

Information Extraction

Wei Xu

(many slides from Greg Durrett, Luheng He, Emma Strubell)

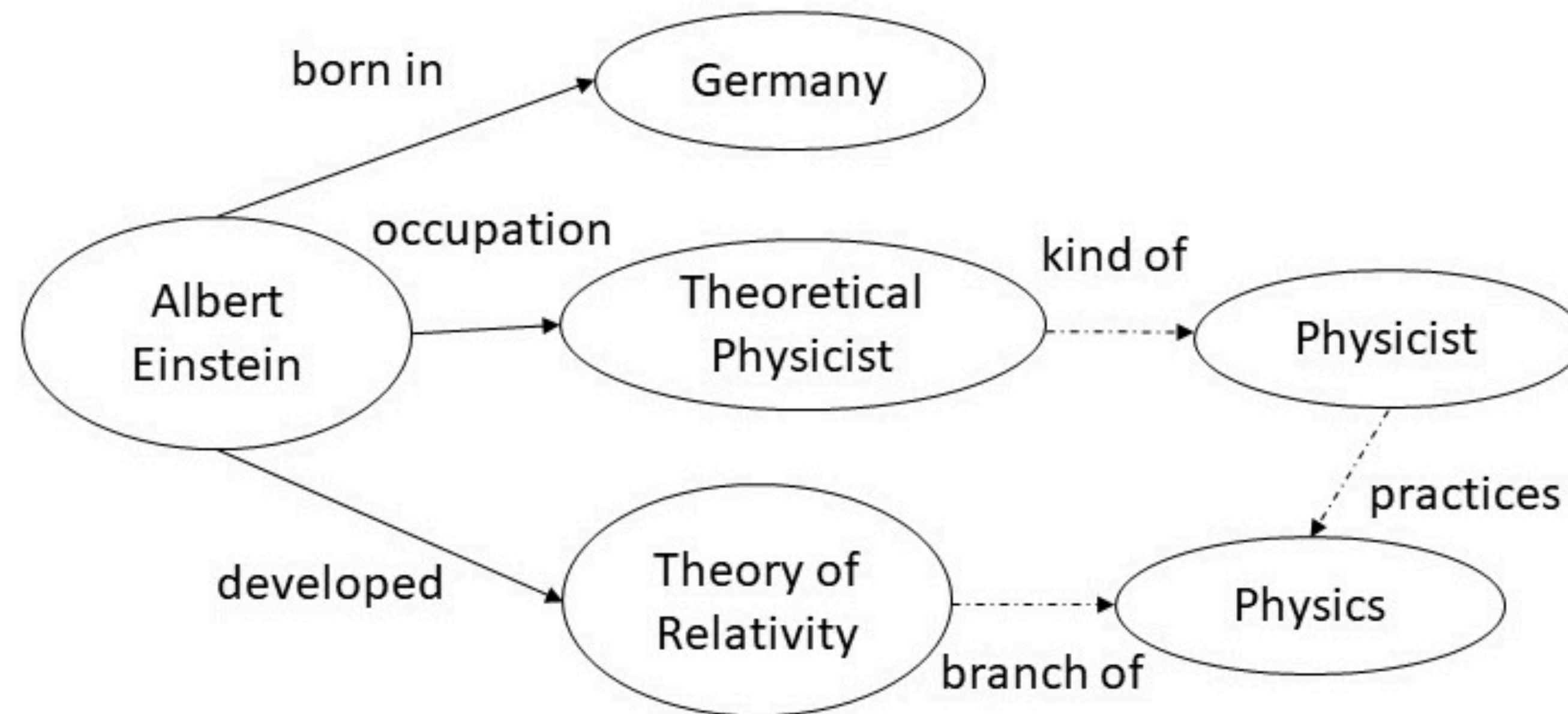
Administrivia

- ▶ Project 3 is released (seq2seq for dialog; can be used for MT)
- ▶ Readings — Eisenstein Chapter 13 & 17
- ▶ Additional readings (if interested) can be found at the right-bottom corner of the slides.

Information Extraction (IE)

- ▶ Extract important, structured information (e.g., a knowledge base) from unstructured texts (often in large quantity)

Albert Einstein was a German-born theoretical physicist who developed the theory of relativity.



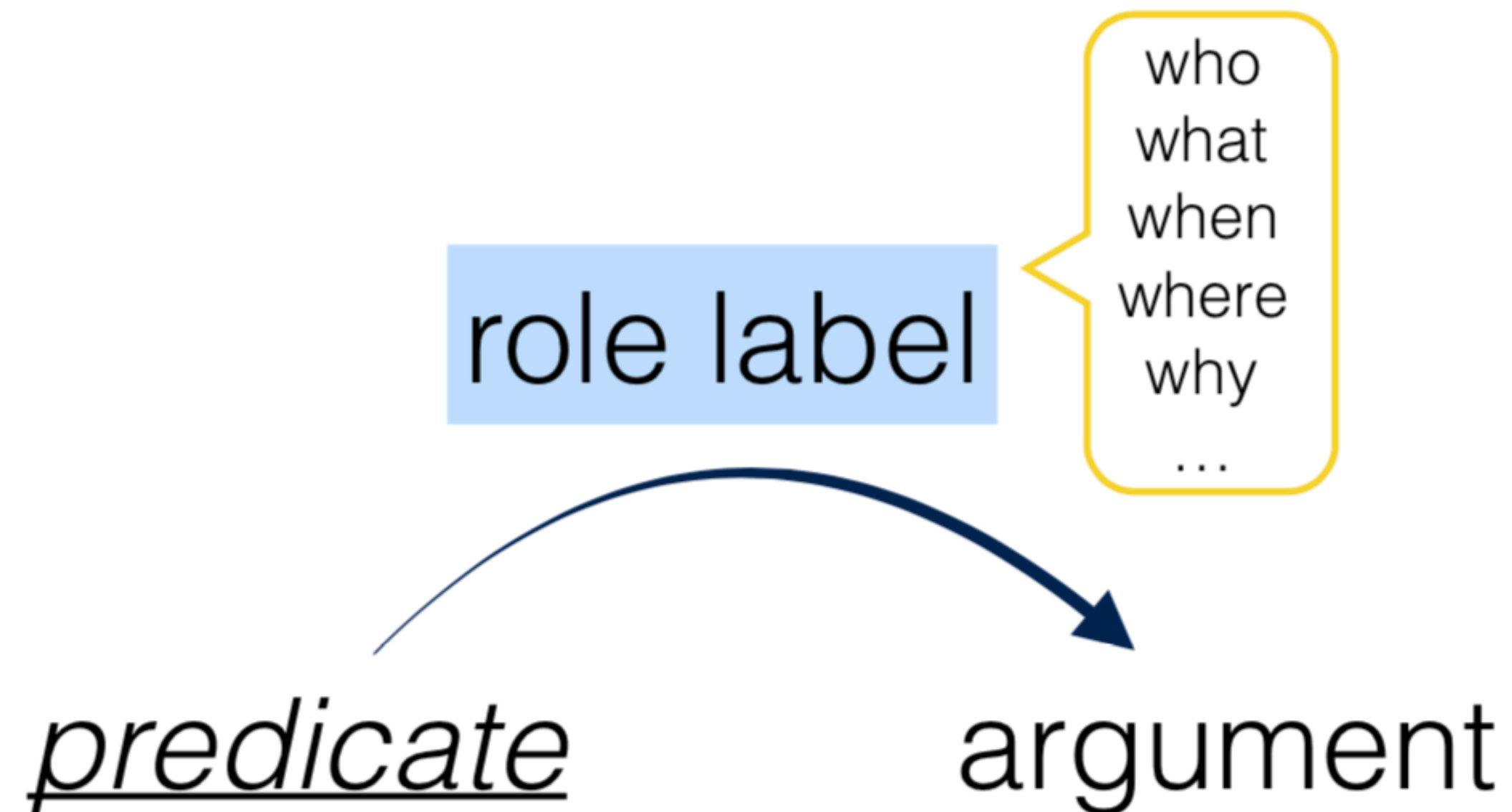
This Lecture

- ▶ How do we represent information for information extraction?
- ▶ Semantic role labeling
- ▶ Relation extraction
- ▶ Open Information Extraction
- ▶ Slot filling

Semantic Role Labeling

Semantic Role Labeling

- ▶ Find out 5W in text — “who did what to whom, when and where”
- ▶ Identify predicate, disambiguate it, identify that predicate’s arguments



Question-Answer Driven SRL

In 1950 Alan M. Turing *published* "Computing machinery and intelligence" in Mind, in which he *proposed* that machines could be *tested* for intelligence *using* questions and answers.

<u>Predicate</u>		<u>Question</u>	<u>Answer</u>
published	1	Who published something?	Alan M. Turing
	2	What was published?	"Computing Machinery and Intelligence"
	3	When was something published?	In 1950
proposed	4	Who proposed something?	Alan M. Turing
	5	What did someone propose?	that machines could be tested for intelligent using questions and answers
	6	When did someone propose something?	In 1950
tested	7	What can be tested?	machines
	8	What can something be tested for?	intelligence
	9	How can something be tested?	using questions and answers
using	10	What was being used?	questions and answers
	11	Why was something being used?	tested for intelligence

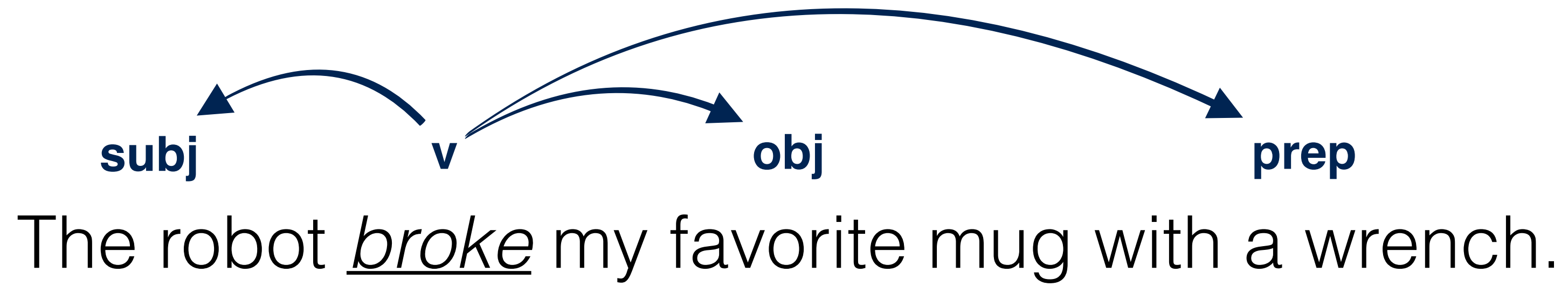
Figure from FitzGerald et al. (2018)

Semantic Role Labeling

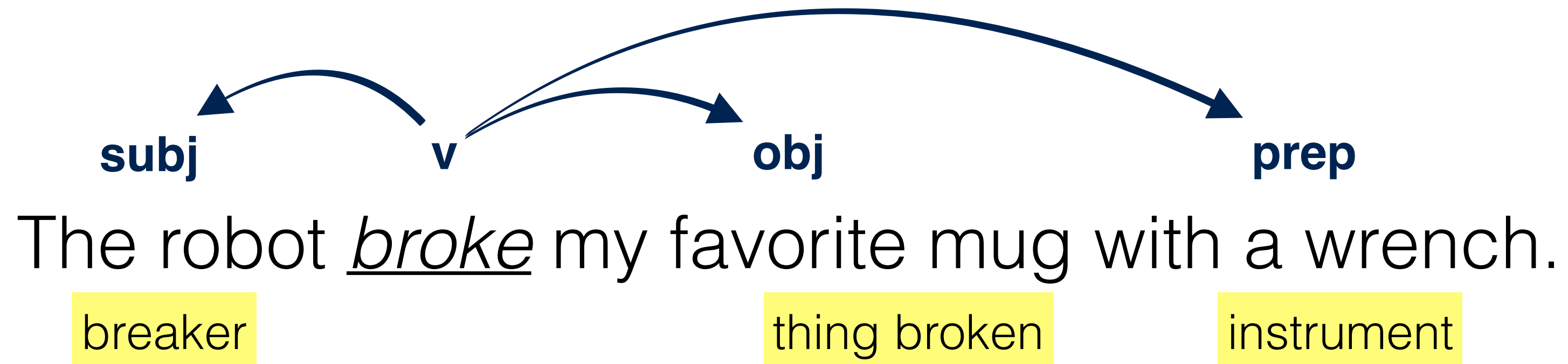
The robot broke my favorite mug with a wrench.

My mug broke into pieces immediately.

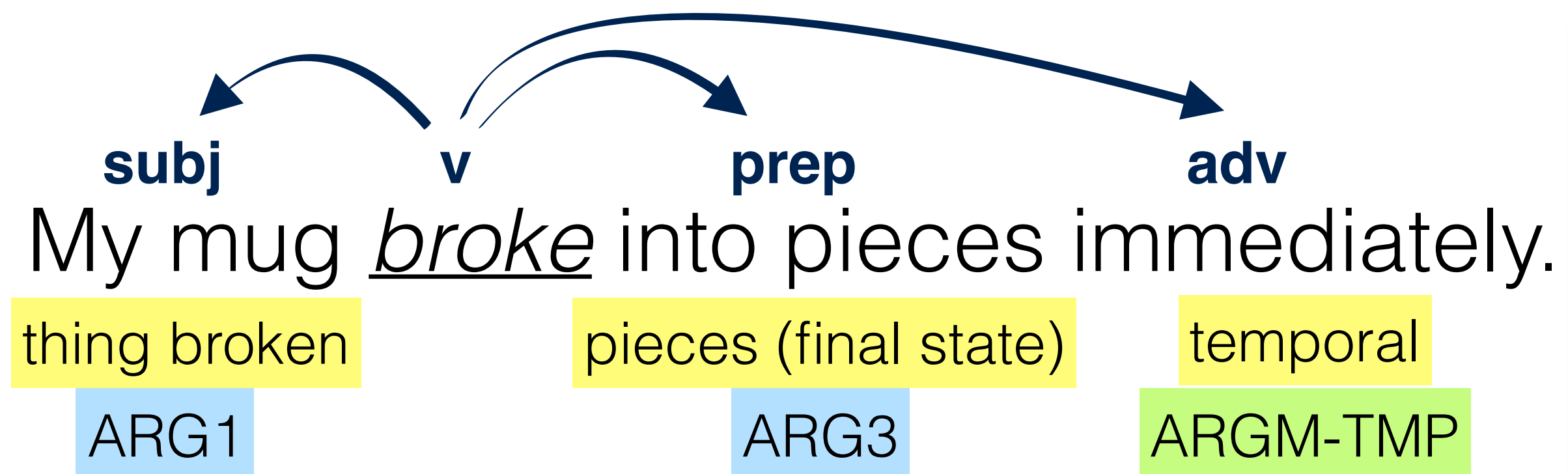
Semantic Role Labeling



Semantic Role Labeling



Semantic Role Labeling



Frame: break.01

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument
ARG3	pieces
ARG4	broken away from what?

The Proposition Bank (PropBank)

Core roles:
Verb-specific roles (ARG0-
ARG5) defined in frame files

Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument

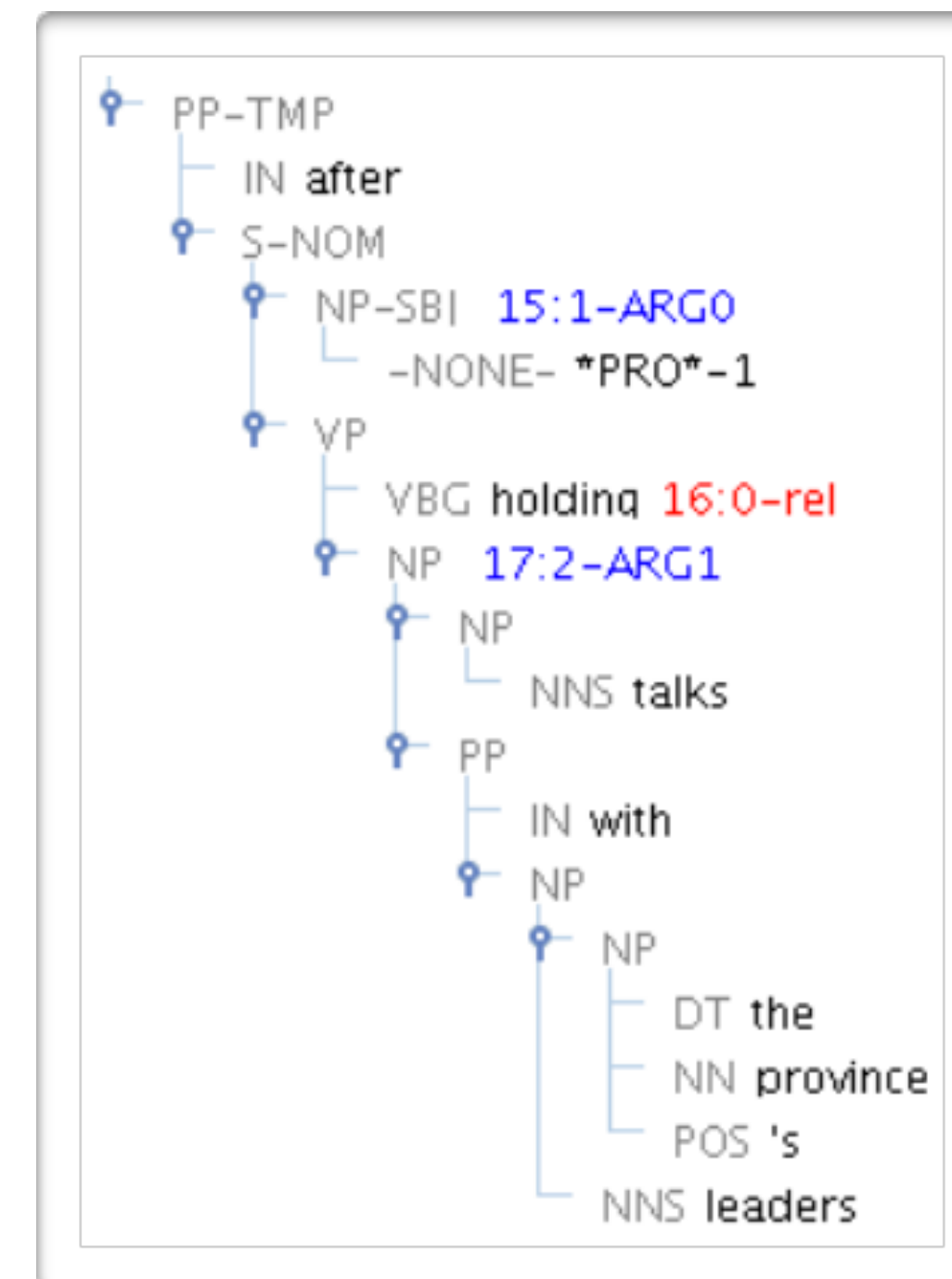
Frame: *buy.01*

role	description
ARG0	buyer
ARG1	thing bough
ARG2	seller
ARG3	price paid
ARG4	benefactive

Adjunct roles:
(ARGM-) shared
across verbs

role	description
TMP	temporal
LOC	location
MNR	manner
DIR	direction
CAU	cause
PRP	purpose
...	

