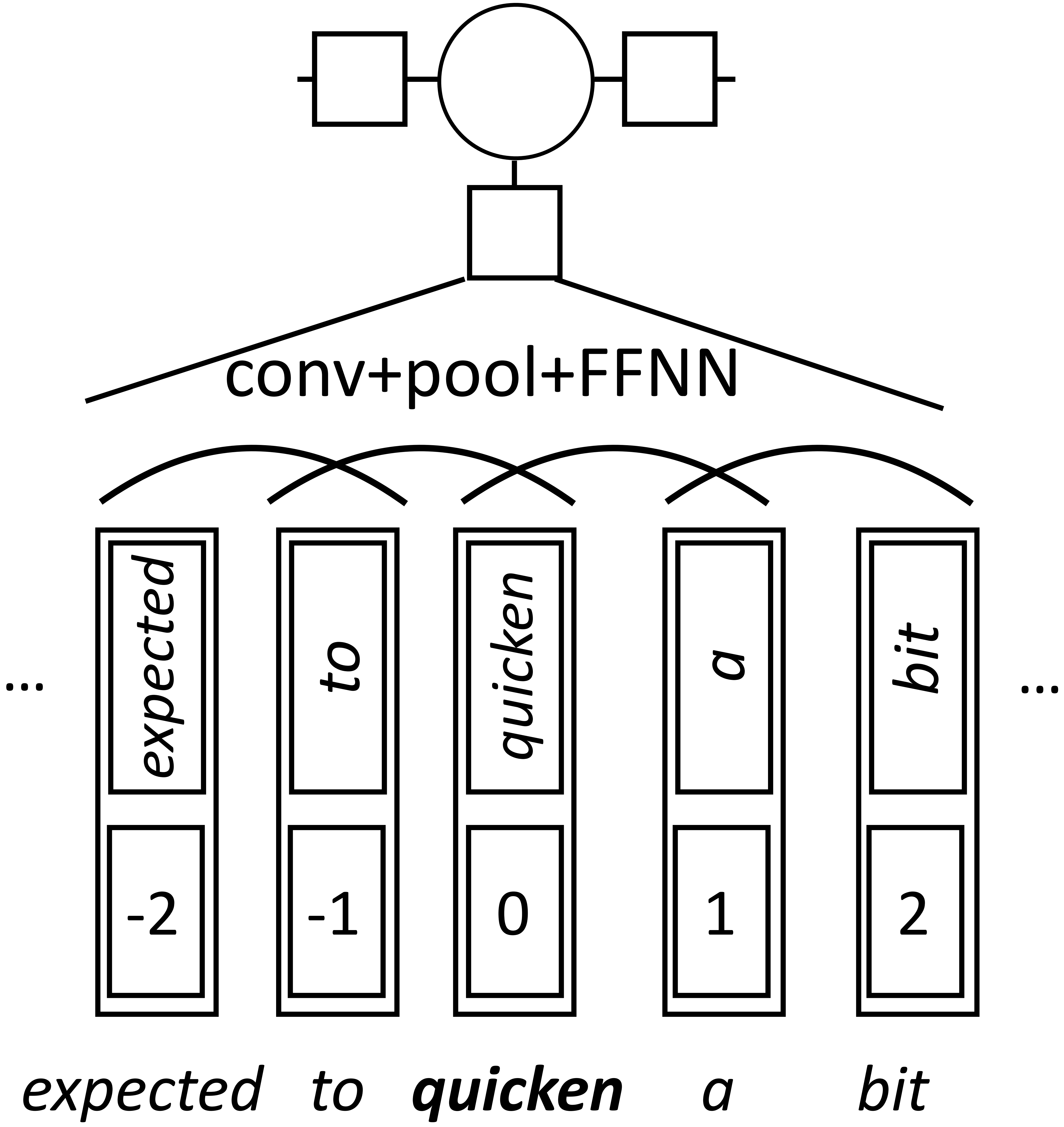


- Tagging problem *with respect to a particular verb*

- Can't do this with feedforward networks efficiently, arguments are too far from the verb to use fixed context window sizes

Figure from He et al. (2017)

CNN Neural CRFs



- Append to each word vector an *embedding of the relative position* of that word
- Convolution over the sentence produces a position-dependent representation
- Use this for SRL: the verb (predicate) is at position 0, CNN looks at the whole sentence “relative” to the verb

CNN NCRFs vs. FFNN NCRFs

Approach	**POS** (PWA)	**CHUNK** (F1)	**NER** (F1)	**SRL** (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
	Window Approach			
NN+SLL+LM2	97.20	93.63	88.67	–
	Sentence Approach			
NN+SLL+LM2	97.12	93.37	88.78	74.15

- Sentence approach (CNNs) is comparable to window approach (FFNNs) except for SRL where they claim it works much better

How "from scratch" was this?

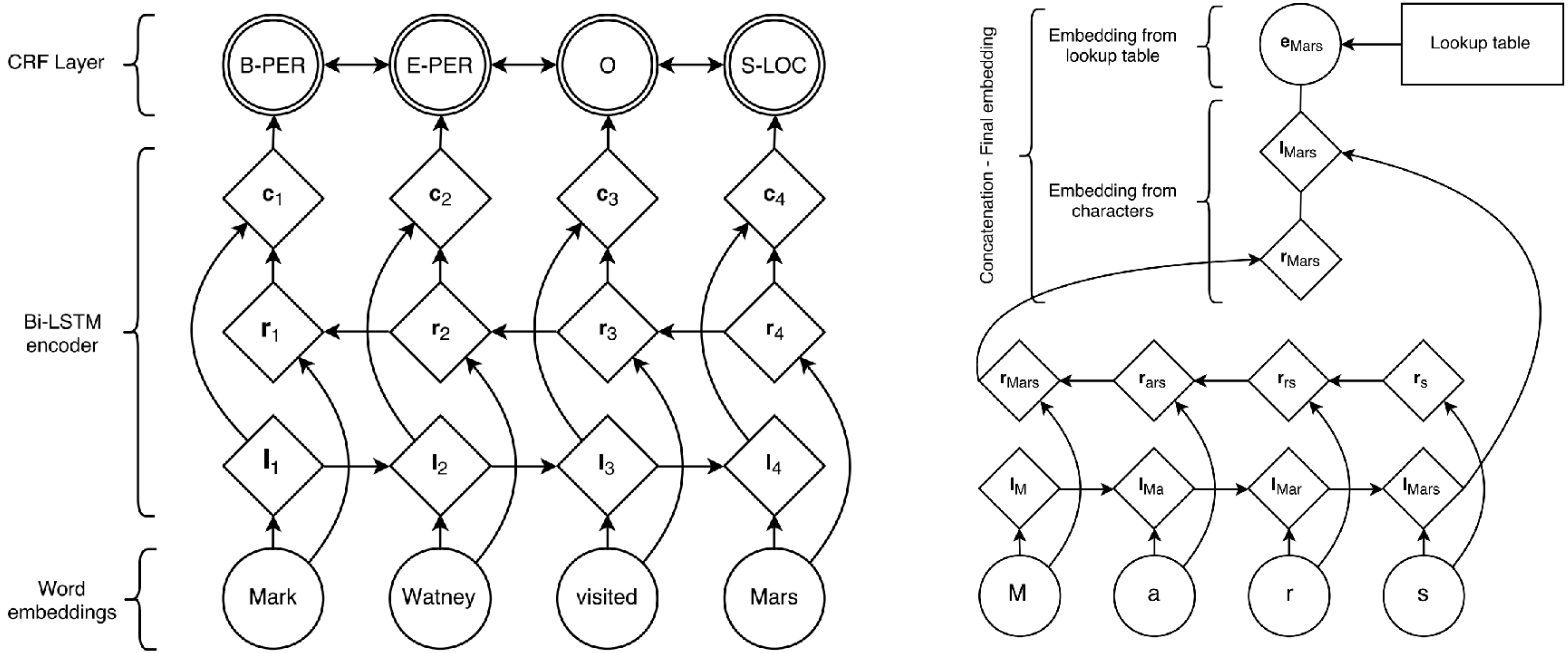
Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
NN+WLL	96.31	89.13	79.53	55.40
NN+SLL	96.37	90.33	81.47	70.99
NN+WLL+LM1	97.05	91.91	85.68	58.18
NN+SLL+LM1	97.10	93.65	87.58	73.84
NN+WLL+LM2	97.14	92.04	86.96	58.34
NN+SLL+LM2	97.20	93.63	88.67	74.15
NN+SLL+LM2+Suffix2	97.29	–	–	–
NN+SLL+LM2+Gazetteer	–	–	89.59	–
NN+SLL+LM2+POS	–	94.32	88.67	–
NN+SLL+LM2+CHUNK	–	–	–	74.72