Annotated on top of the
Penn Treebank Syntax



PropBank Annotation Guidelines,
Bonial et al., 2010

Figure from He et al. (2017)

Syntax vs. Semantics

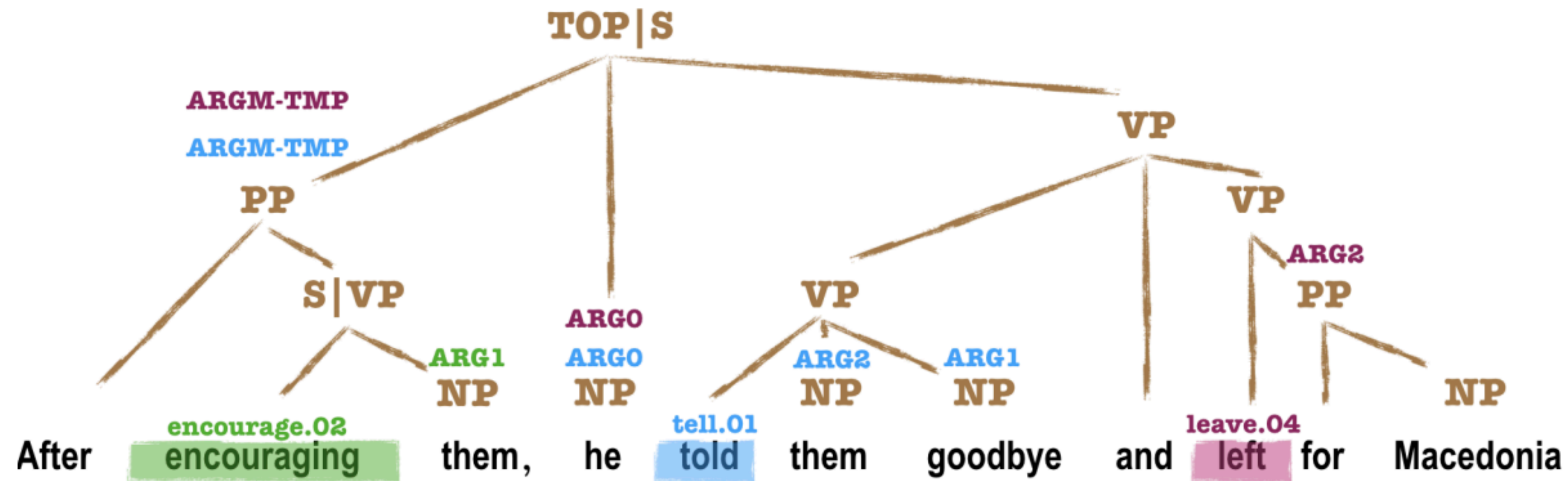
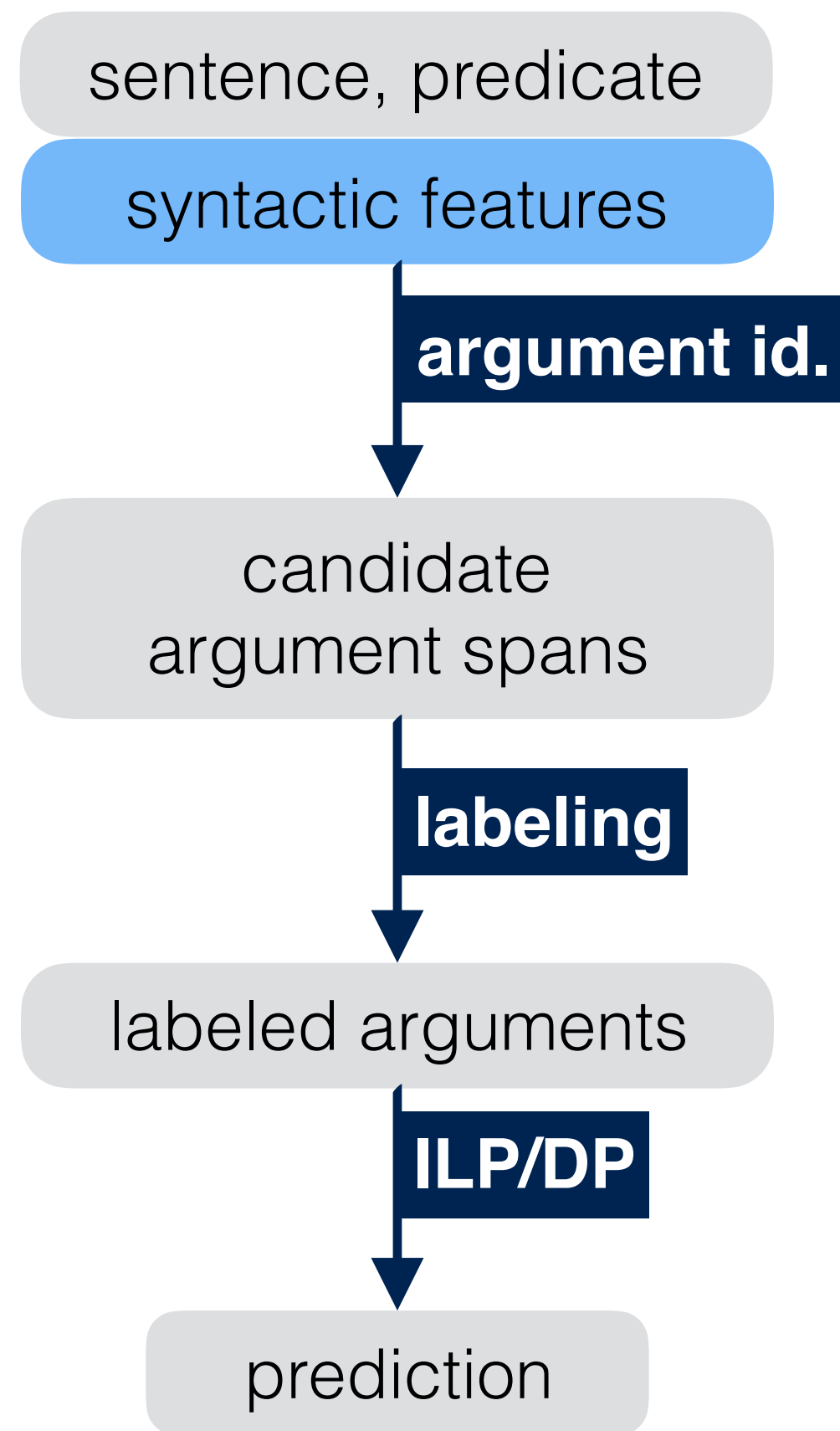


Figure 1.2: Syntax and semantics are closely related. The phrase-syntactic tree is shown in brown above the sentence. Semantic role labeling (SRL) structures from PropBank (Palmer et al., 2005) are shown alongside, in green, blue and magenta. Under SRL, words in the sentence that indicate stand-alone events are selected as predicates. These are shown as highlighted leaf nodes—"encouraging", "told" and "left". Each predicate is disambiguated to its relevant sense shown above it. Arguments to the predicates are annotated on top of syntactic nodes, with the role labels color-coded by the predicate. SRL substructures (predicates, arguments) thus fully overlap with phrase-syntactic nodes.

SRL Systems

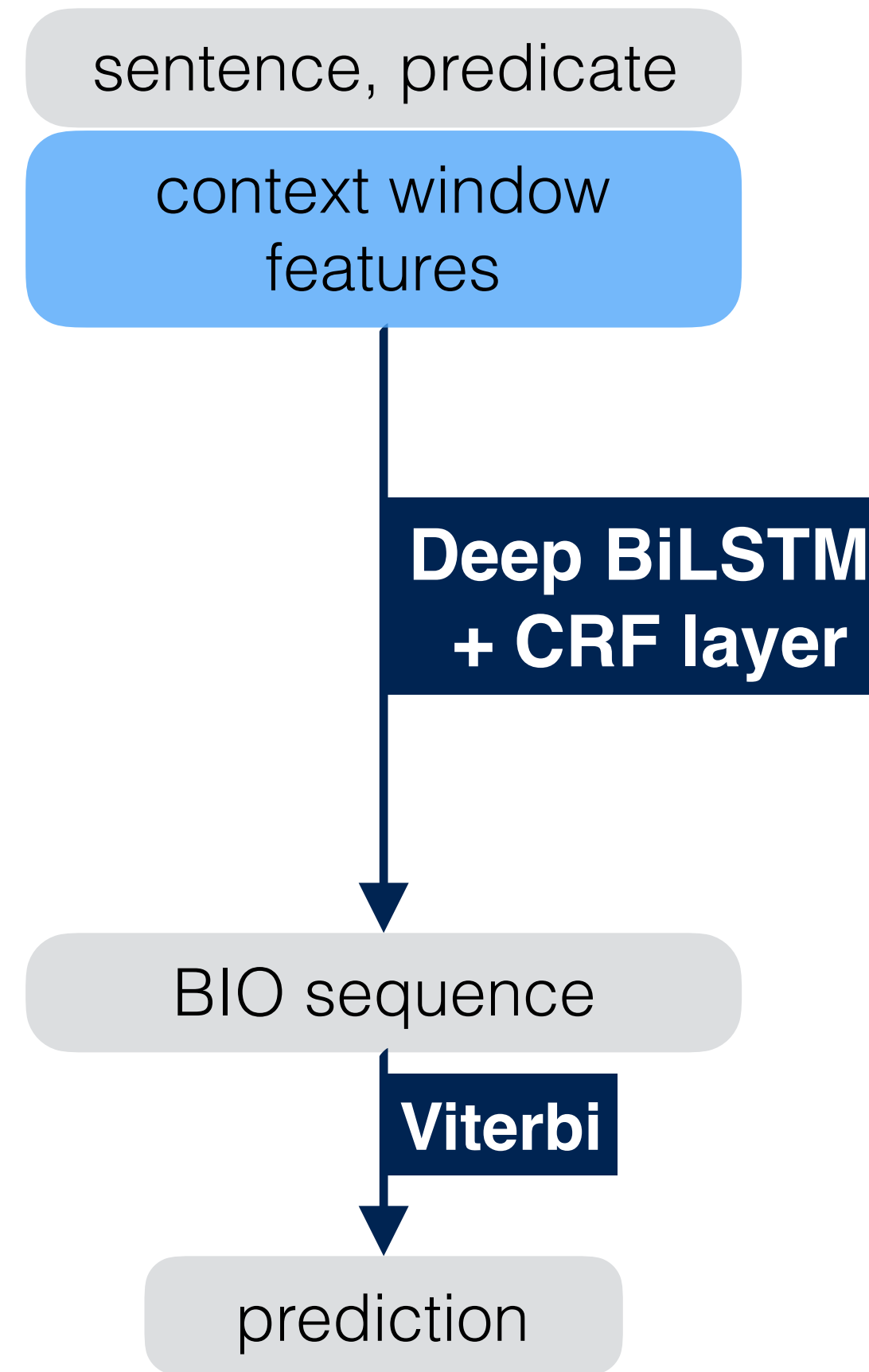
(syntax-based)

Pipeline Systems



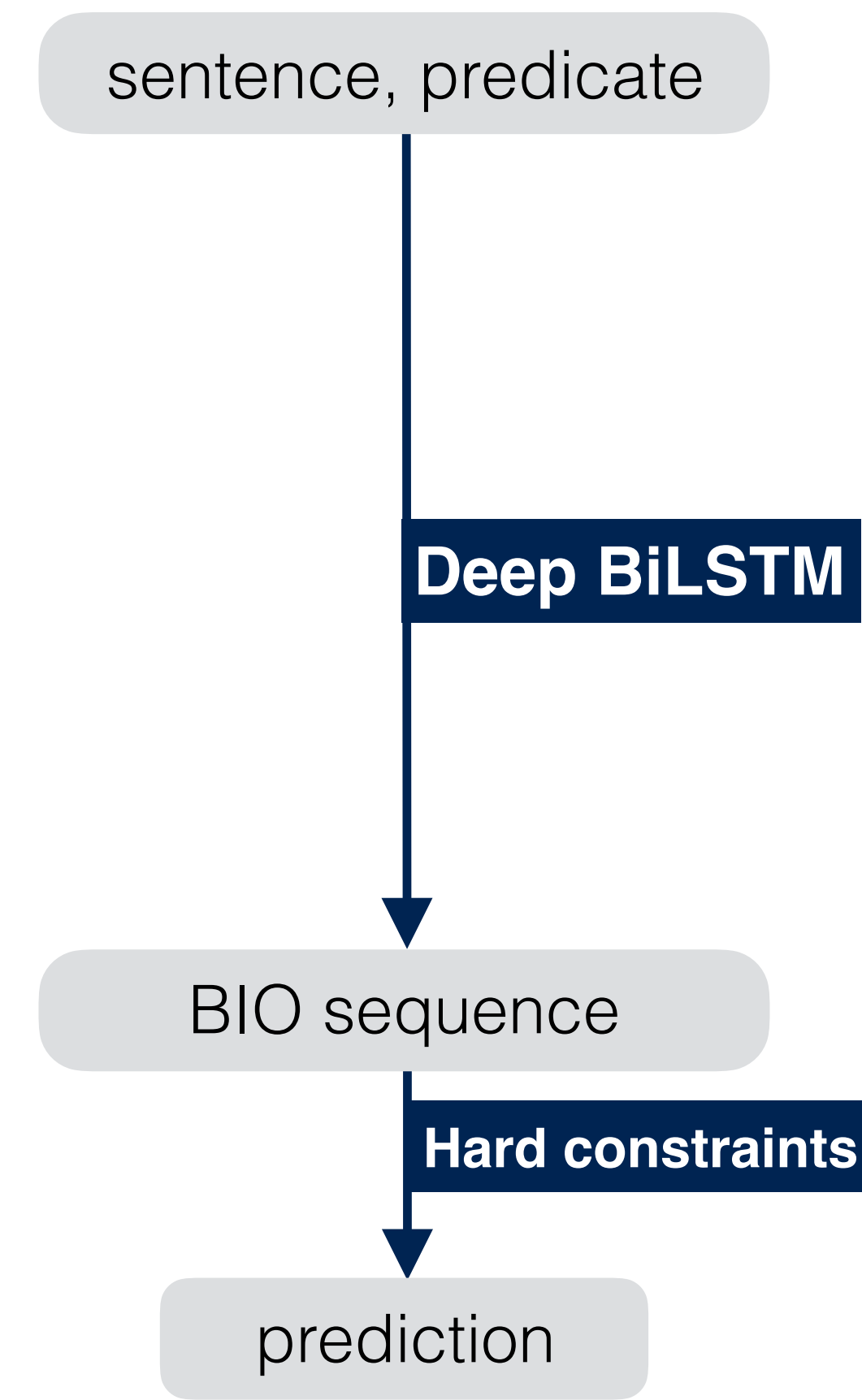
Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015

End-to-end Systems



Collobert et al., 2011
Zhou and Xu, 2015
Wang et. al, 2015

He et al. (2017)



He et al., 2017

Figure from He et al. (2017)

Semantic Role Labeling

- ▶ Identify predicates (*love*) using a classifier (not shown)
- ▶ Identify ARG0, ARG1, etc. as a tagging task with a BiLSTM conditioned on *love*
- ▶ Other systems incorporate syntax, joint predicate-argument finding

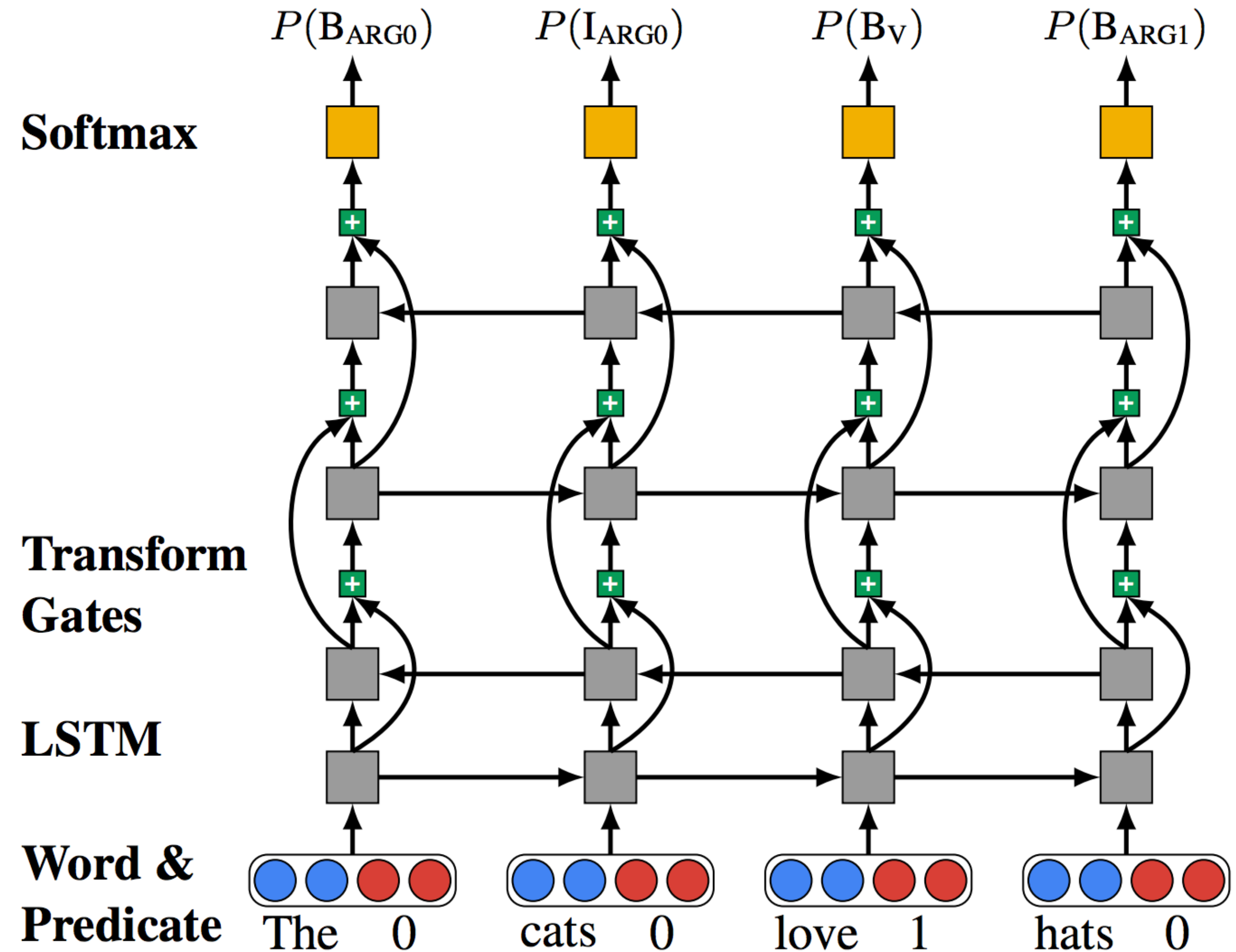


Figure from He et al. (2017)

Semantic Role Labeling

- ▶ Used highway connections, variational dropout, Viterbi decoding with hard constraints.

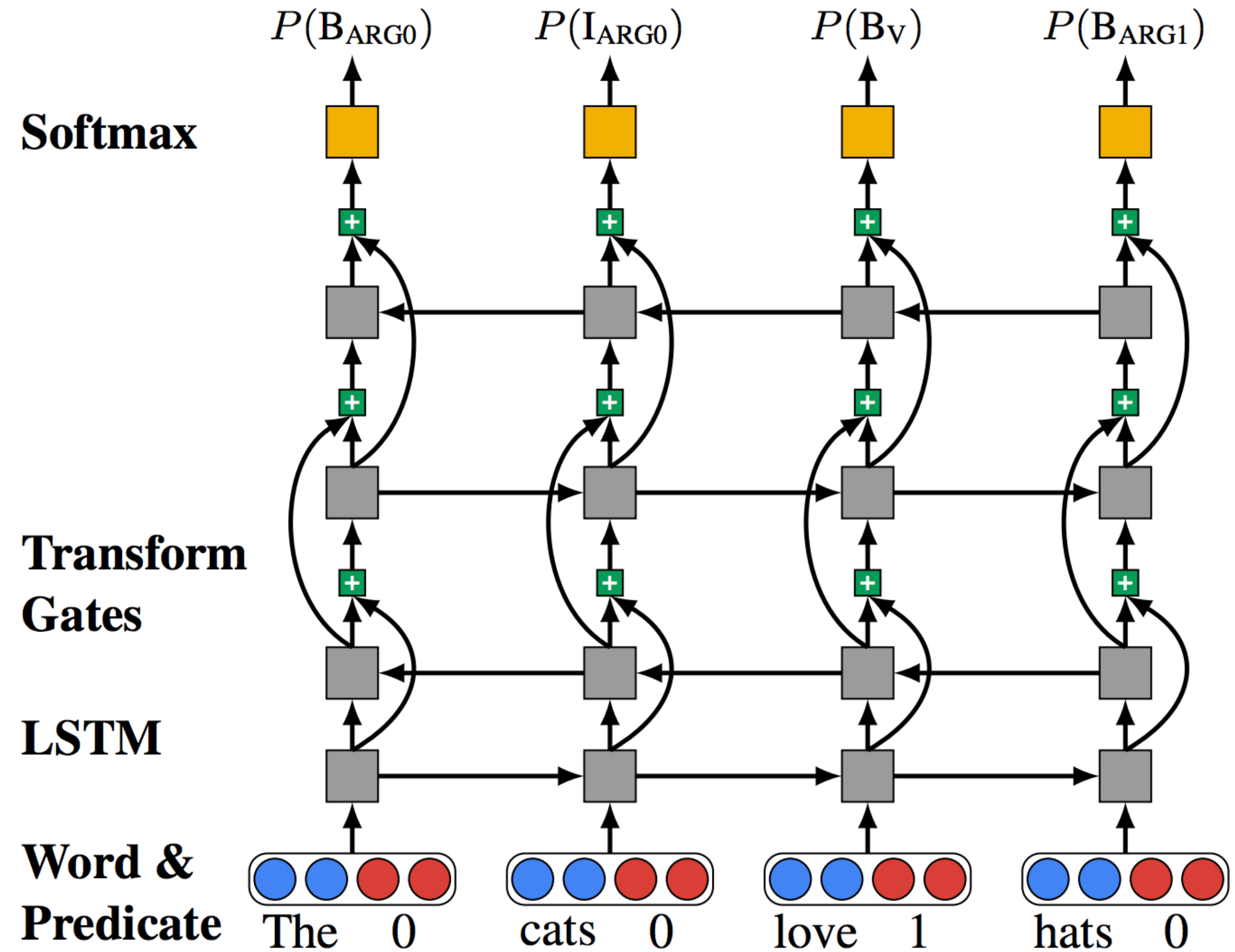
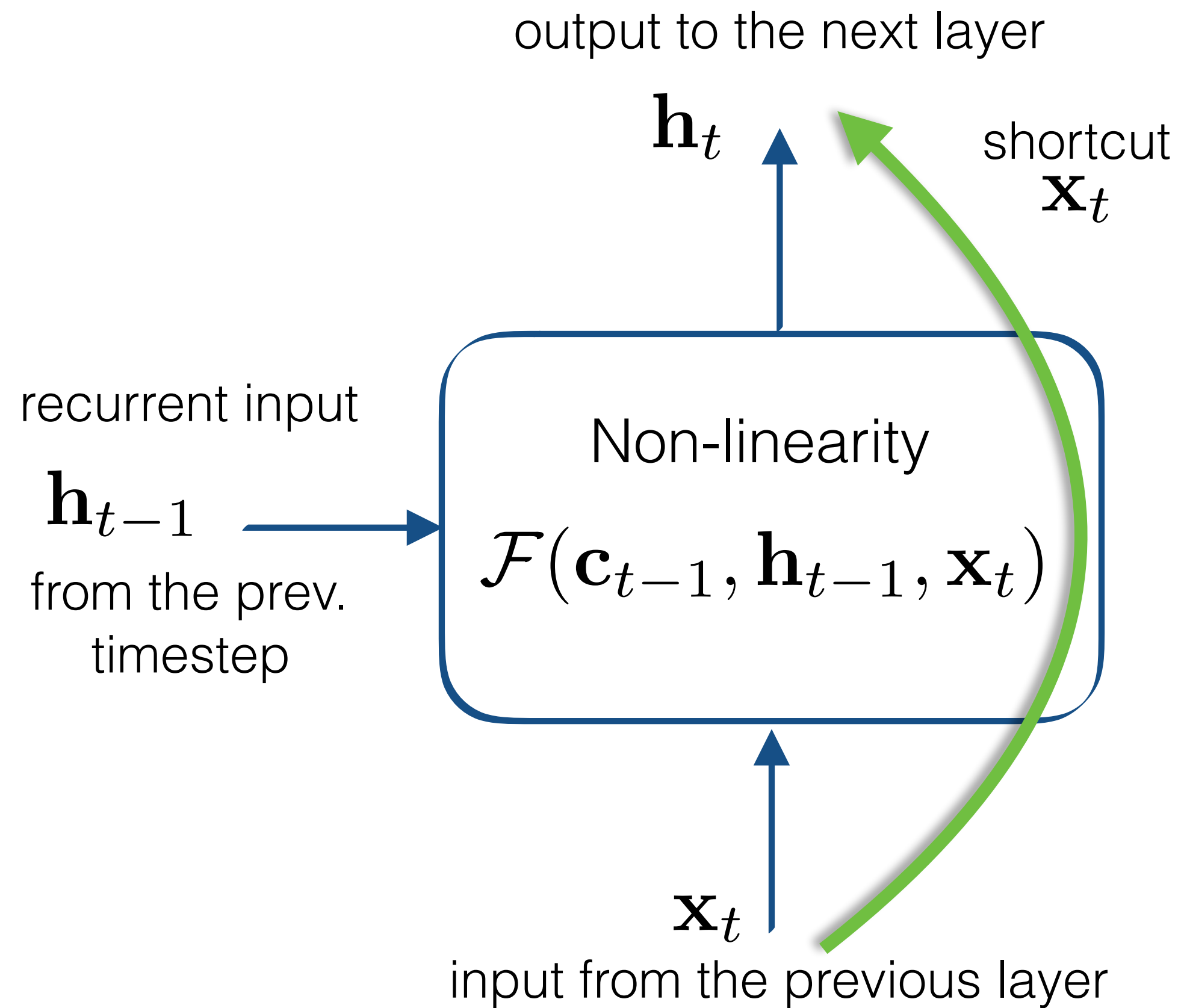


Figure from He et al. (2017)

Residual / Highway Connections



new output:

residual net $\mathbf{h}_t + \mathbf{x}_t$

gated highway network:

$$\mathbf{r}_t \circ \mathbf{h}_t + (1 - \mathbf{r}_t) \circ \mathbf{x}_t$$
$$\mathbf{r}_t = \sigma(f(\mathbf{h}_{t-1}, \mathbf{x}_t))$$

References:

Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
Training Very Deep Networks, Srivastava et al., 2015

Figure from He et al. (2017)

Variational Dropout for RNN

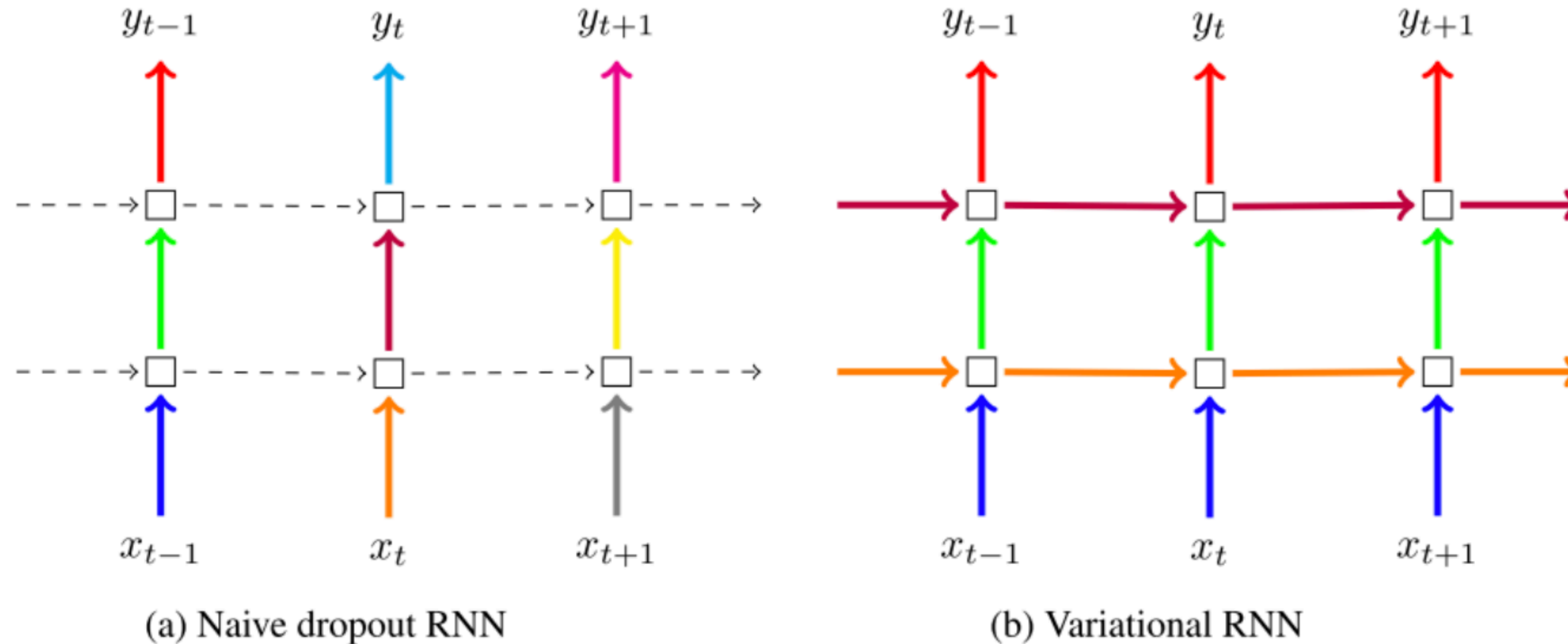


Fig 9. after Gal and Ghahramani (2015). Naive dropout (a) (eg Zaremba et al., 2014) uses different masks at different time steps, with no dropout on the recurrent layers. Variational Dropout (b) uses the same dropout mask at each time step, including the recurrent layers (colours representing dropout masks, solid lines representing dropout, dashed lines representing standard connections with no dropout).

10 Years of PropBank SRL

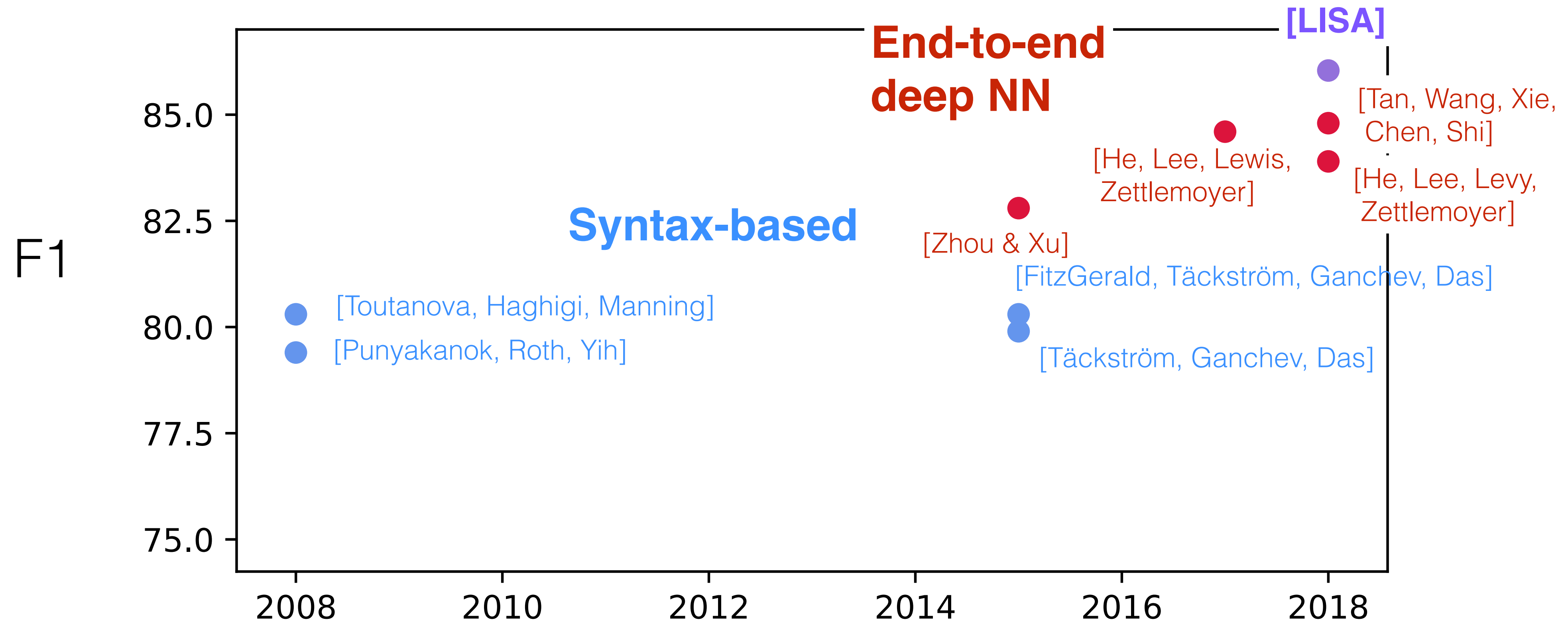
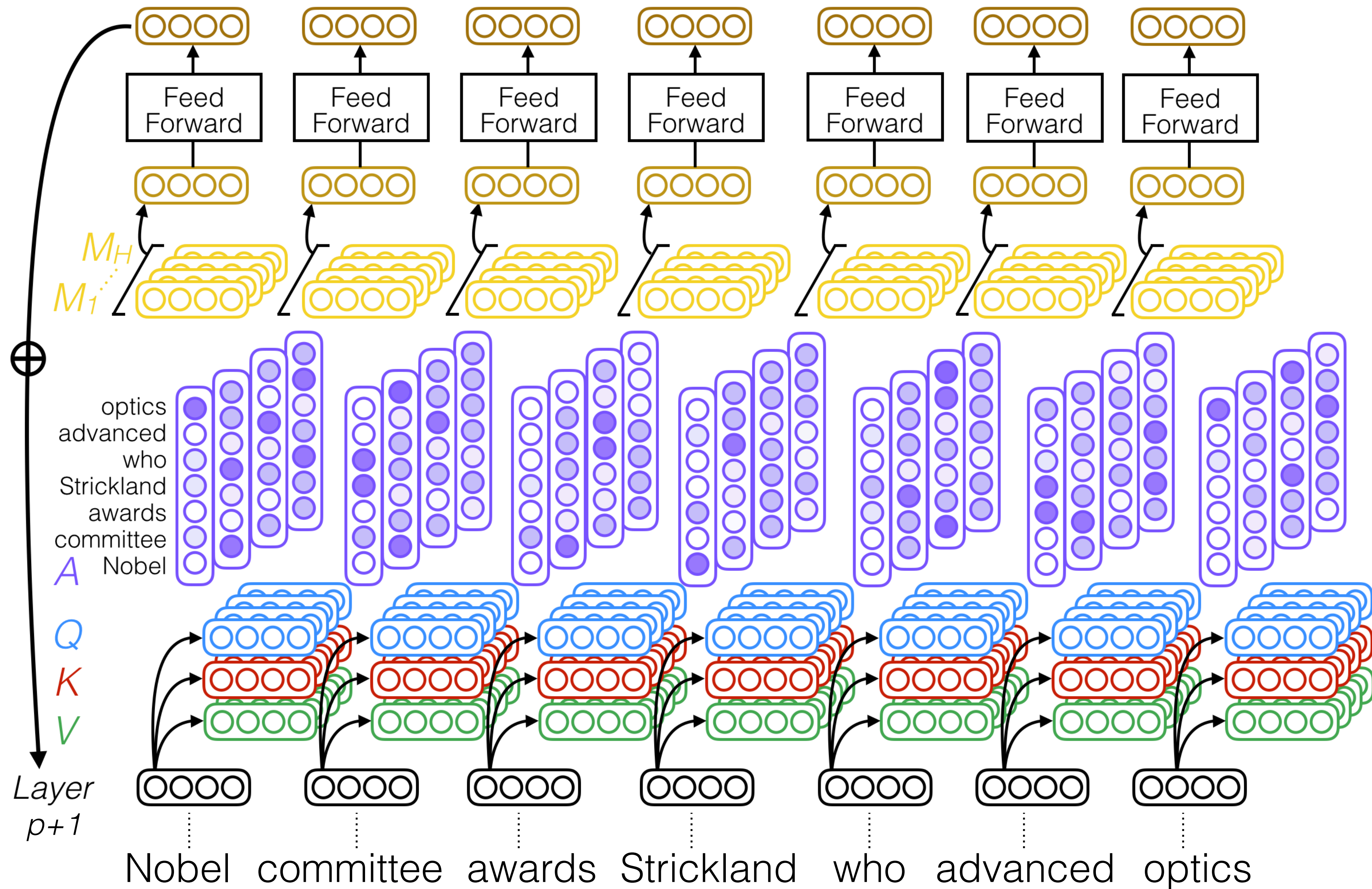


Figure from Strubell et al. (2018)

Semantic Role Labeling

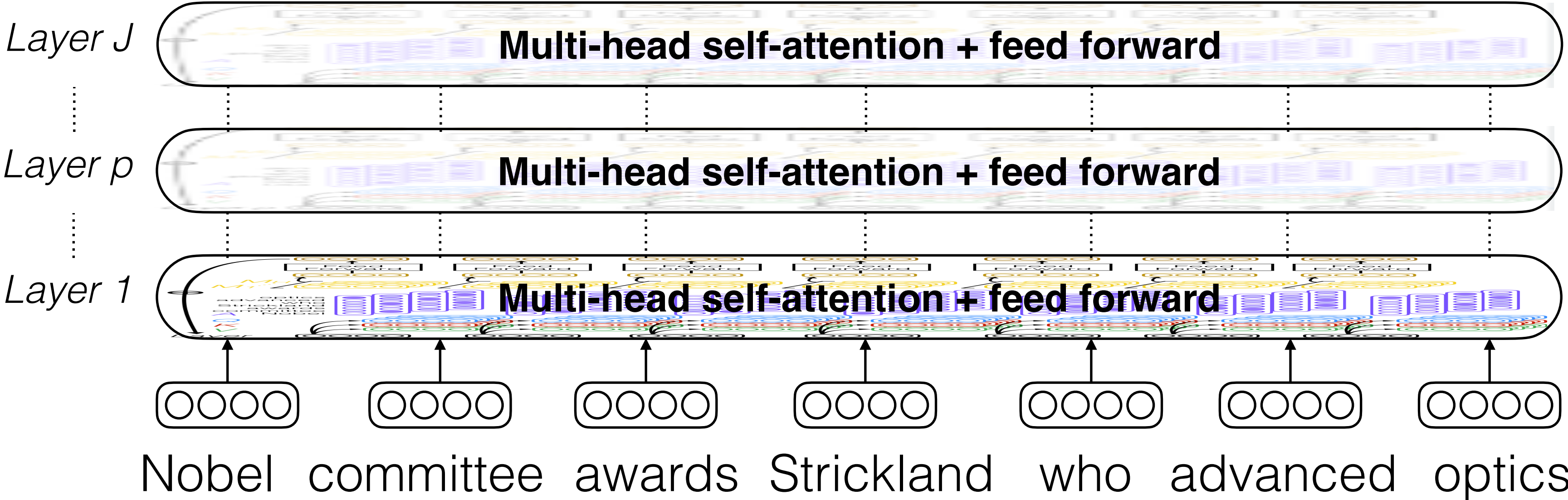
- ▶ Can we combine the two approaches — incorporate syntactic information into neural networks?
- ▶ Multi-task learning with related tasks, e.g., part-of-speech tagging, dependency parsing ...
- ▶ Syntactically-informed self-attention: use the Transformer to encode the sentence; in one head, token attends to its likely syntactic parents; in next layer, tokens observe all other parents.