- NN+SLL isn't great
- LM2: trained for 7 weeks on Wikipedia+Reuters — very expensive!
- Sparse features needed to get best performance on NER+SRL anyway
- No use of sub-word features…

Collobert and Weston 2008, 2011

Neural CRFs with LSTMs

- Neural CRF using character LSTMs to compute word representations



Chiu and Nichols (2015), Lample et al. (2016)

Neural CRFs with LSTMs

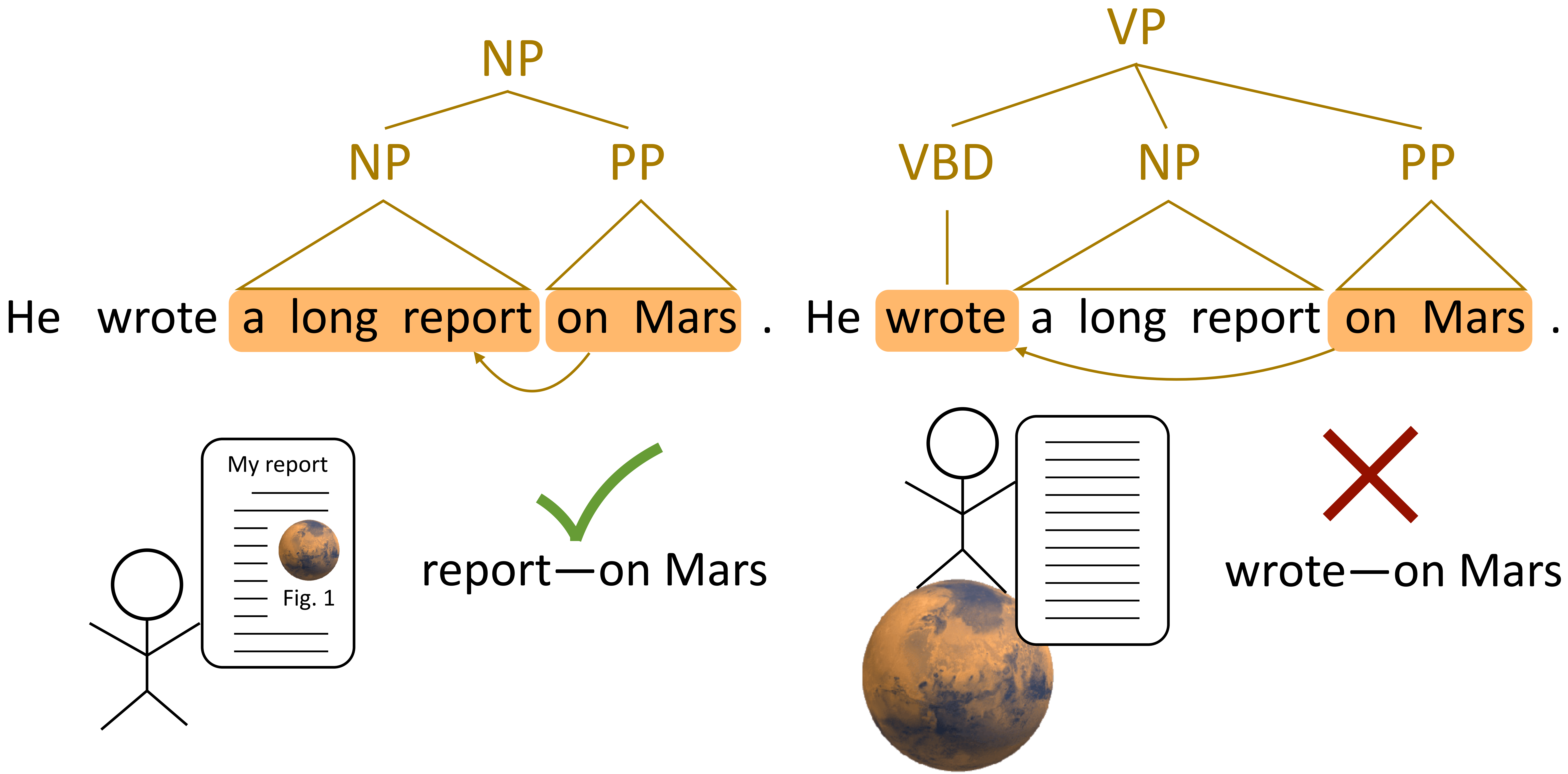
- Chiu+Nichols: character CNNs instead of LSTMs
- Lin/Passos/Luo: use external resources like Wikipedia
- LSTM-CRF captures the important aspects of NER: word context (LSTM), sub-word features (character LSTMs), outside knowledge (word embeddings)

Model	$\mathbf{F_1}$
Collobert et al. (2011)*	89.59
Lin and Wu (2009)	83.78
Lin and Wu (2009)*	90.90
Huang et al. (2015)*	90.10
Passos et al. (2014)	90.05
Passos et al. (2014)*	90.90
Luo et al. (2015)* + gaz	89.9
Luo et al. (2015)* + gaz + linking	**91.2**
Chiu and Nichols (2015)	90.69
Chiu and Nichols (2015)*	90.77
LSTM-CRF (no char)	90.20
LSTM-CRF	**90.94**