Recall: Transformer (multi-head self-attention)

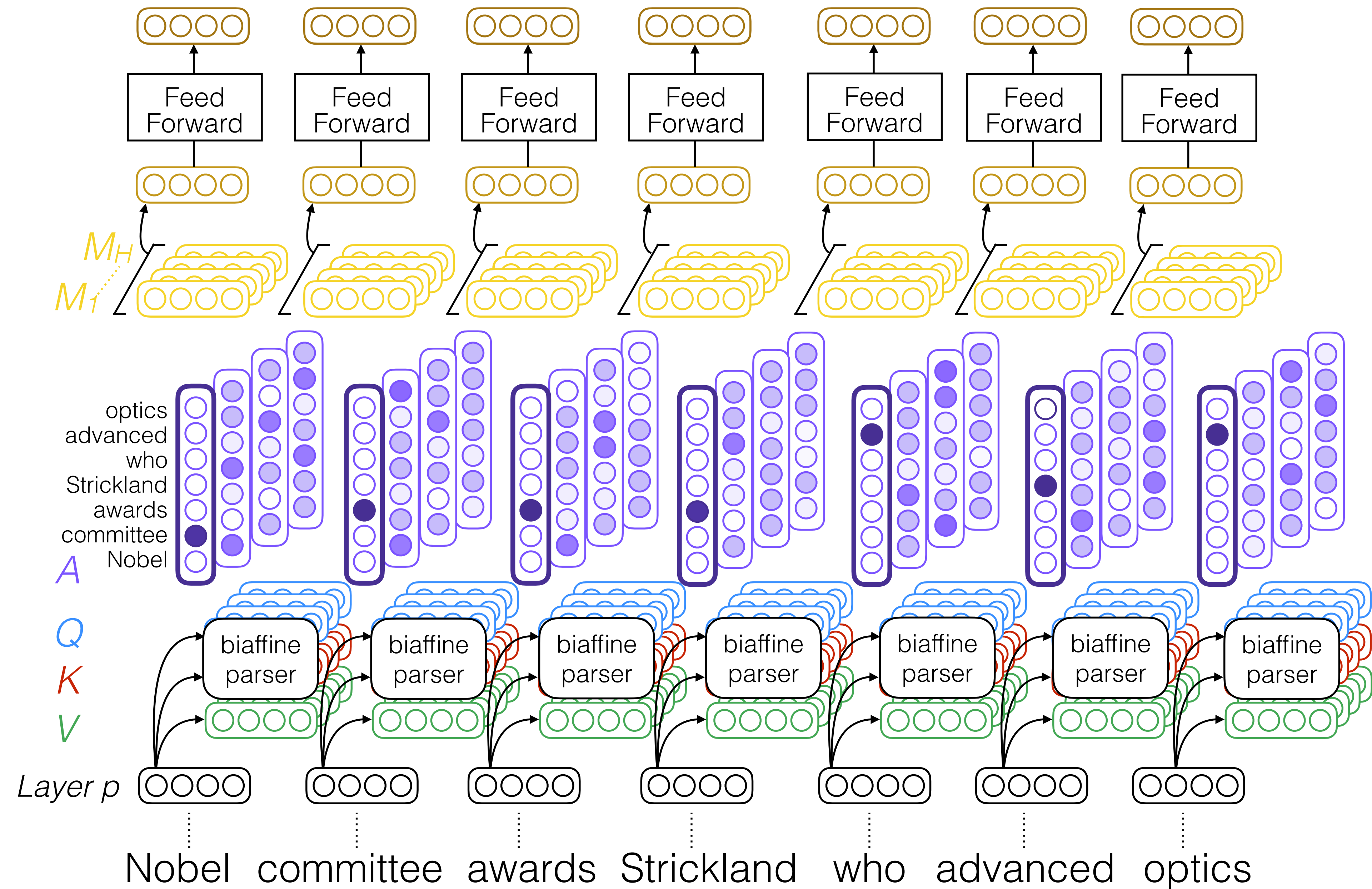


Recall: Transformer (multi-head self-attention)