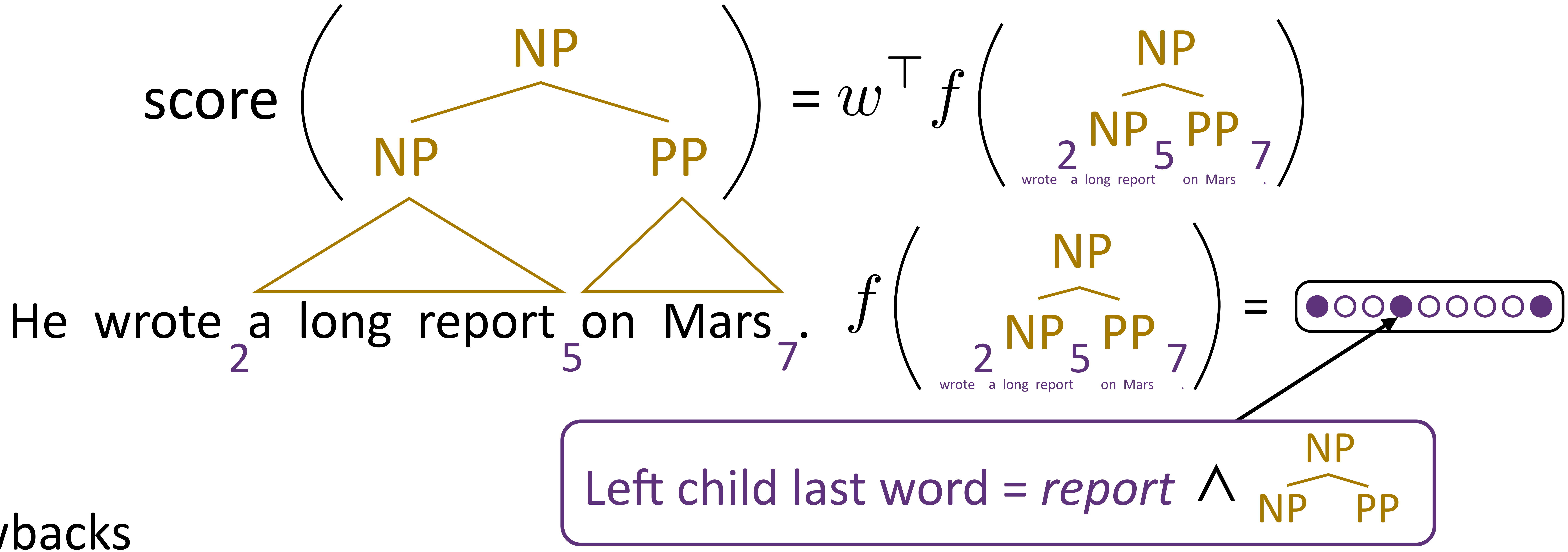
Chiu and Nichols (2015), Lample et al. (2016)

Neural CRFs for Parsing

Constituency Parsing



Discrete Parsing

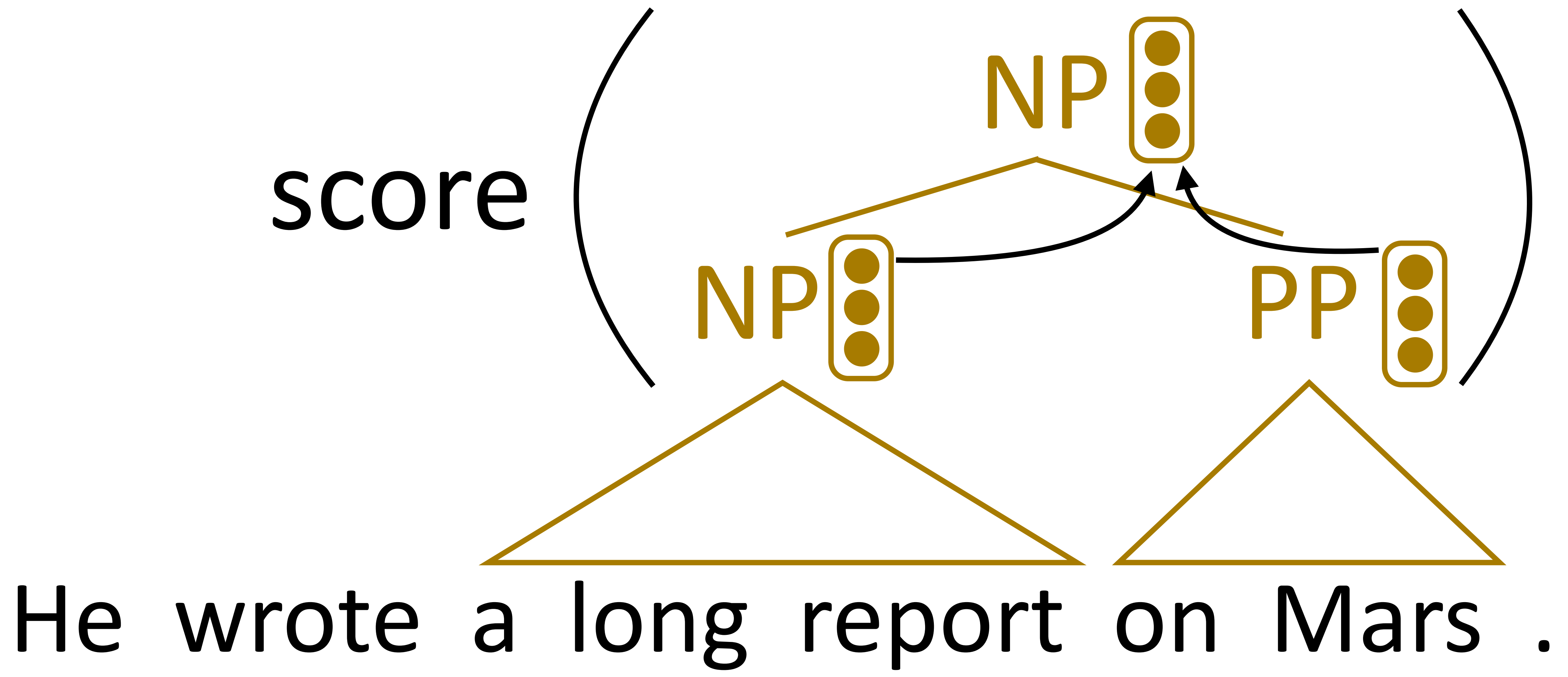


Drawbacks

- Need to learn each word's properties individually
- Hard to learn feature conjunctions (*report on* X)

Taskar et al. (2004)
Hall, Durrett, and Klein (ACL 2014)