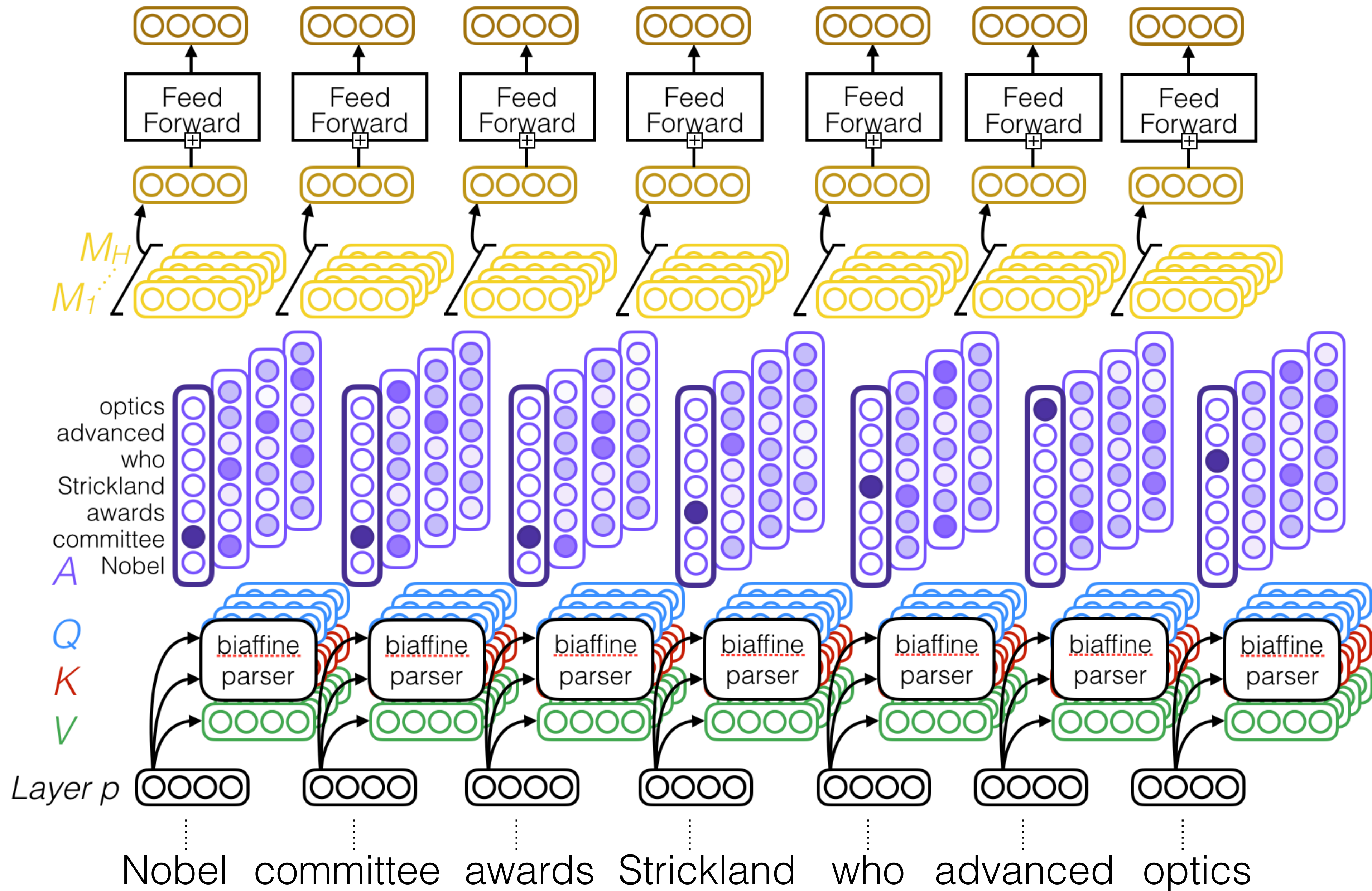
Slide Credit: Emma Strubell



Syntactically-Informed Self-Attention

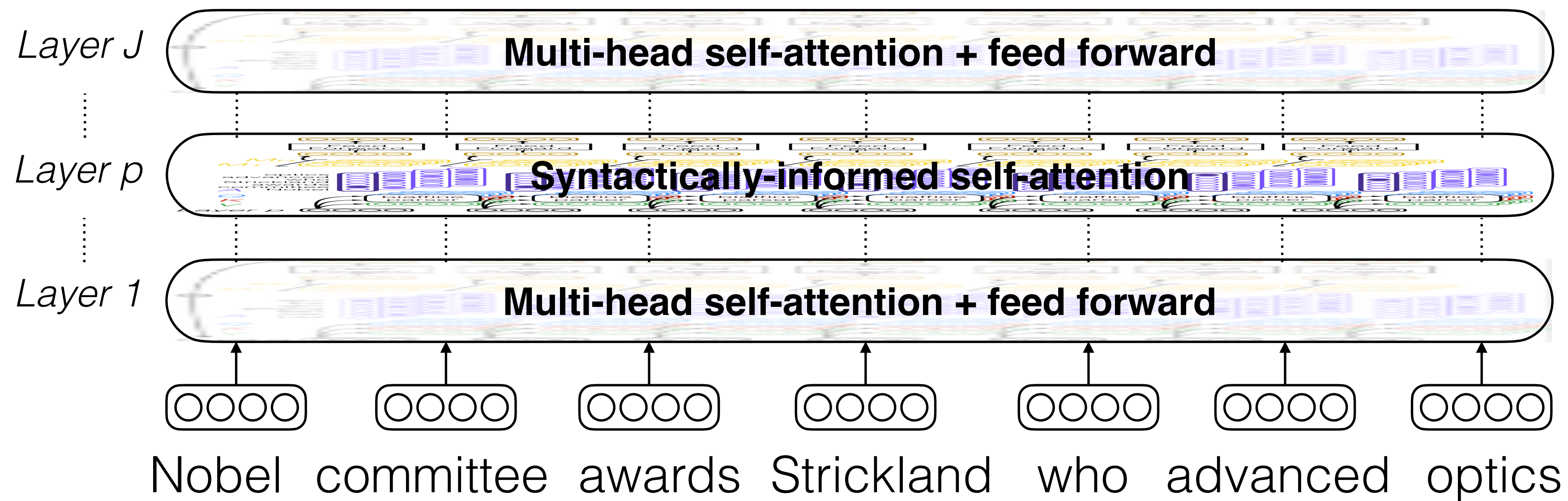


Syntactically-Informed Self-Attention



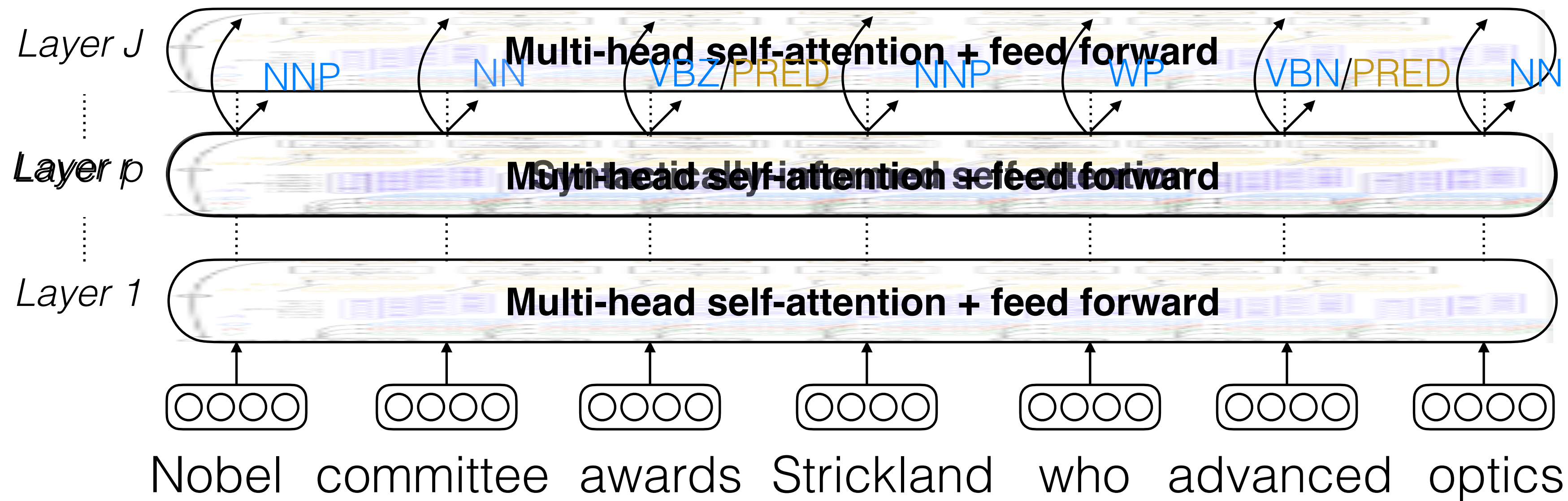
Linguistically-Informed Self-Attention

Slide Credit: Emma Strubell

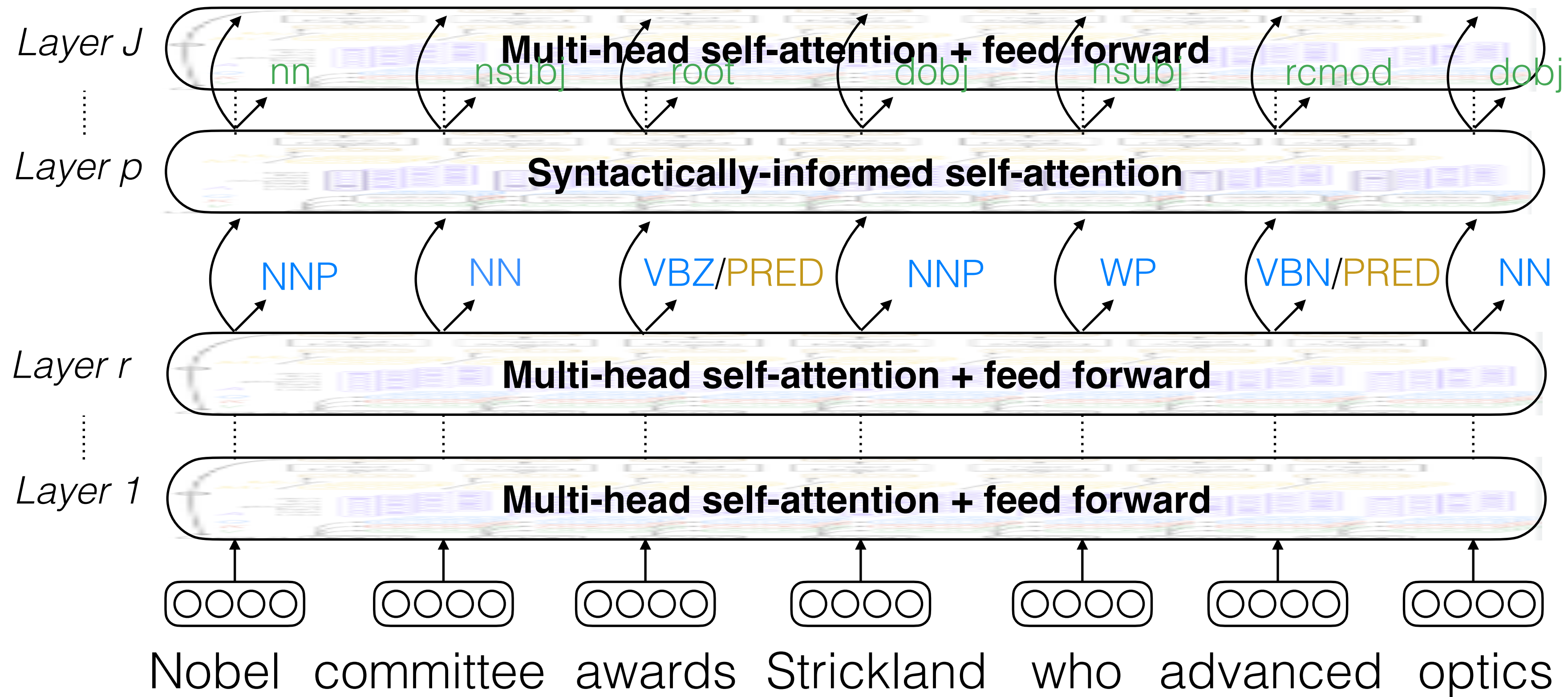


Linguistically-Informed Self-Attention

Slide Credit: Emma Strubell

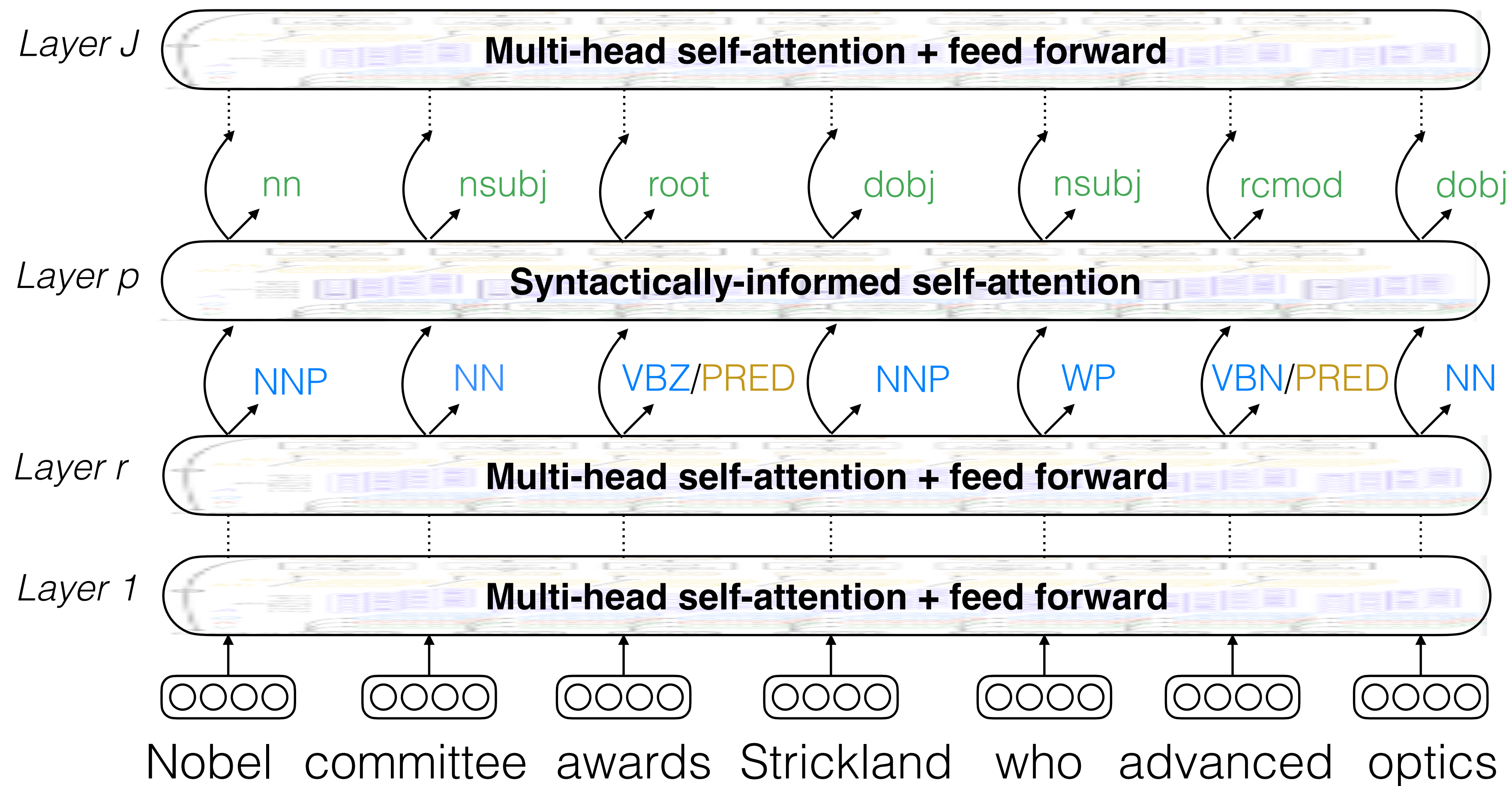


Linguistically-Informed Self-Attention



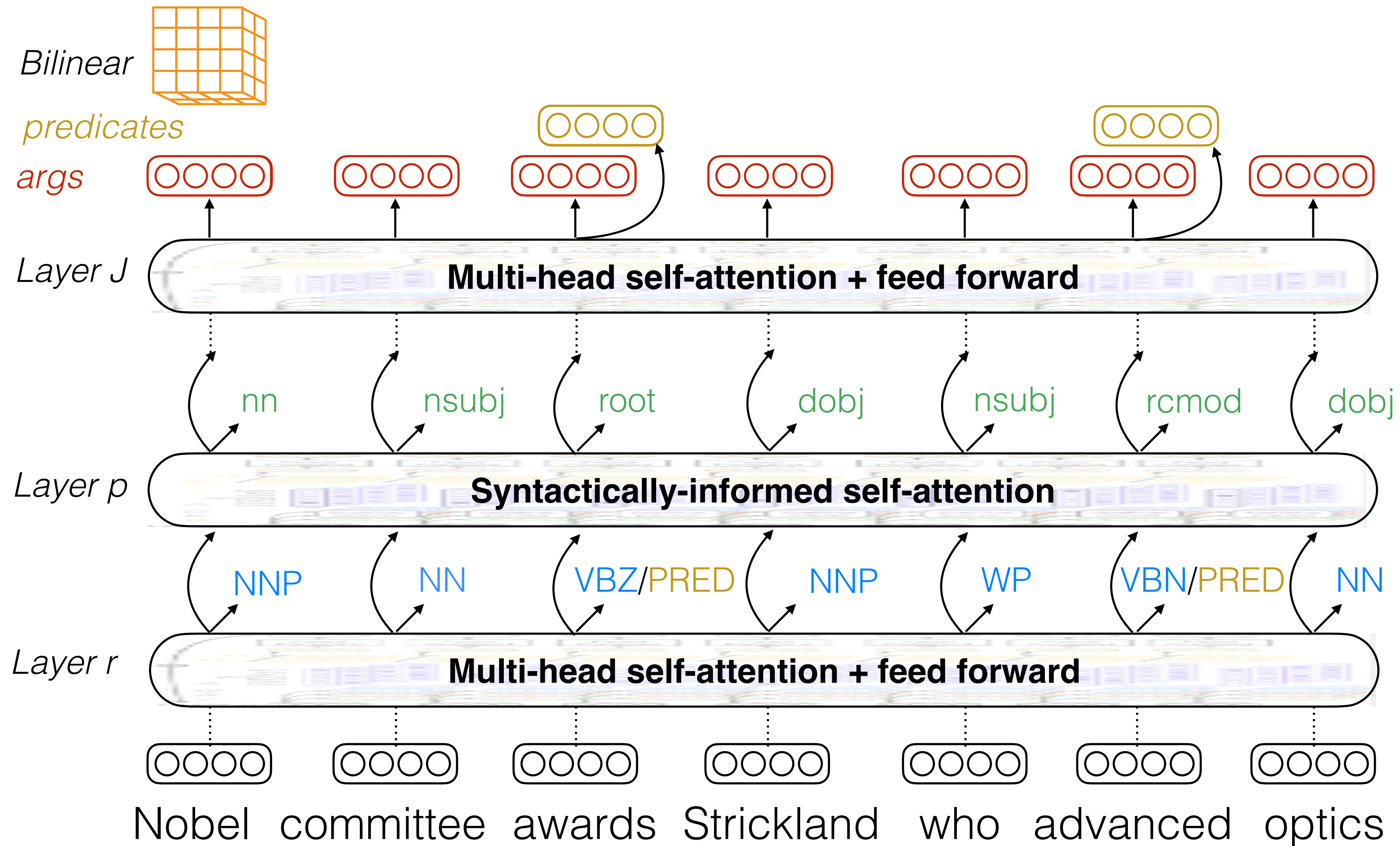
Linguistically-Informed Self-Attention

Slide Credit: Emma Strubell



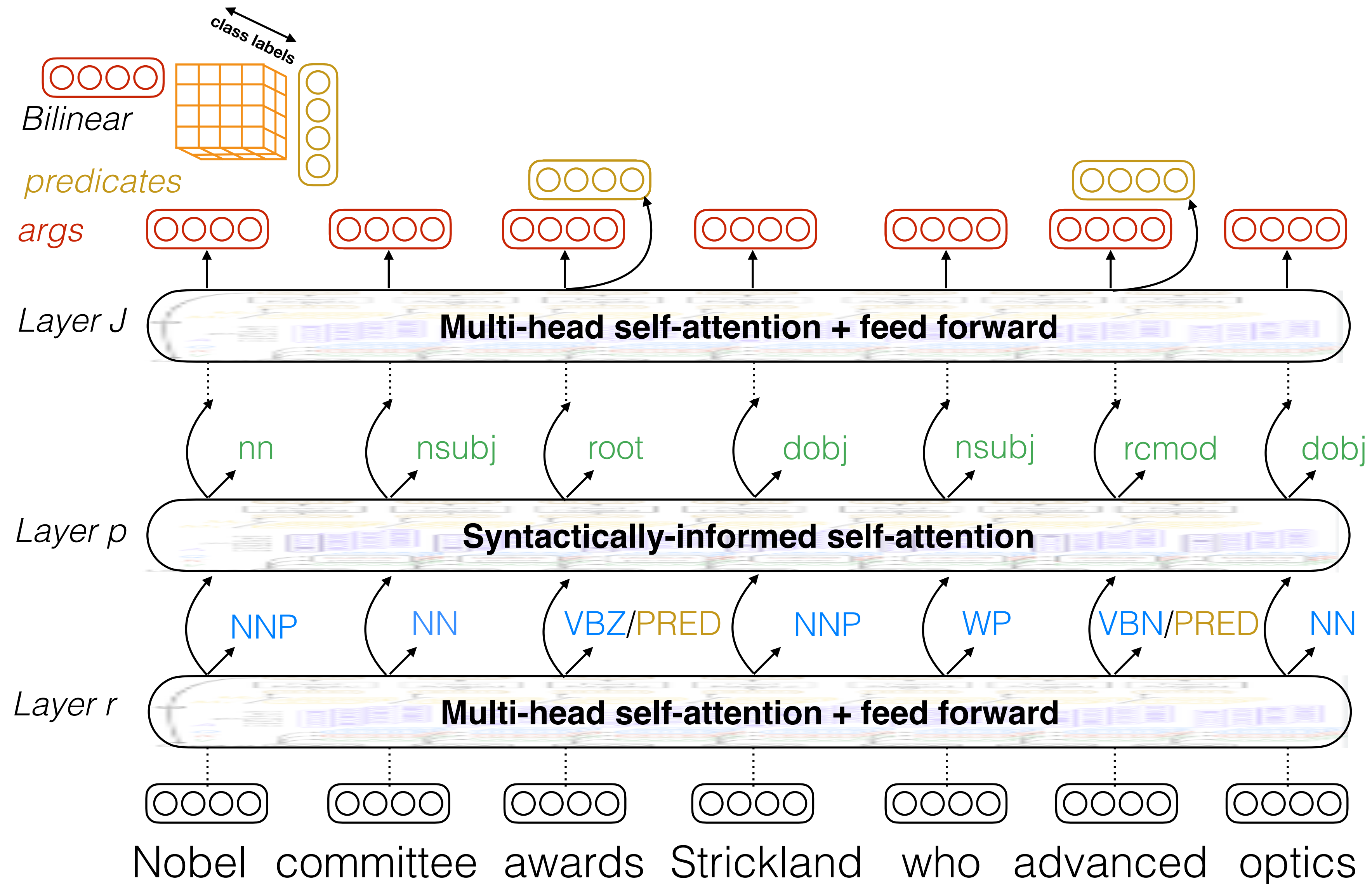
Linguistically-Informed Self-Attention

Slide Credit: Emma Strubell

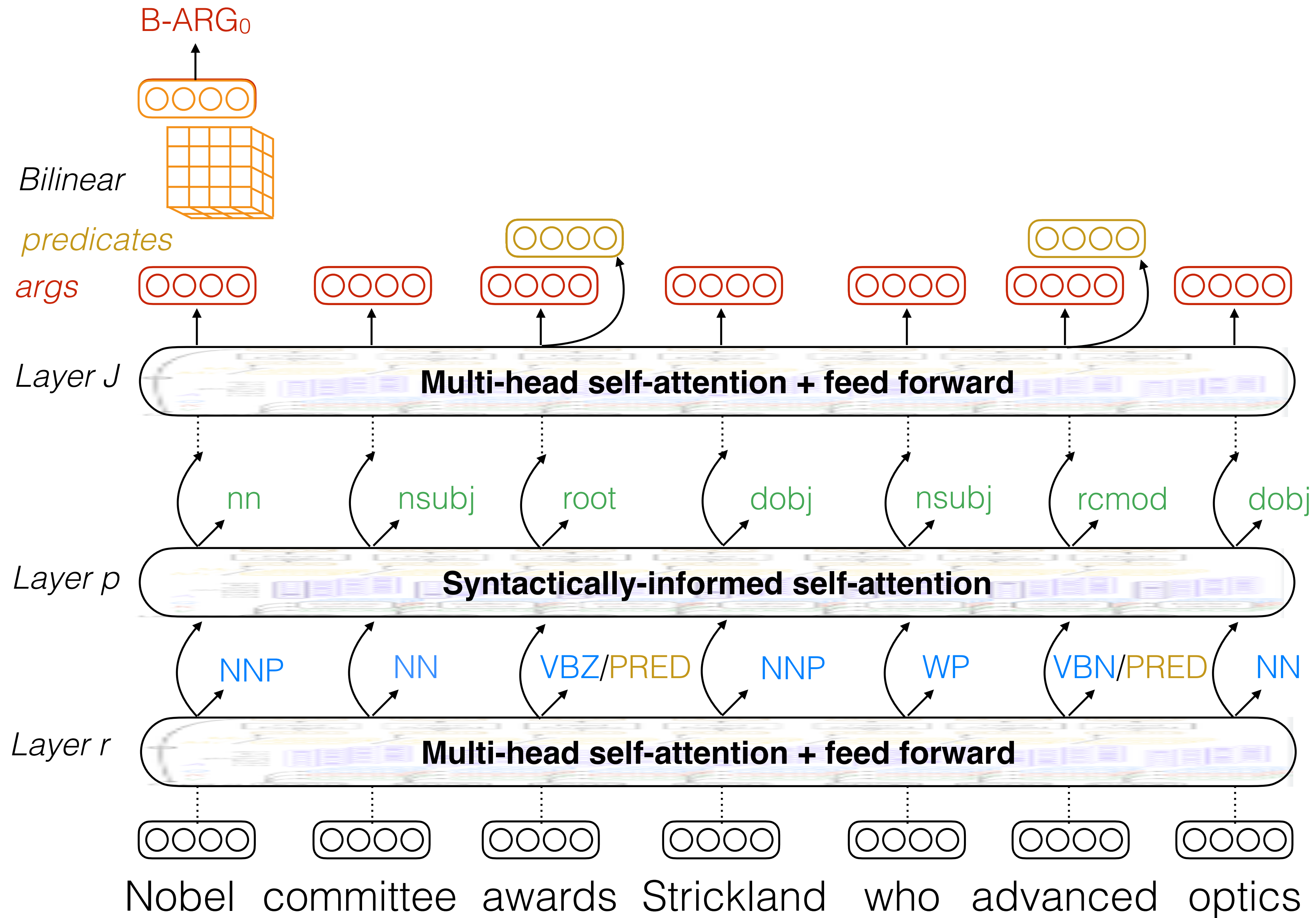


Linguistically-Informed Self-Attention

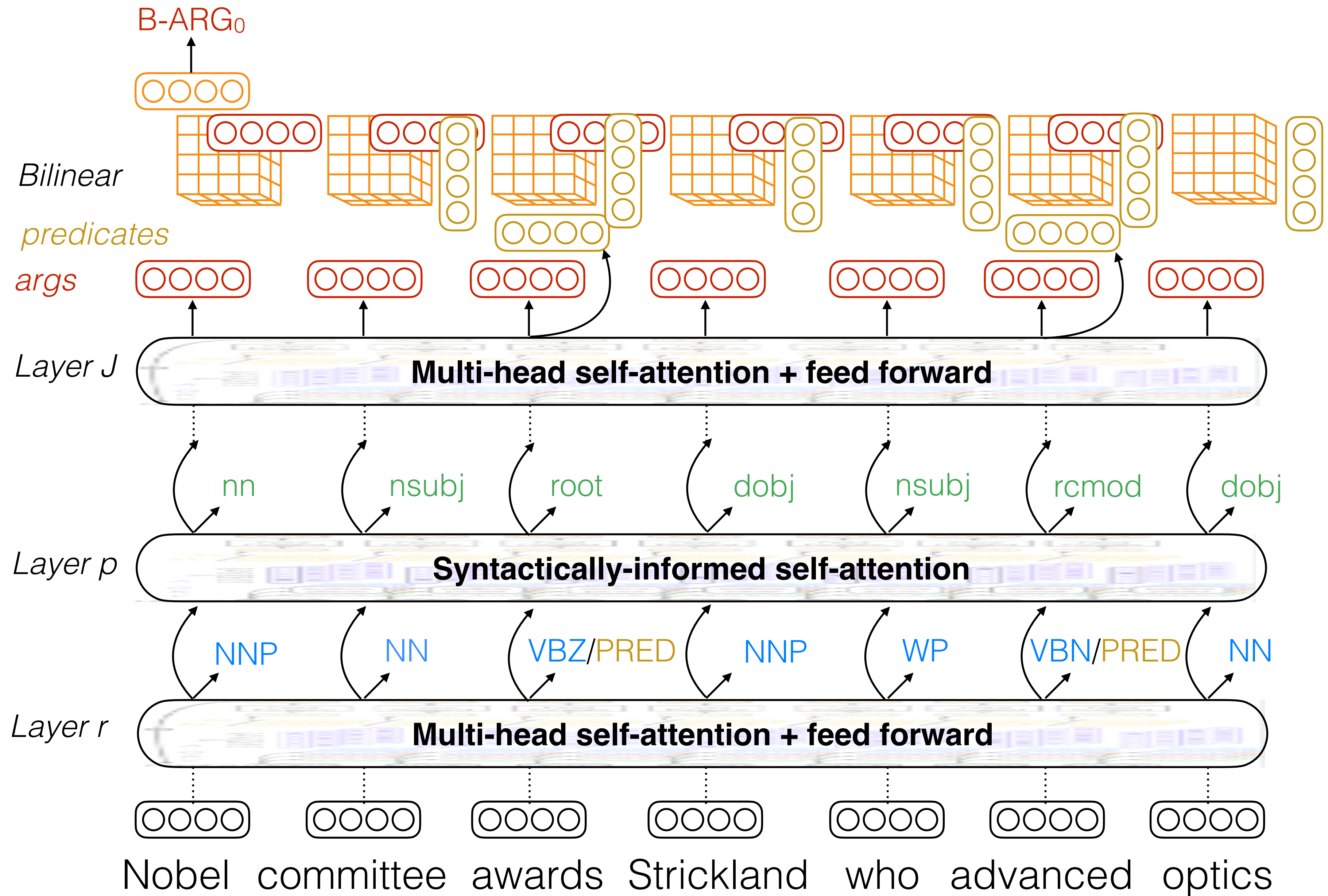
Slide Credit: Emma Strubell



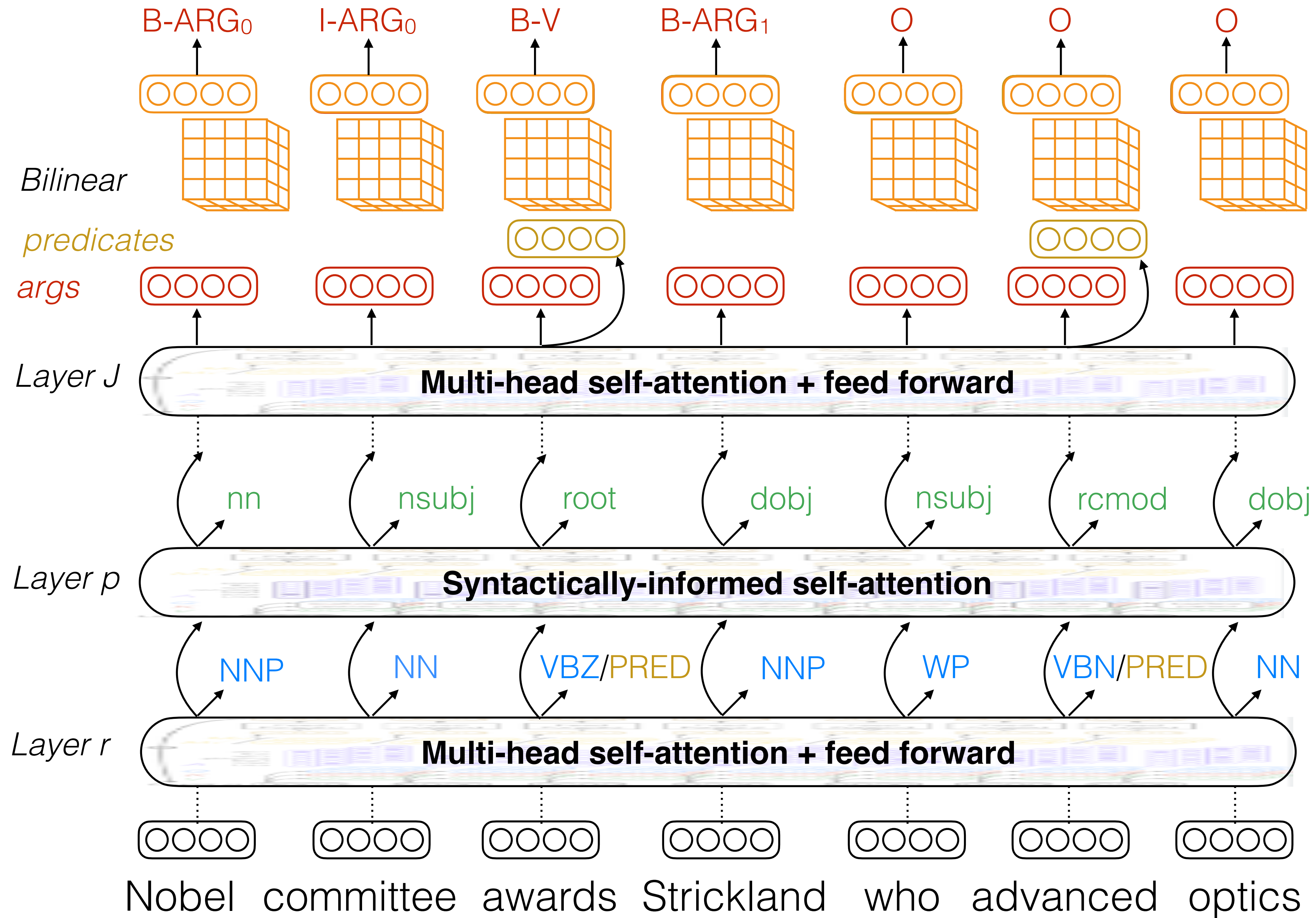
Linguistically-Informed Self-Attention



Linguistically-Informed Self-Attention

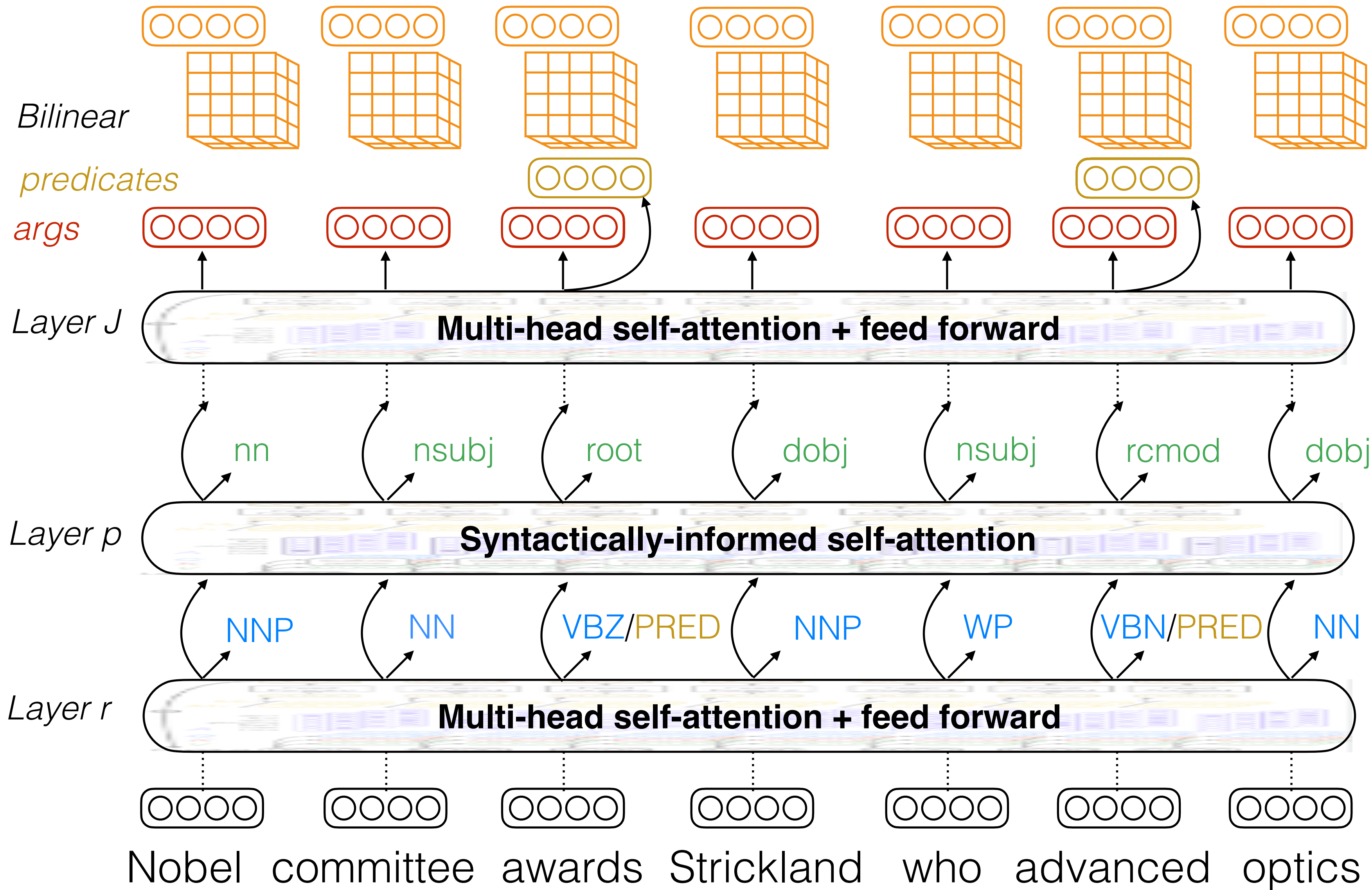


Linguistically-Informed Self-Attention



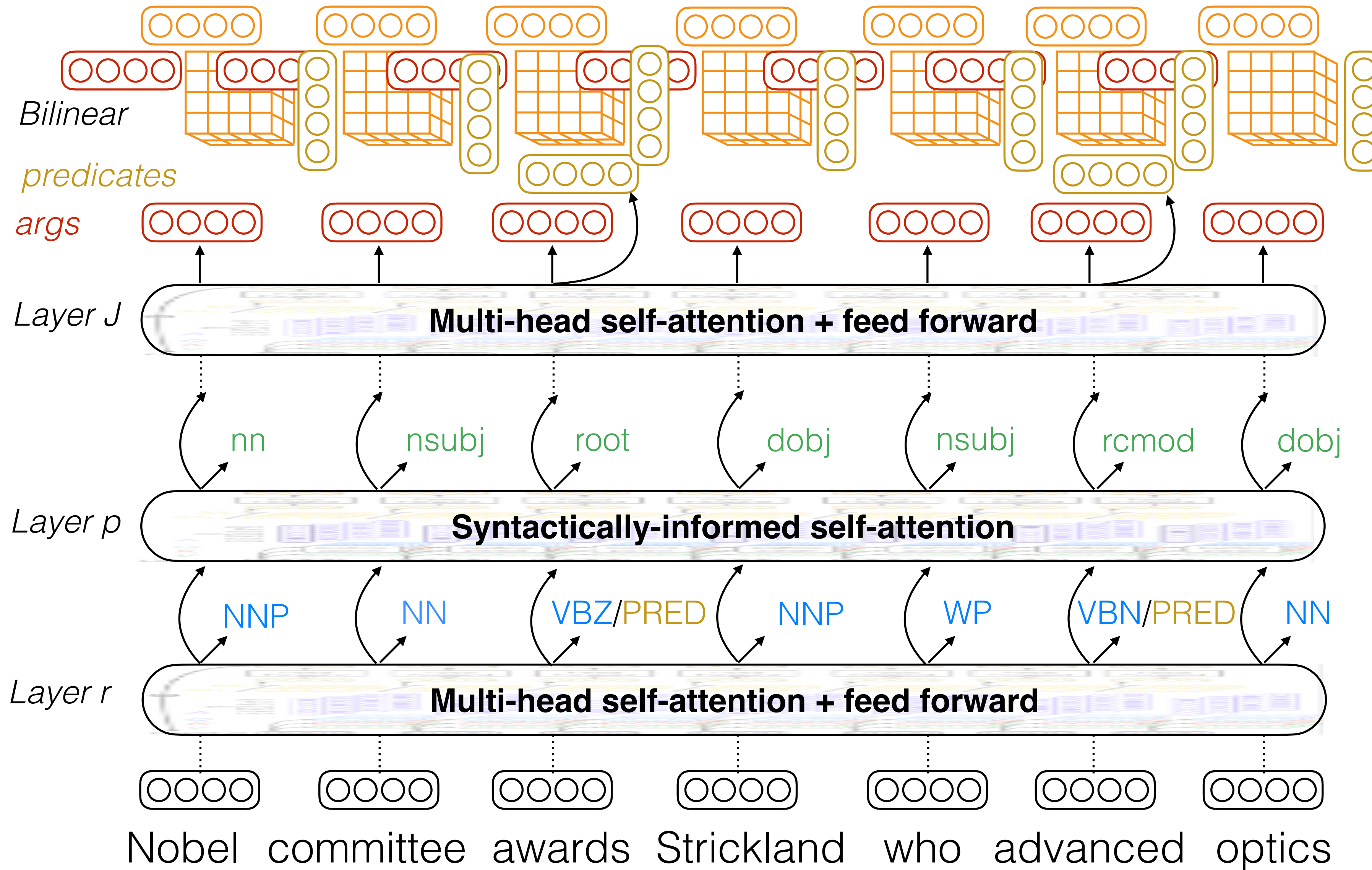
Linguistically-Informed Self-Attention