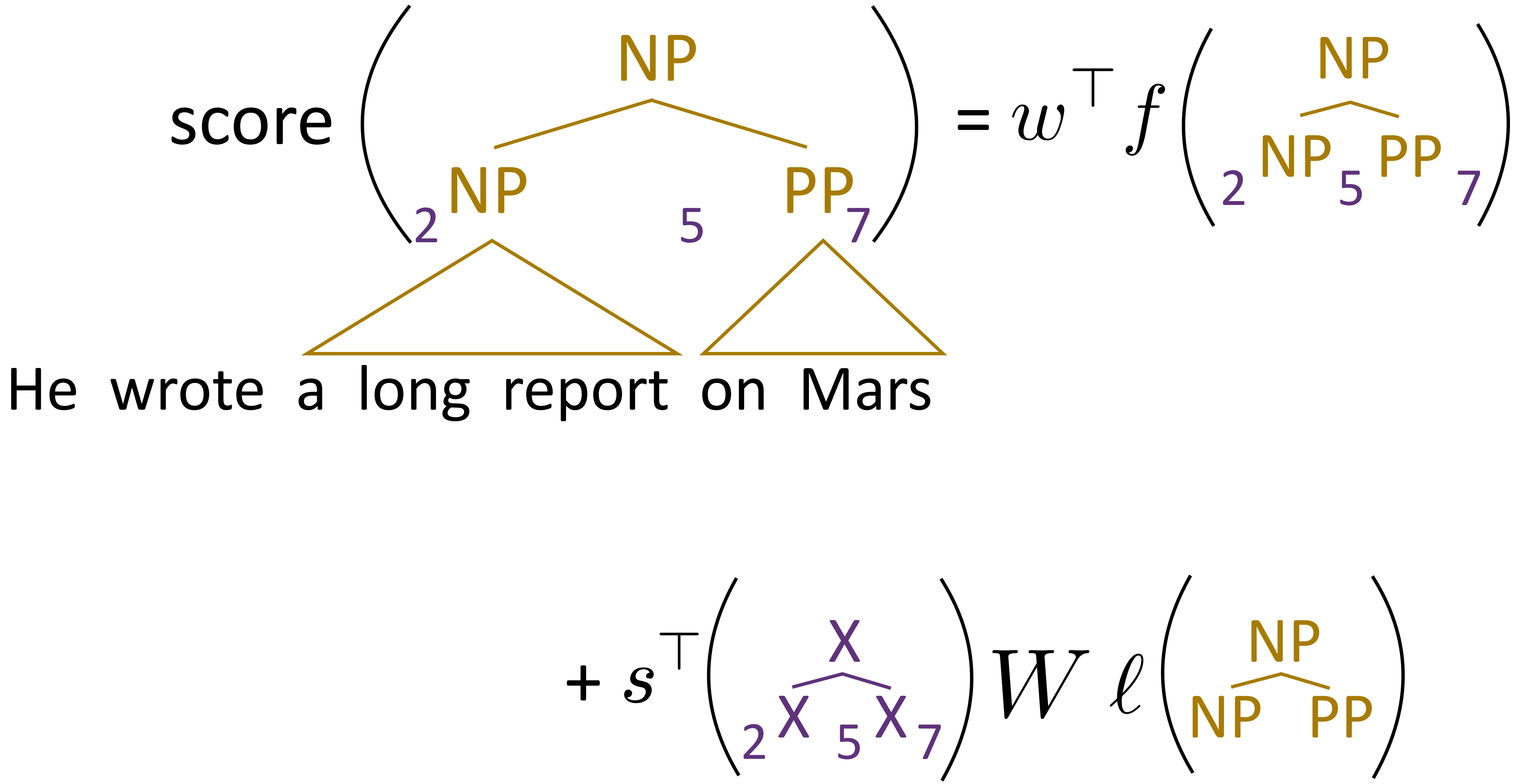
Continuous-State Grammars



Powerful nonlinear featurization, but inference is intractable

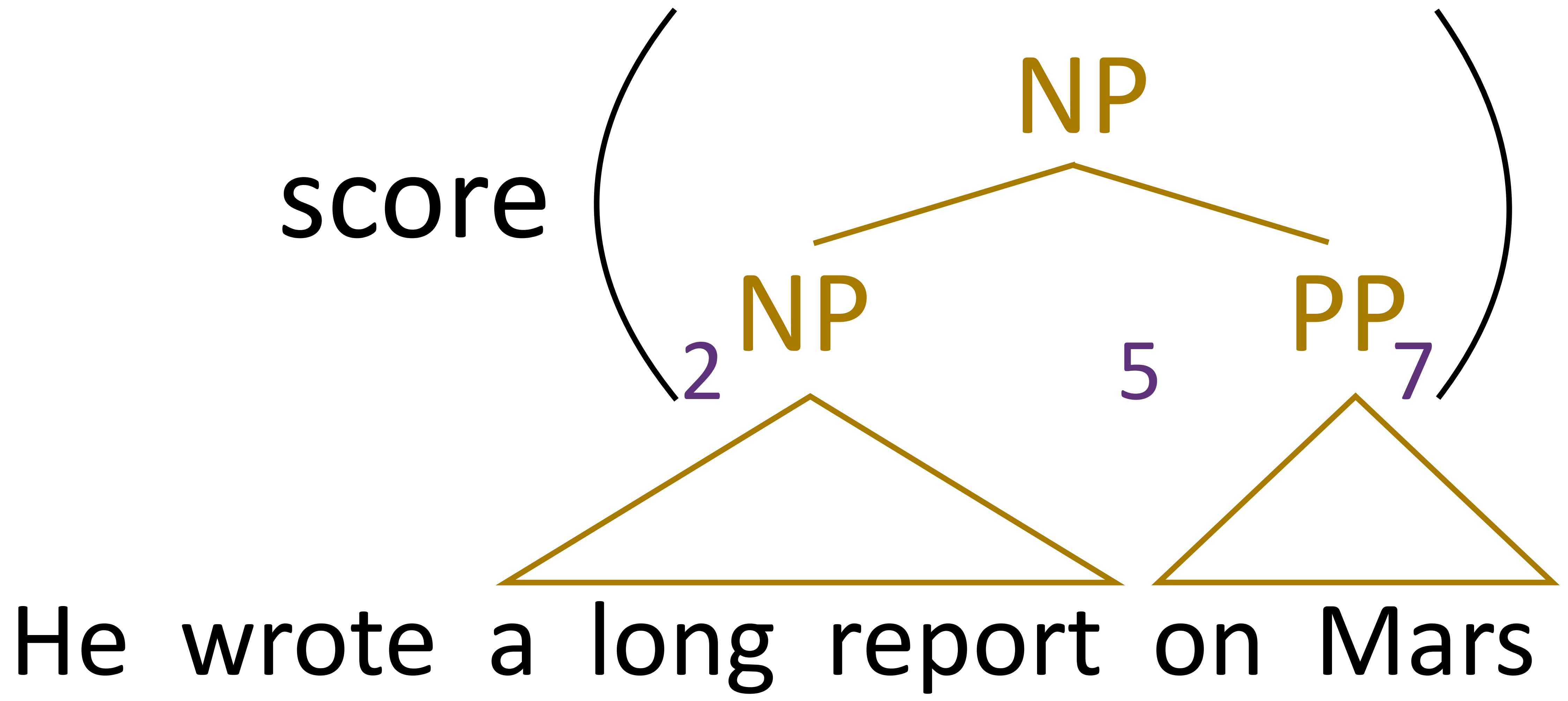
Socher et al. (2013)

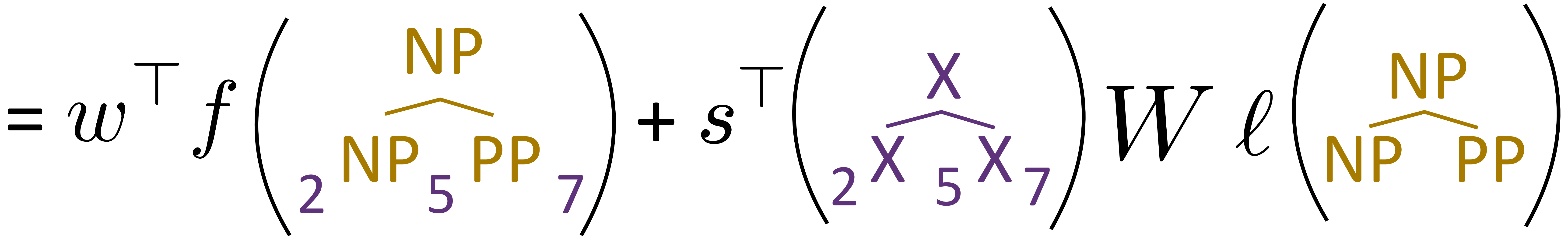
Joint Discrete and Continuous Parsing



Durrett and Klein (ACL 2015)

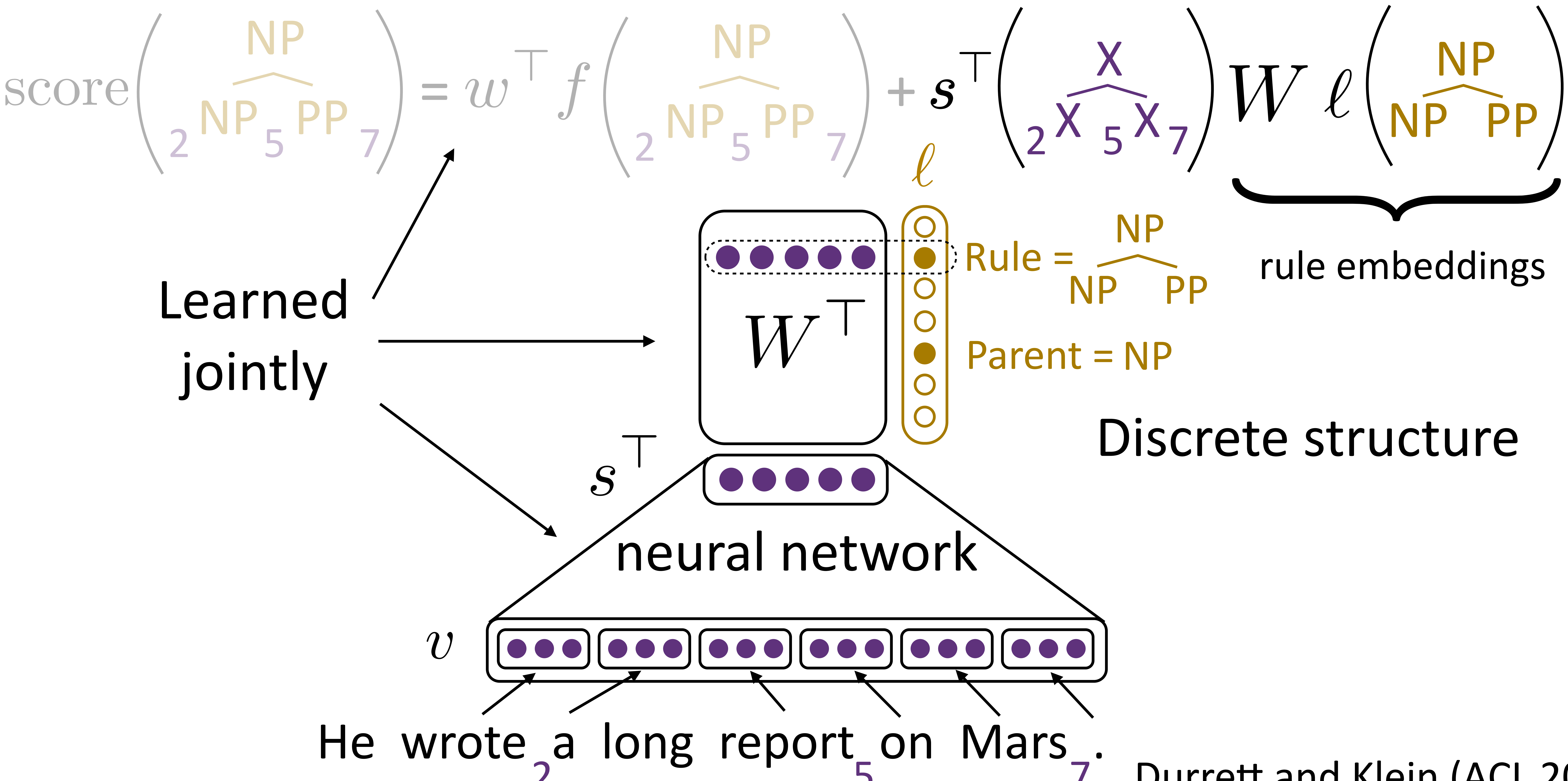
Joint Discrete and Continuous Parsing





Durrett and Klein (ACL 2015)