B-ARG₀ I-ARG₀ B-V B-ARG₁ O O O



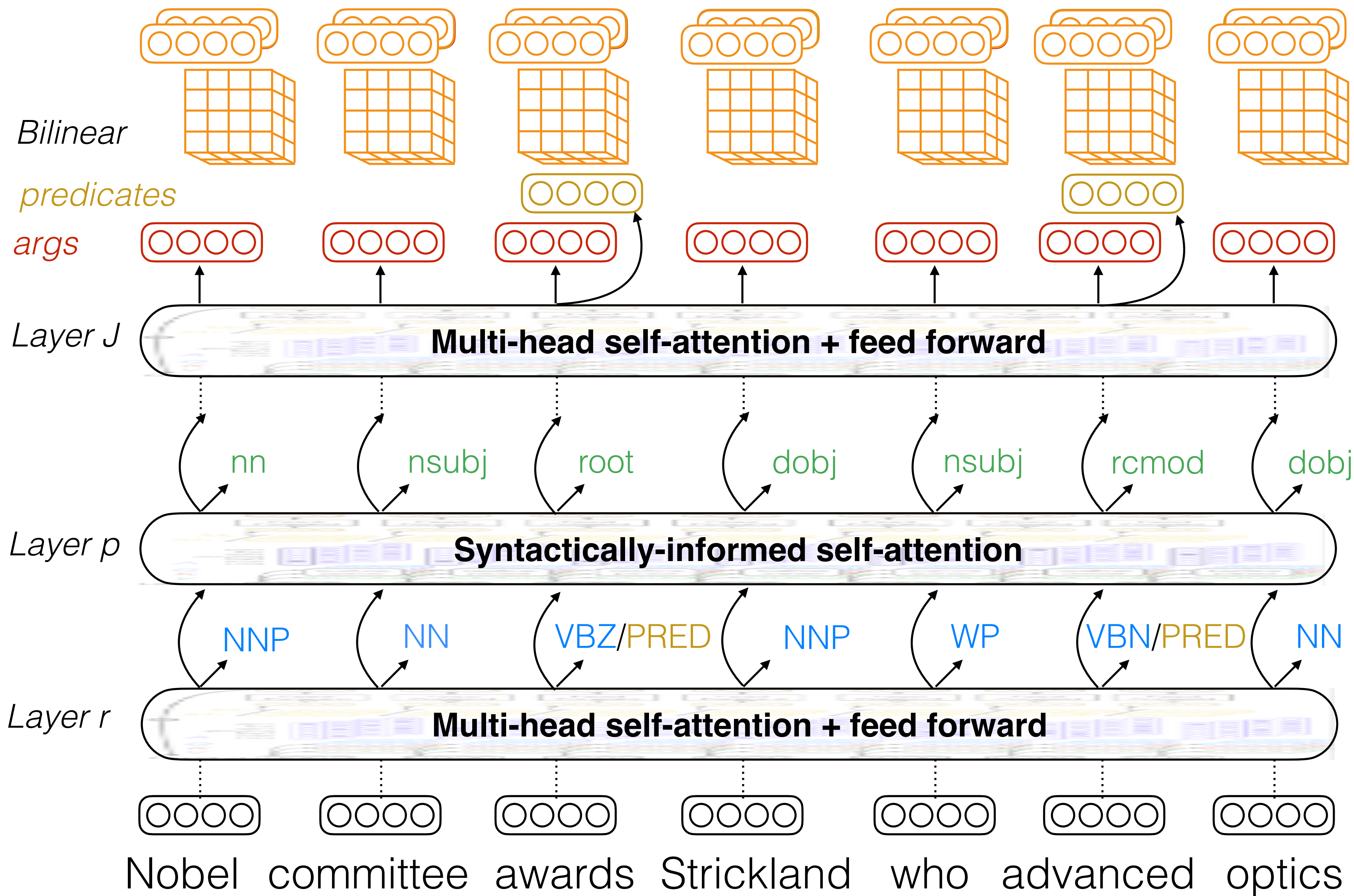
Linguistically-Informed Self-Attention

B-ARG₀ I-ARG₀ B-V B-ARG₁ O O O

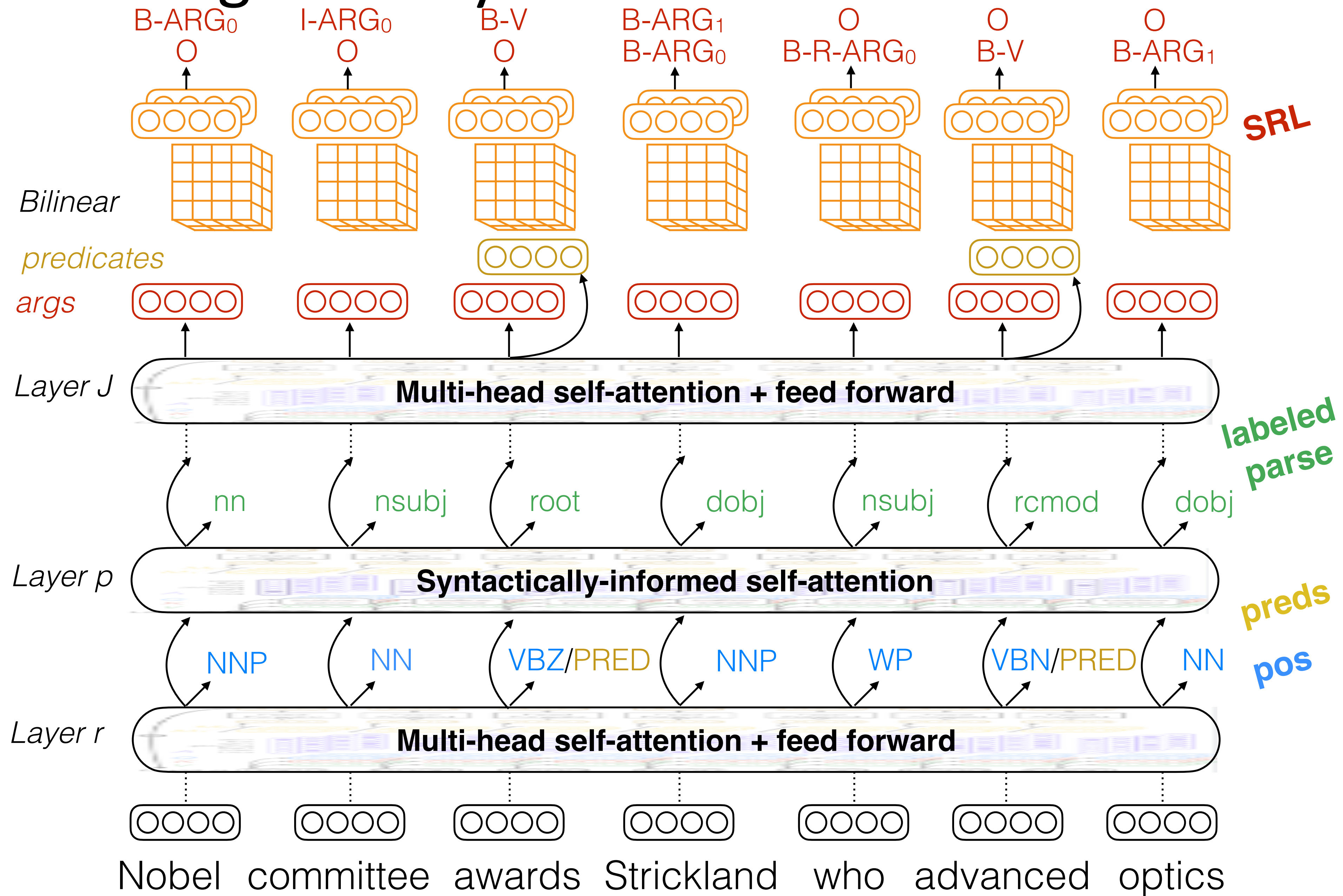


Linguistically-Informed Self-Attention



B-ARG₀ I-ARG₀ B-V B-ARG₁ O O O



Linguistically-Informed Self-Attention



Linguistically-Informed Self-Attention

	 GloVe	ELMo 
	in-domain (dev)	in-domain (dev)
He et al. 2017	81.5	---
He et al. 2018	81.6	85.3
SA	82.39	85.26
LISA	82.24	85.35
+D&M	83.58	85.17
<i>+Gold</i>	<i>86.81</i>	<i>87.63</i>

Biaffine Attention for Dependency Parsing

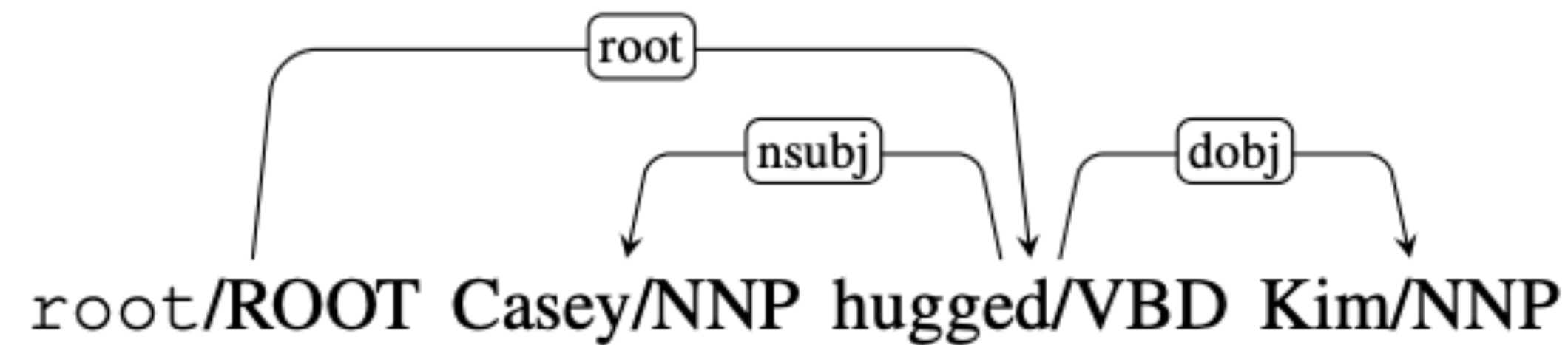


Figure 1: A dependency tree parse for *Casey hugged Kim*, including part-of-speech tags and a special `root` token. Directed edges (or arcs) with labels (or relations) connect the verb to the root and the arguments to the verb head.

Biaffine Attention for Dependency Parsing

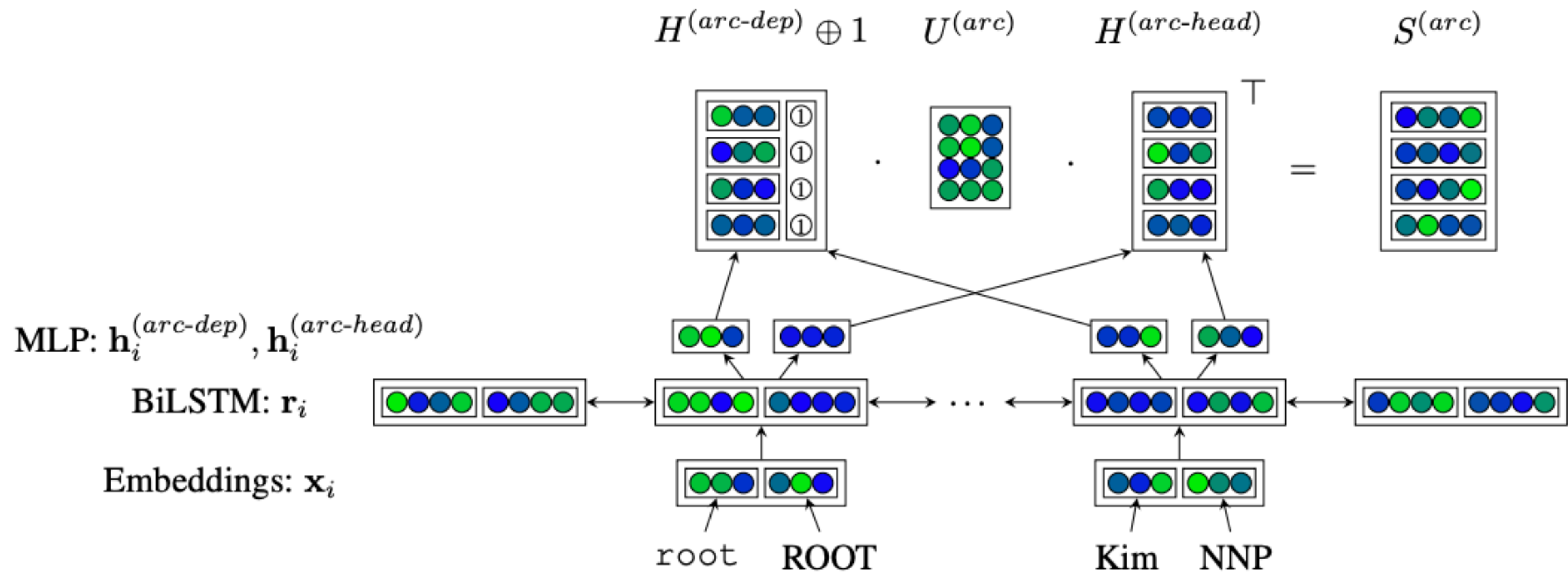
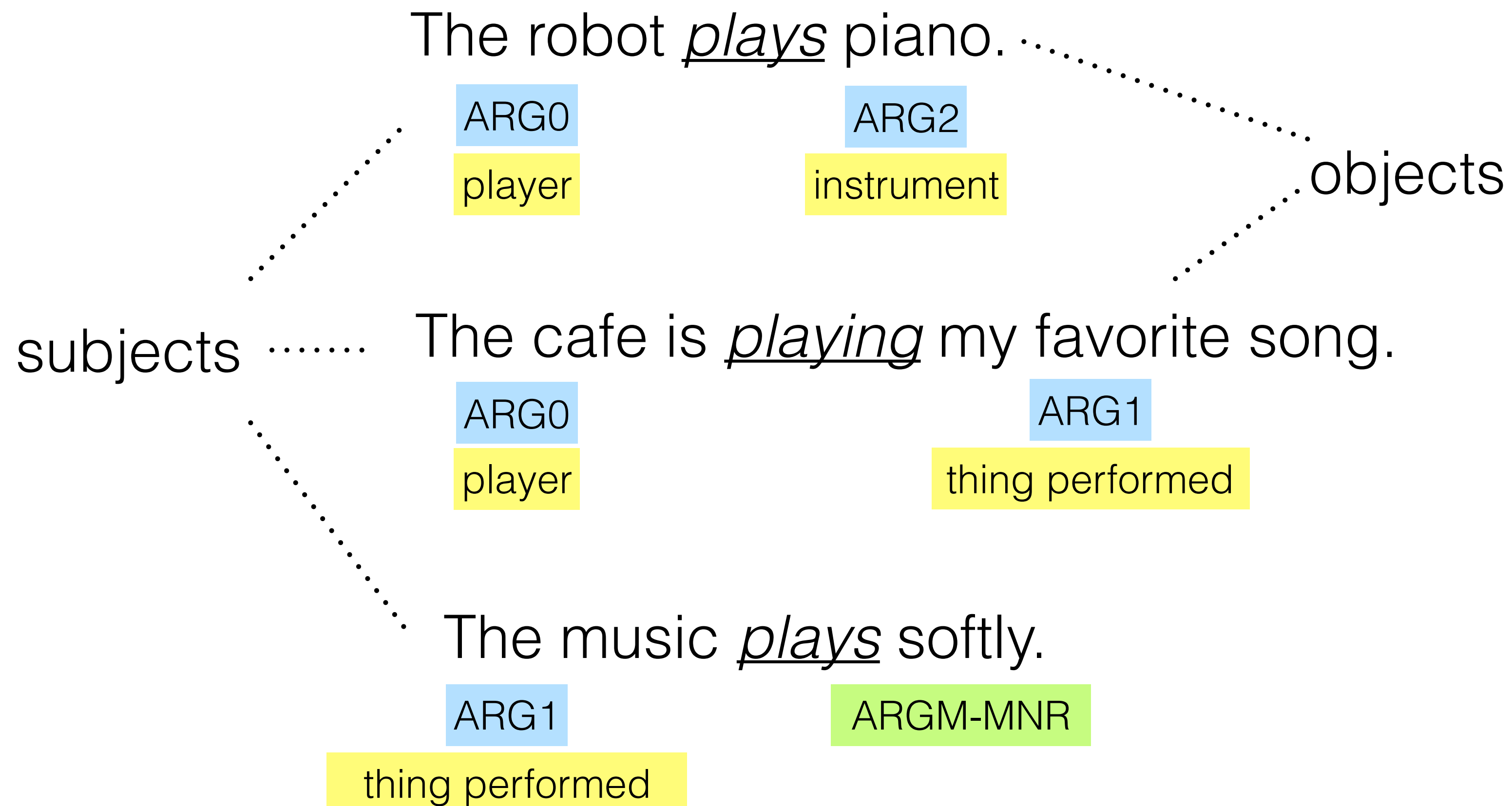


Figure 2: BiLSTM with deep biaffine attention to score each possible head for each dependent, applied to the sentence “Casey hugged Kim”. We reverse the order of the biaffine transformation here for clarity.