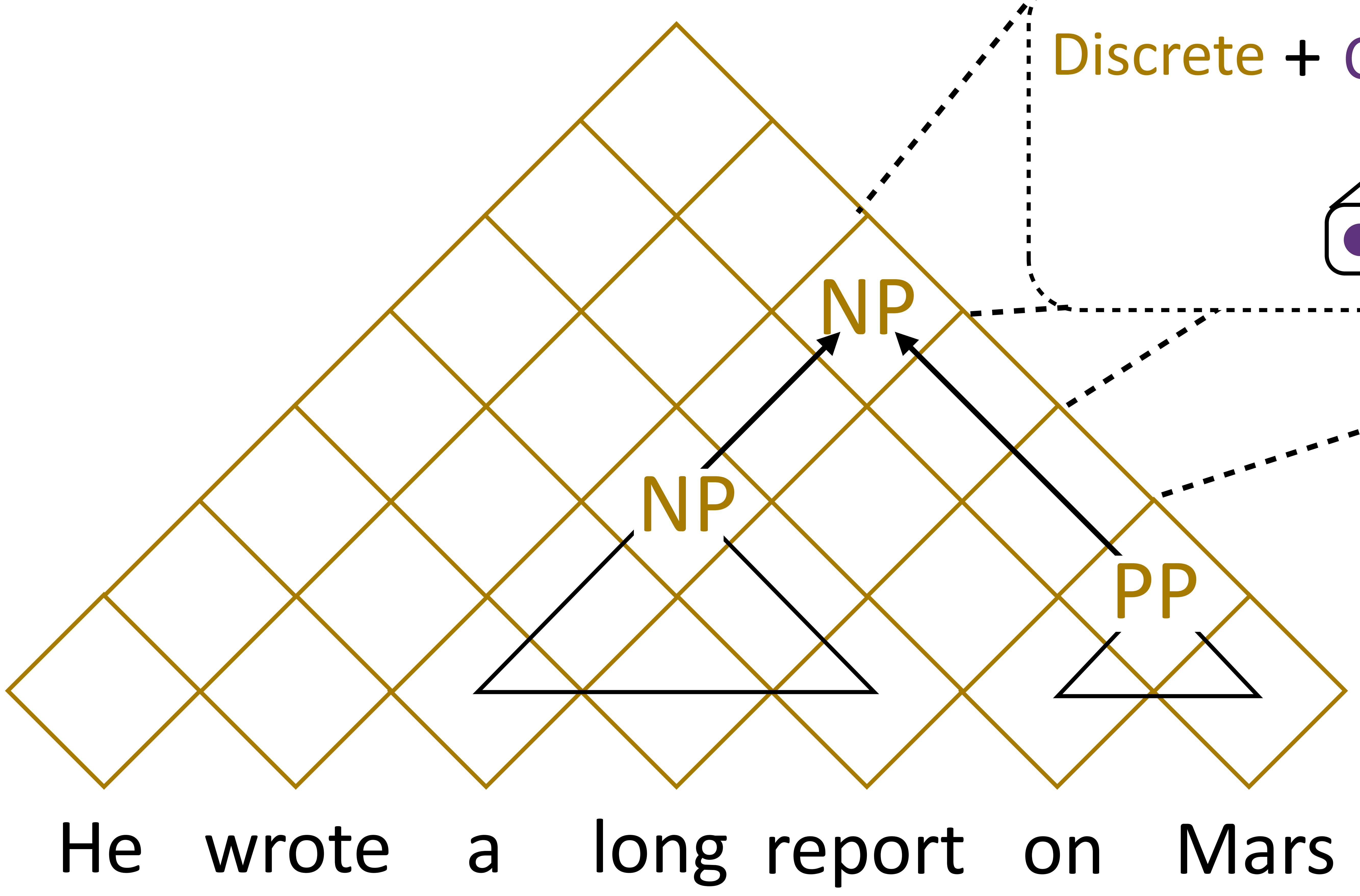
Joint Discrete and Continuous Parsing

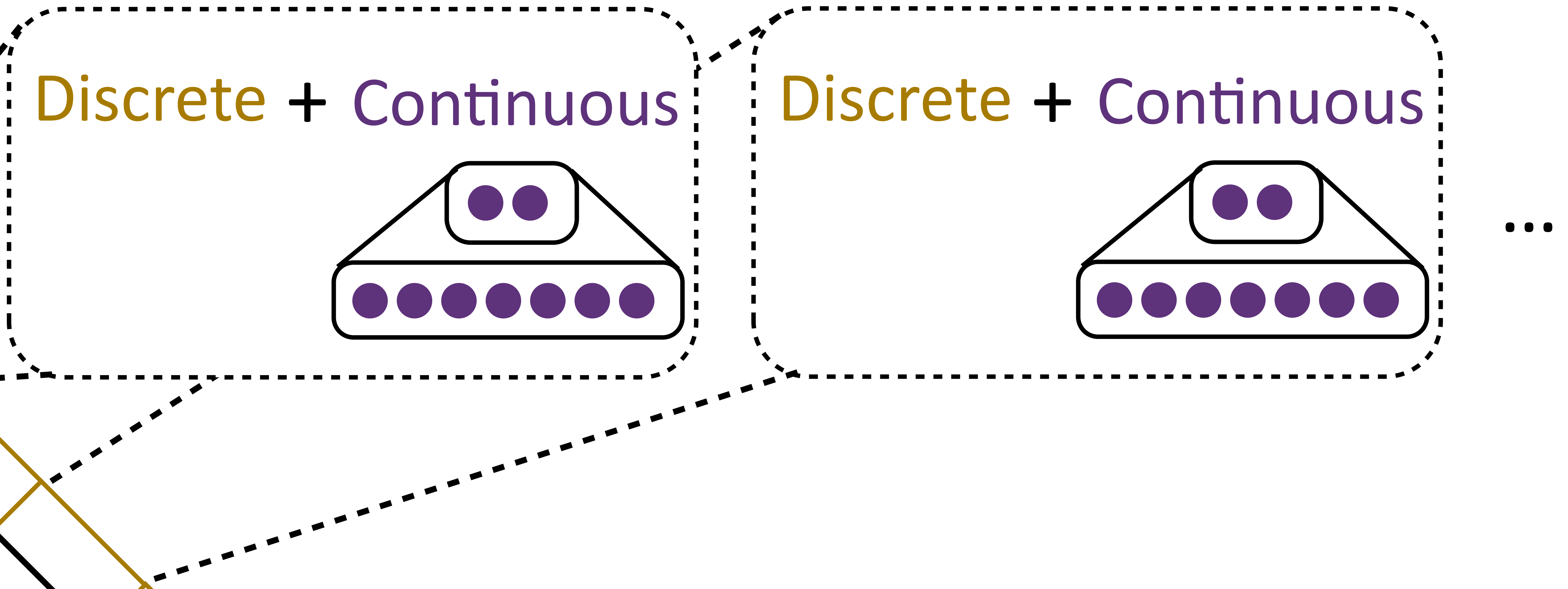


Durrett and Klein (ACL 2015)

Joint Discrete and Continuous Parsing

- Chart remains discrete!



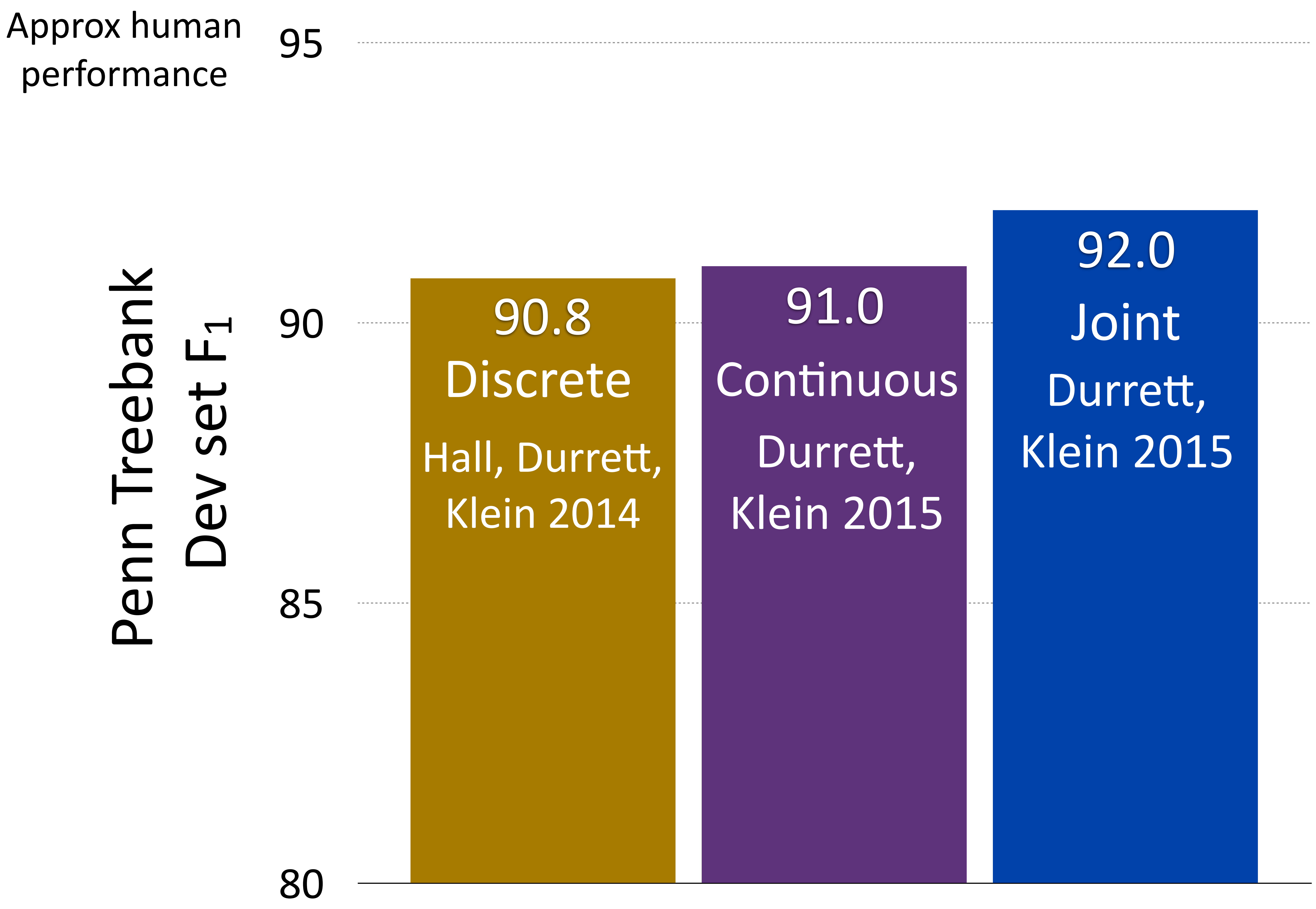


Parsing a sentence:

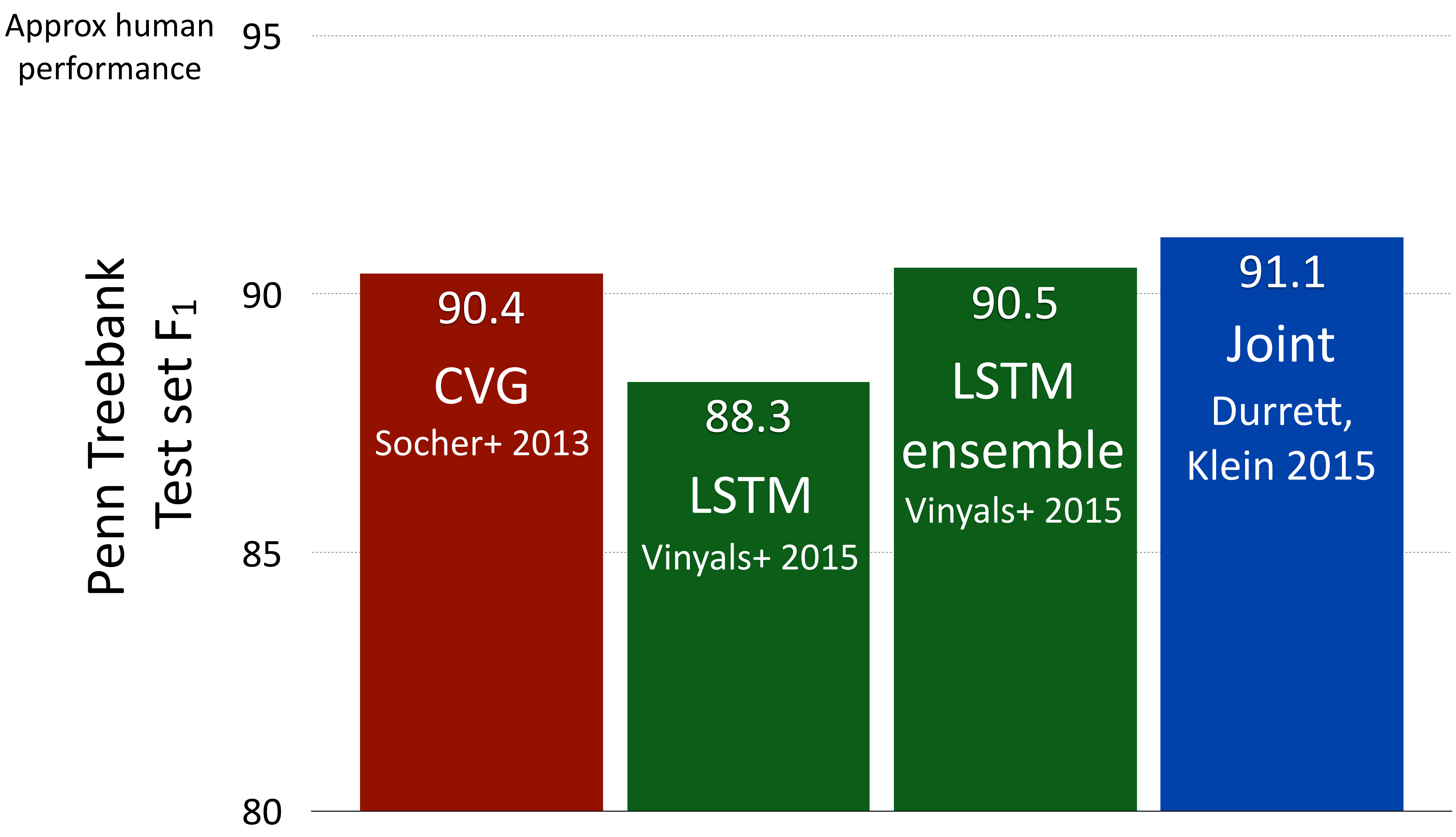
- Feedforward pass on nets
- Discrete feature computation
- Run CKY dynamic program

Durrett and Klein (ACL 2015)

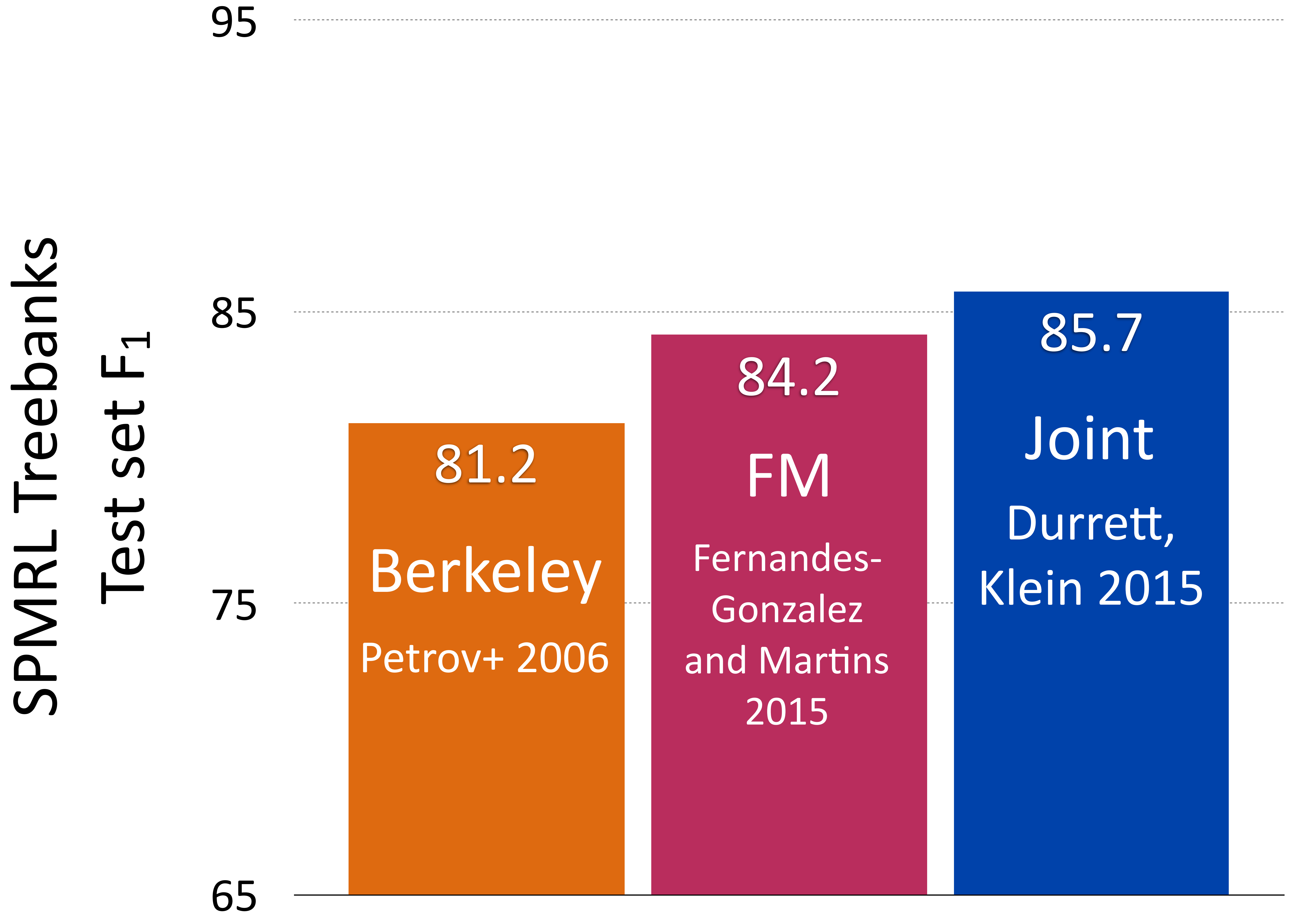
Joint Modeling Helps!