Why SRL is difficult? or NLP in general

- ▶ Syntactic Alternation



Why SRL is difficult? or NLP in general

- ▶ Prepositional Phrase (PP) Attachment

I eat [pasta] [with delight].

ARG0	ARG1	ARGM-MNR
eater	meal	manner



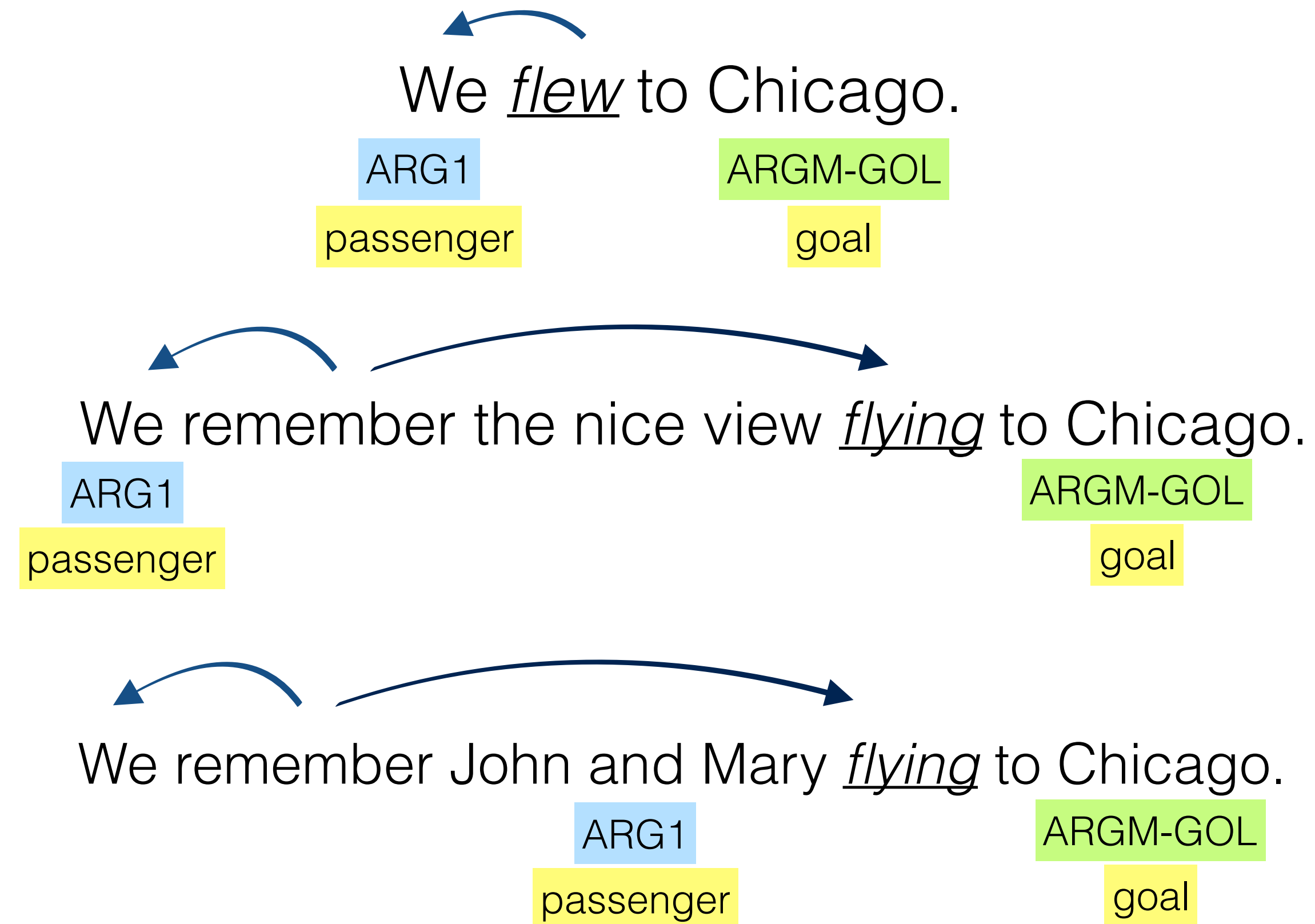
I eat [pasta with broccoli].

ARG0	ARG1
eater	meal



Why SRL is difficult? or NLP in general

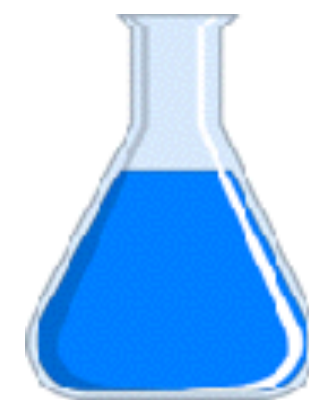
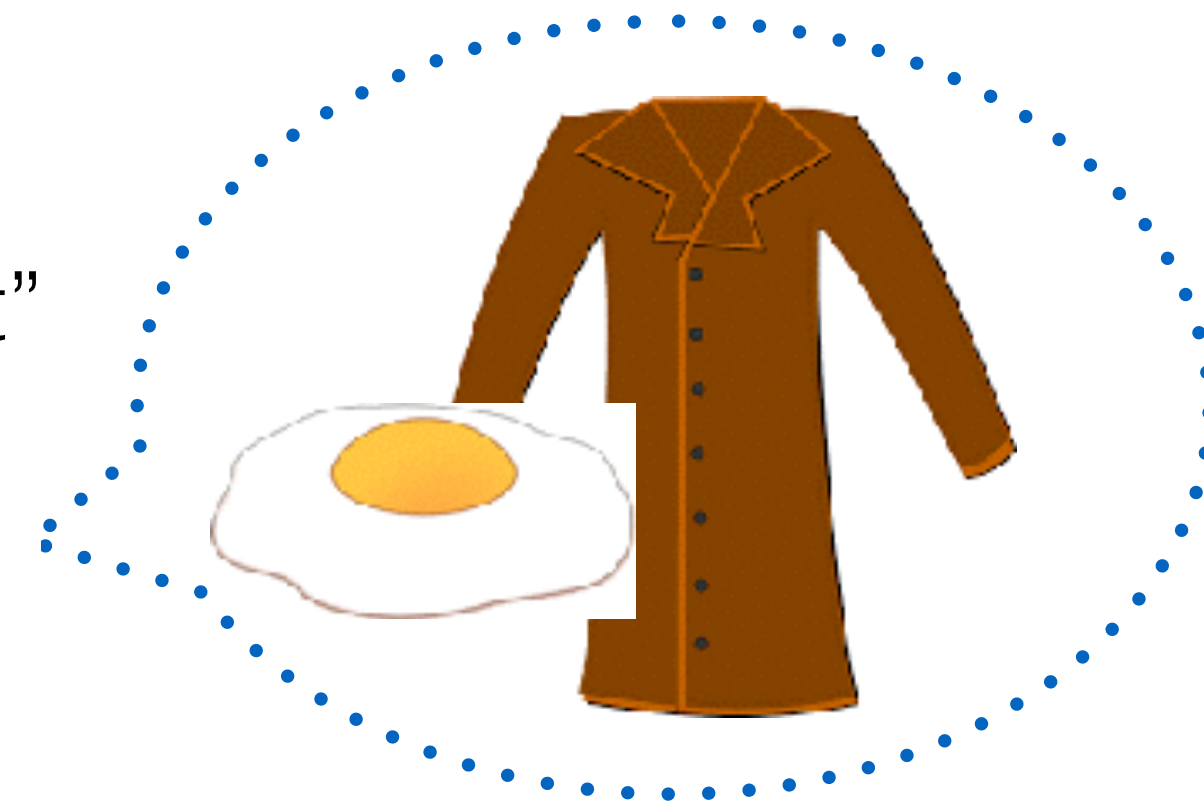
- ▶ Long Dependencies



Why SRL is difficult? or NLP in general

- ▶ Even harder for out-of-domain data

“Dip chicken breasts into eggs to coat”



Active, Ser133-phosphorylated CREB effects transcription of CRE-dependent genes via interaction with the 265-kDa ...

Slot Filling

Slot Filling: MUC

Template

(a)

SELLER	BUSINESS	ACQUIRED	PURCHASER
CSR Limited	Oil and Gas	Delhi Fund	Esso Inc.

Document

(b) [S CSR] has said that [S it] has sold [S its] [B oil interests] held in [A Delhi Fund]. [P Esso Inc.] did not disclose how much [P they] paid for [A Dehli].

- ▶ Key aspect: need to combine information across multiple mentions of an entity using coreference

Slot Filling

- ▶ Most conservative, narrow form of IE

magnitude

time

Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

epicenter

Speaker: Alan Clark speaker

“Gender Roles in the Holy Roman Empire” title

Allagher Center Main Auditorium location

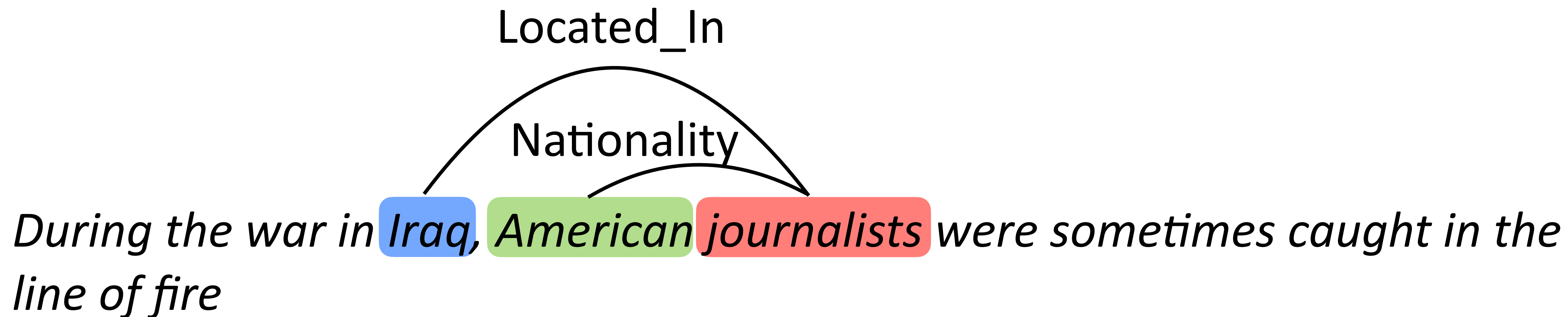
This talk will discuss...

- ▶ Old work: HMMs, later CRFs trained per role

Relation Extraction

Relation Extraction

- ▶ Extract entity-relation-entity triples from a fixed inventory



- ▶ Use NER-like system to identify entity spans, classify relations between entity pairs with a classifier
- ▶ Systems can be feature-based or neural, look at surface words, syntactic features (dependency paths), semantic roles
- ▶ Problem: limited data for scaling to big ontologies ACE (2003-2005)

Hearst Patterns

- ▶ Syntactic patterns especially for finding hypernym-hyponym pairs (“is a” relations)

Y is a X

Berlin is a city

X such as [list]

cities such as Berlin, Paris, and London.

other X including Y

other cities including Berlin

- ▶ Totally unsupervised way of harvesting world knowledge for tasks like parsing and coreference (Bansal and Klein, 2011-2012)

Distant Supervision

- ▶ Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data
- ▶ If two entities in a relation appear in the same sentence, assume the sentence expresses the relation

Director

[Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers' story

Allison co-produced the Academy Award-winning [Saving Private Ryan], directed by [Steven Spielberg]

Director

Distant Supervision

- ▶ Learn decently accurate classifiers for ~100 Freebase relations
- ▶ Could be used to crawl the web and expand our knowledge base

Relation name	100 instances			1000 instances		
	Syn	Lex	Both	Syn	Lex	Both
/film/director/film	0.49	0.43	0.44	0.49	0.41	0.46
/film/writer/film	0.70	0.60	0.65	0.71	0.61	0.69
/geography/river/basin_countries	0.65	0.64	0.67	0.73	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	0.70	0.72	0.68	0.72
/location/location/contains	0.81	0.89	0.84	0.85	0.83	0.84
/location/us_county/county_seat	0.51	0.51	0.53	0.47	0.57	0.42
/music/artist/origin	0.64	0.66	0.71	0.61	0.63	0.60
/people/deceased_person/place_of_death	0.80	0.79	0.81	0.80	0.81	0.78
/people/person/nationality	0.61	0.70	0.72	0.56	0.61	0.63
/people/person/place_of_birth	0.78	0.77	0.78	0.88	0.85	0.91
Average	0.67	0.66	0.69	0.68	0.67	0.67

Distant Supervision

- ▶ Inherently have noise in training data, need special methods (e.g., multi-instance learning) to handle false positives AND false negatives.

Freebase

Entity 1	Entity 2	Relation
Thailand	Bangkok	/location/country/capital

Sentences mentioning the two entities:

1. *Bangkok* is the most populous city of *Thailand*.
2. *Bangkok* grew rapidly during the 1960s through the 1980s and now exerts a significant impact among *Thailand's* politics, economy, education, media and modern society.
3. The nation of *Thailand* is about to get its very first visit ever from a president this weekend, President Obama, so the American Embassy in *Bangkok* is understandably very excited right now.

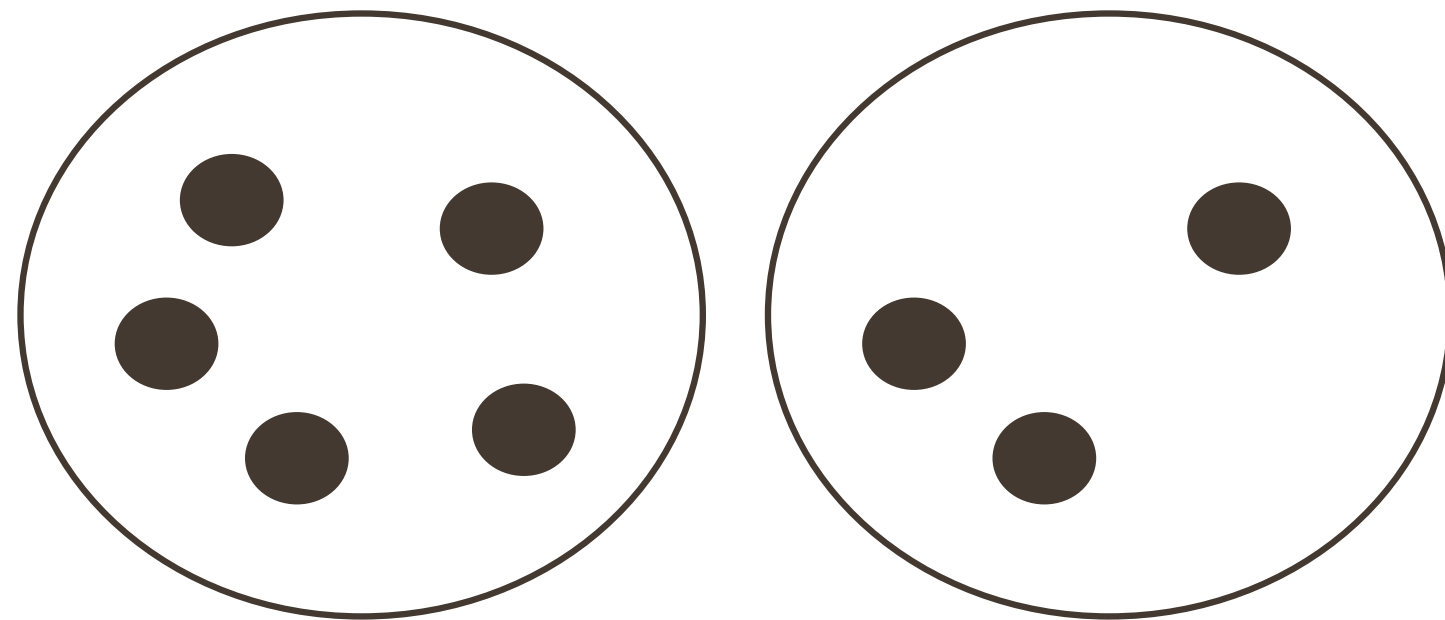
false positive

???

Multi-instance Learning

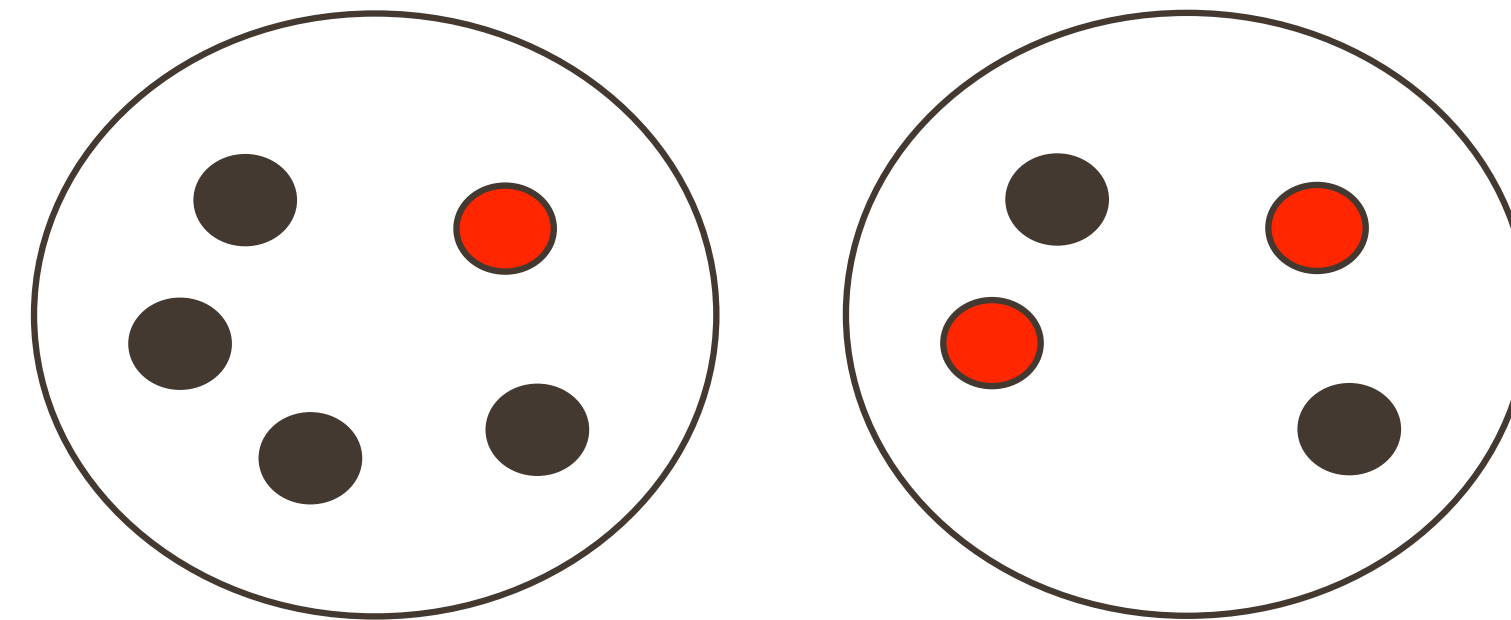
- ▶ Instead of labels on each individual instance, the learner only observes labels on bags of instances.