Comparison to Neural Parsers

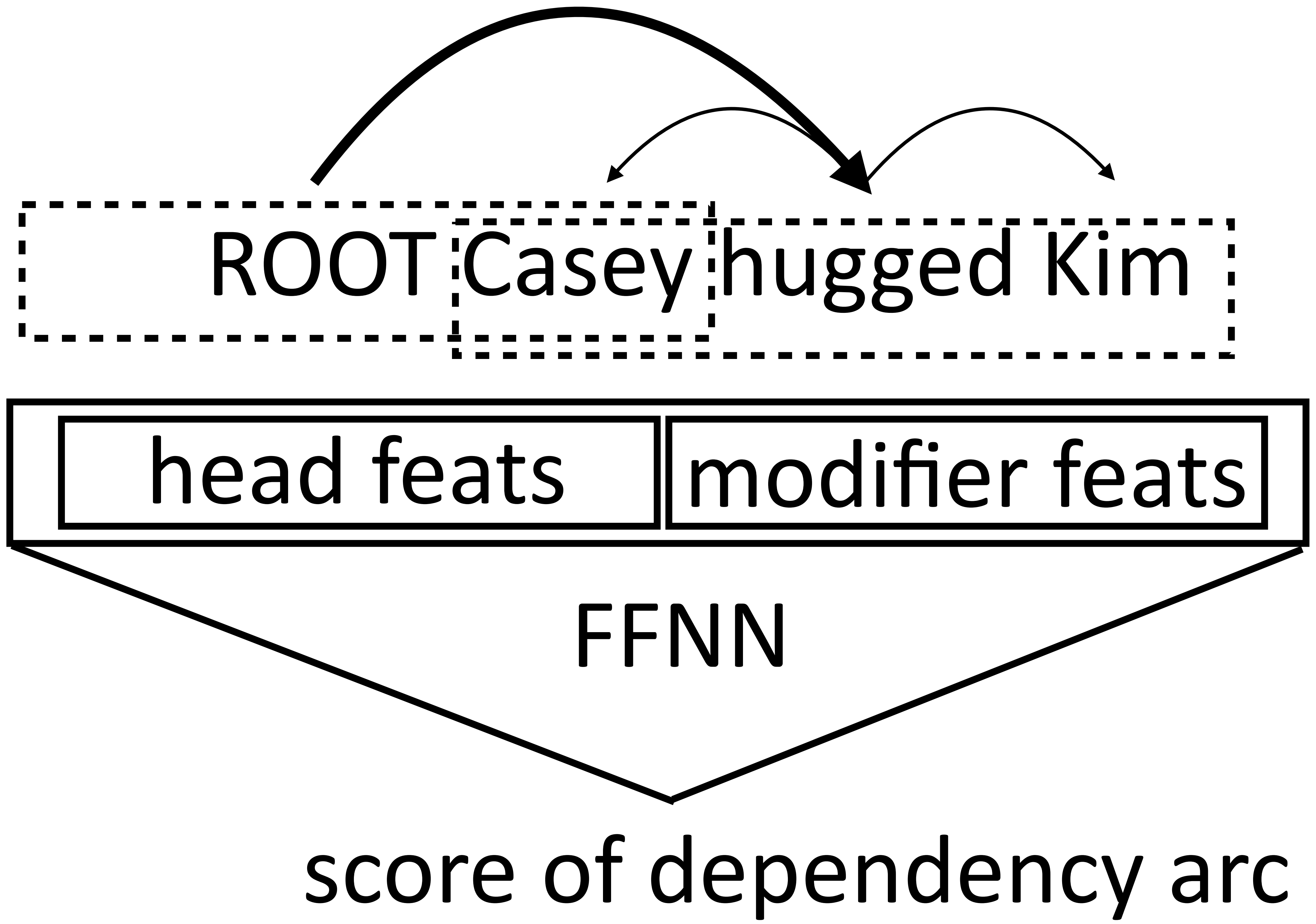


Results: 8 languages



Dependency Parsing

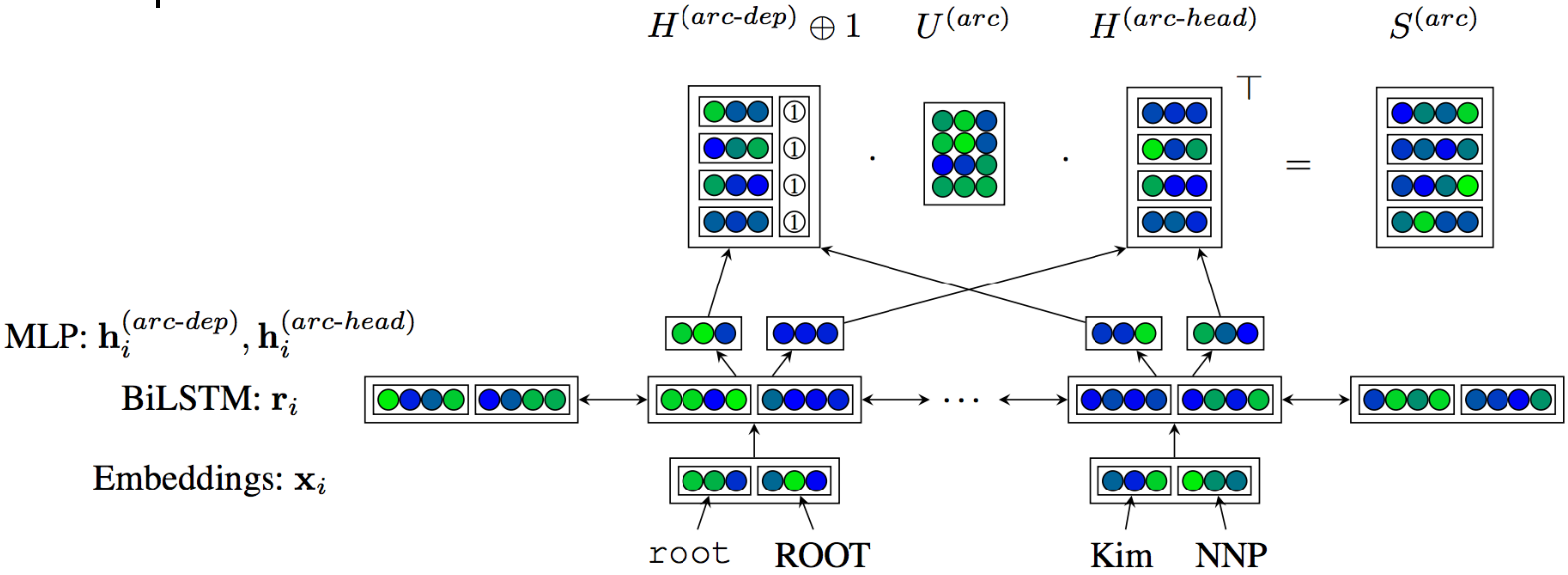
- Score each head-child pair in a dependency parse, use Eisner's algorithm or MST to assemble a parse
- Feedforward neural network approach: use features on head and modifier



Pei et al. (2015), Kiperwasser and Goldberg (2016), Dozat and Manning (2017)

Dependency Parsing

- Biaffine approach: condense each head and modifier separately, compute score $h^\top Um$



Dozat and Manning (2017)

Results

Type	Model	English PTB-SD 3.3.0 UAS	English PTB-SD 3.3.0 LAS	Chinese PTB 5.1 UAS	Chinese PTB 5.1 LAS
Transition	Ballesteros et al. (2016)	93.56	91.42	87.65	86.21
	Andor et al. (2016)	94.61	92.79	–	–
	Kuncoro et al. (2016)	**95.8**	**94.6**	–	–
Graph	Kiperwasser & Goldberg (2016)	93.9	91.9	87.6	86.1
	Cheng et al. (2016)	94.10	91.49	88.1	85.7
	Hashimoto et al. (2016)	94.67	92.90	–	–
	Deep Biaffine	95.74	94.08	**89.30**	**88.23**

- Biaffine approach works well (other neural CRFs are also strong)

Dozat and Manning (2017)

Neural CRFs

- State-of-the-art for:
  - POS
  - NER without extra data (Lample et al.)
  - Dependency parsing (Dozat and Manning)
  - Semantic Role Labeling (He et al.)
- Why do they work so well?
  - Word-level LSTMs compute features based on the word + context
  - Character LSTMs/CNNs extract features per word
  - Pretrained embeddings capture external semantic information
  - CRF handles structural aspects of the problem

Takeaways

- Any structured model / dynamic program + any neural network to compute potentials = neural CRF

- Can incorporate transition potentials or other scores over the structure like grammar rules

- State-of-the-art for many text analysis tasks