Negative Bags



A bag is labeled negative, if **all** the examples in it are negative

Positive Bags



A bag is labeled positive, if there is **at least one** positive example

Multi-instance Learning

- ▶ Handle false positives (sentences contain the entity pair but not the relation) and false negatives (due to incomplete knowledge base)

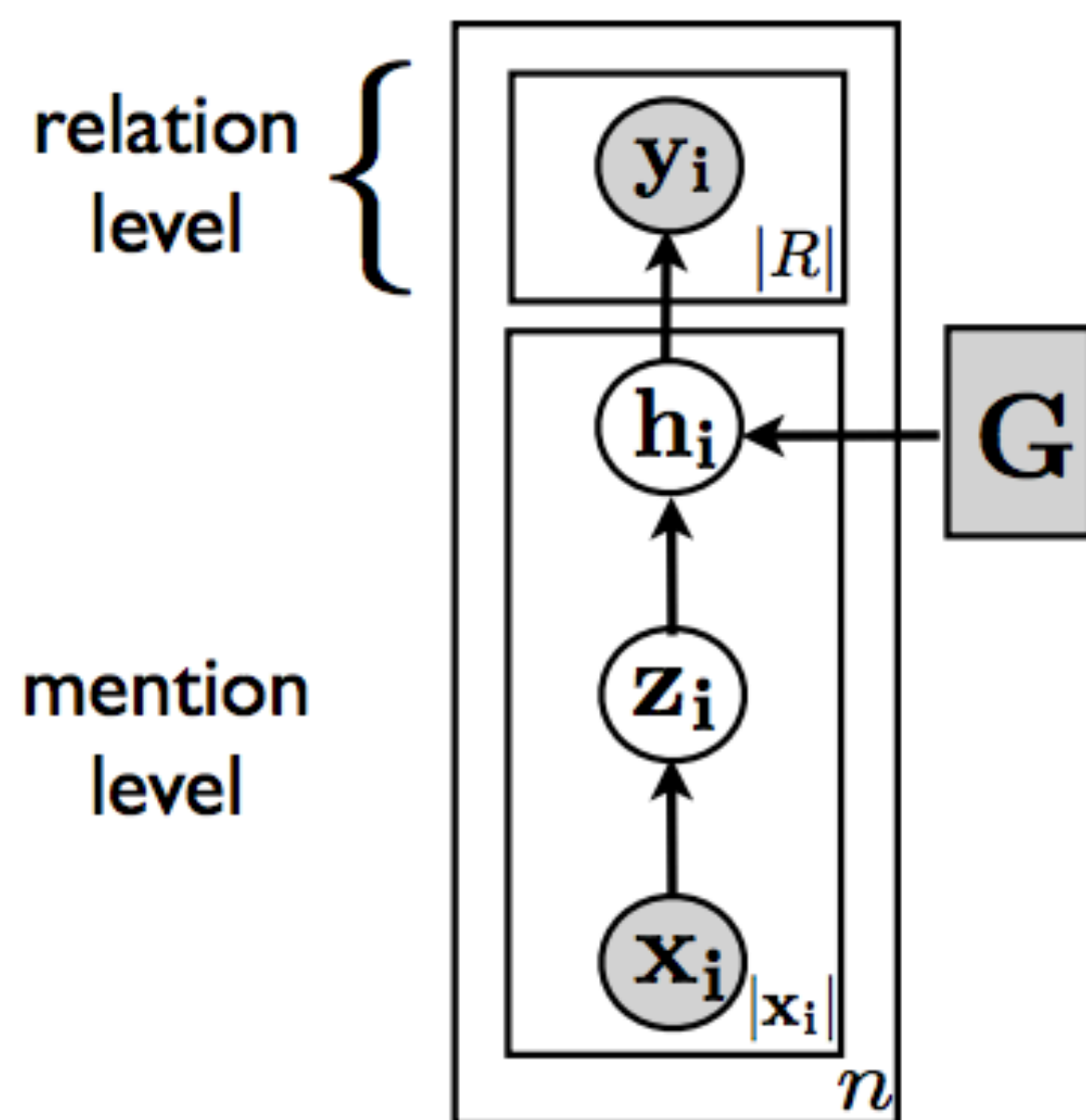
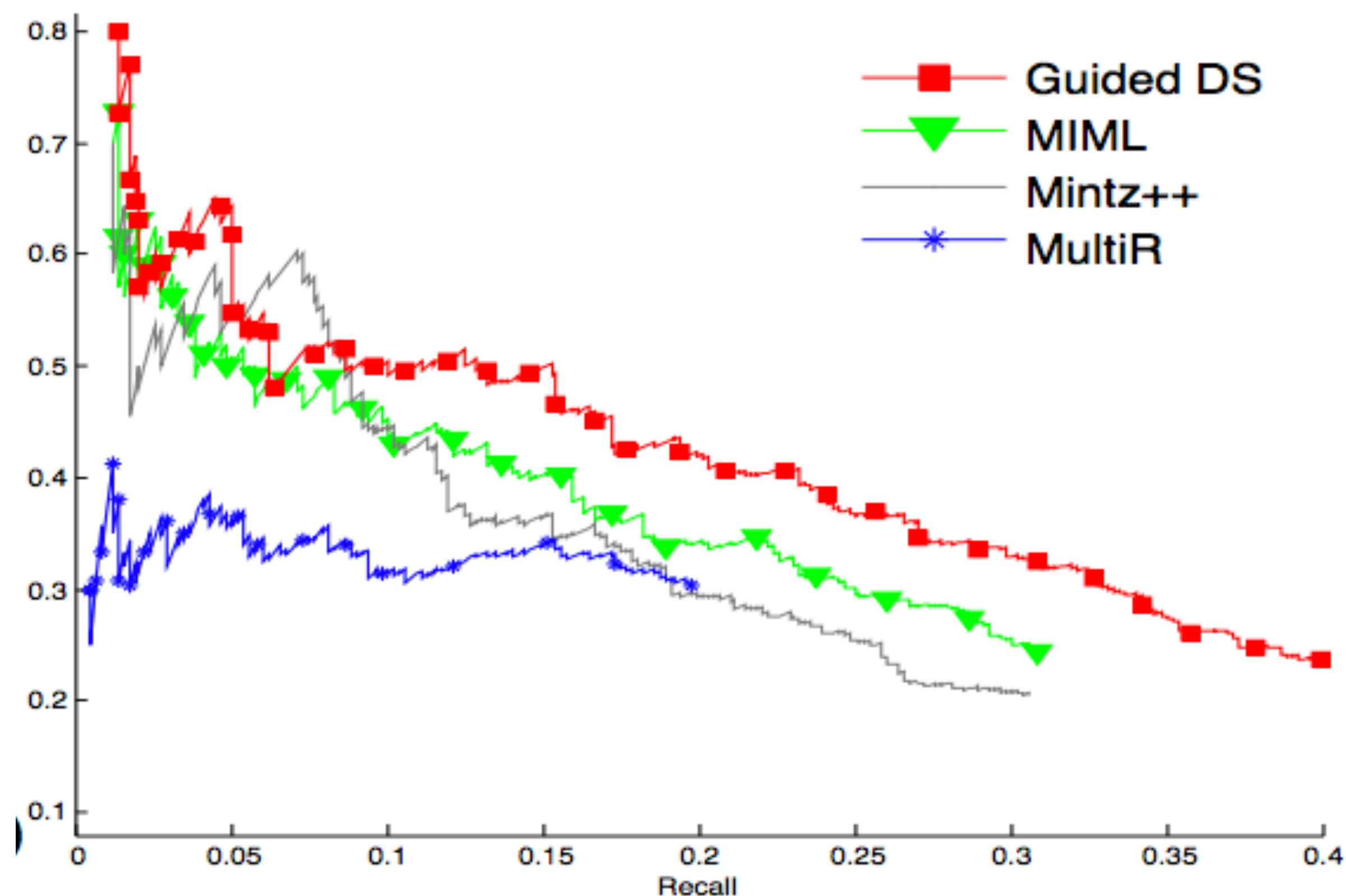


Figure 1: Plate diagram of Guided DS



Xu et al. (2013), Pershina et al. (2014), Tabassum (2016)

Multi-instance Learning

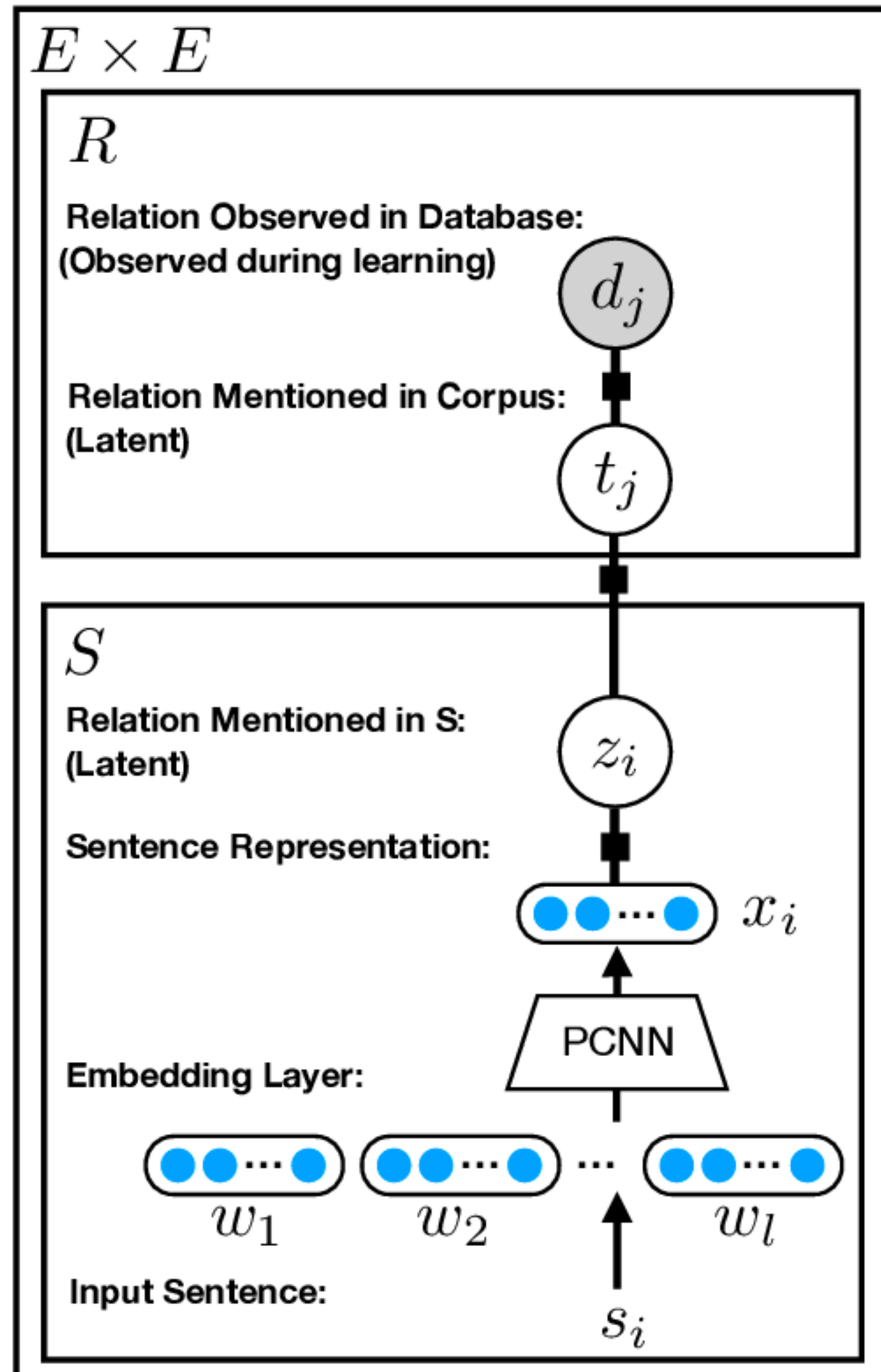


Figure 1: Plate representation of our proposed model. Plates represent replication; $E \times E$ is the number of entity pairs in the dataset, S is the number of sentences mentioning each entity pair and R is the number of relations. Arrows represent functions from input to output. Latent variables are represented as unshaded nodes. Factors over variables are represented as boxes.

Piecewise CNN (PCNN)

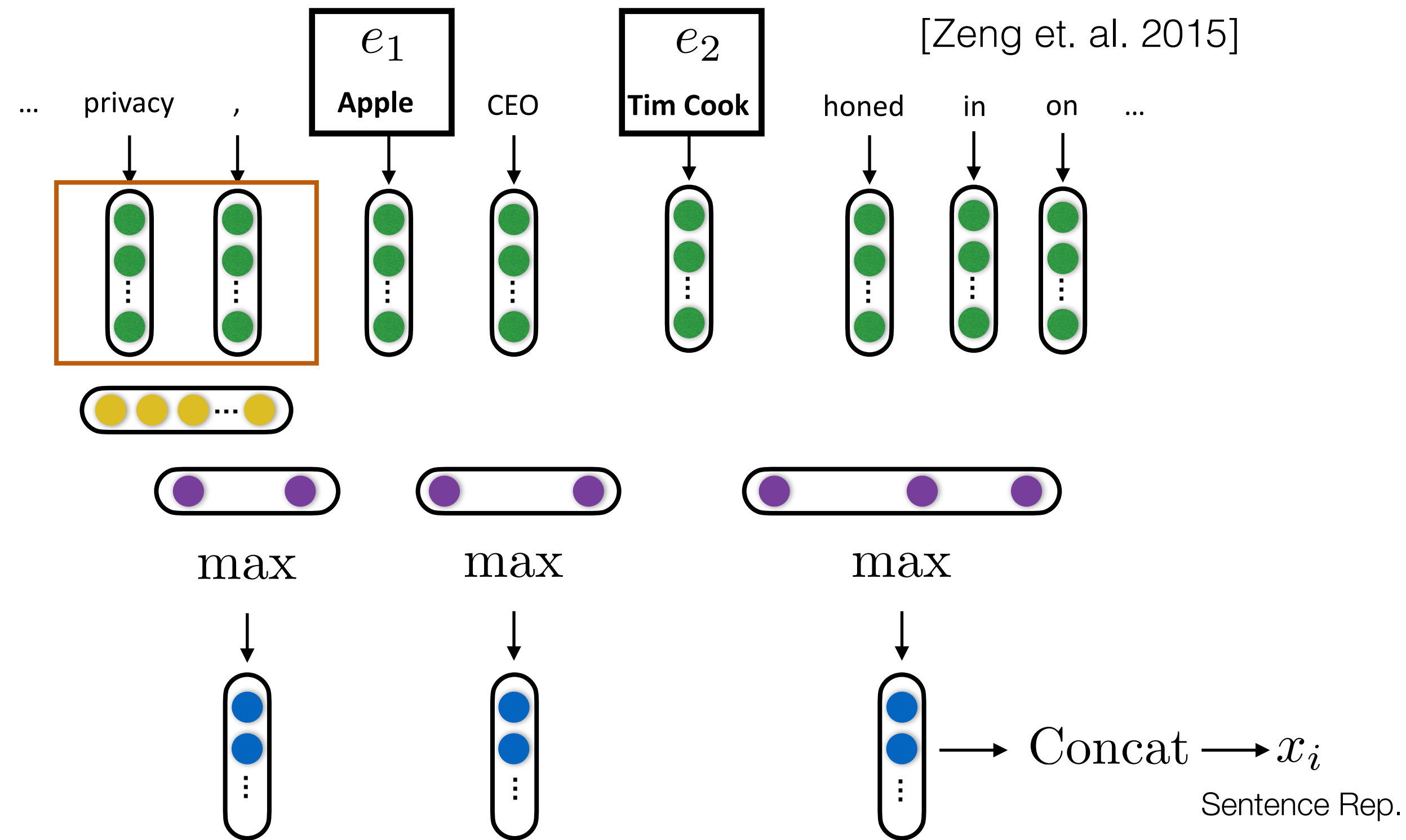


Figure from Bai and Ritter (2019), Model from Zeng et al. (2015)

Open IE

Open Information Extraction

- ▶ “Open”ness — want to be able to extract all kinds of information from open-domain text
- ▶ Acquire commonsense knowledge just from “reading” about it, but need to process lots of text (“machine reading”)
- ▶ Typically no fixed relation inventory

TextRunner

- ▶ Extract positive examples of (e, r, e) triples via parsing and heuristics
- ▶ Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

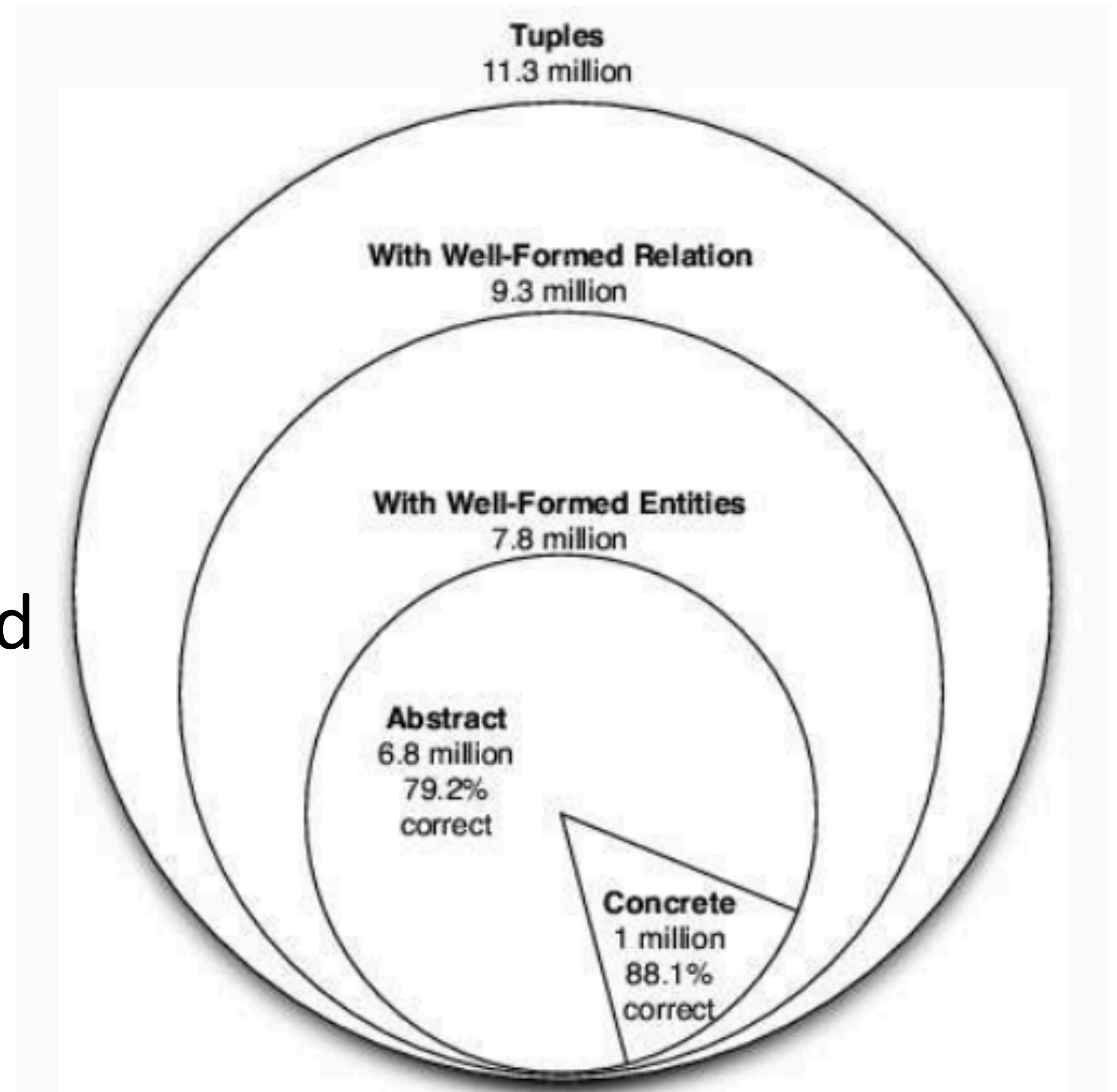
Barack Obama, 44th president of the United States, was born on August 4, 1961 in Honolulu

=> Barack_Obama, was born in, Honolulu

- ▶ 80x faster than running a parser (which was slow in 2007...)
- ▶ Use multiple instances of extractions to assign probability to a relation

Exploiting Redundancy

- ▶ 9M web pages / 133M sentences
- ▶ 2.2 tuples extracted per sentence, filter based on probabilities
- ▶ Concrete: definitely true
Abstract: possibly true but underspecified
- ▶ Hard to evaluate: can assess precision of extracted facts, but how do we know recall?



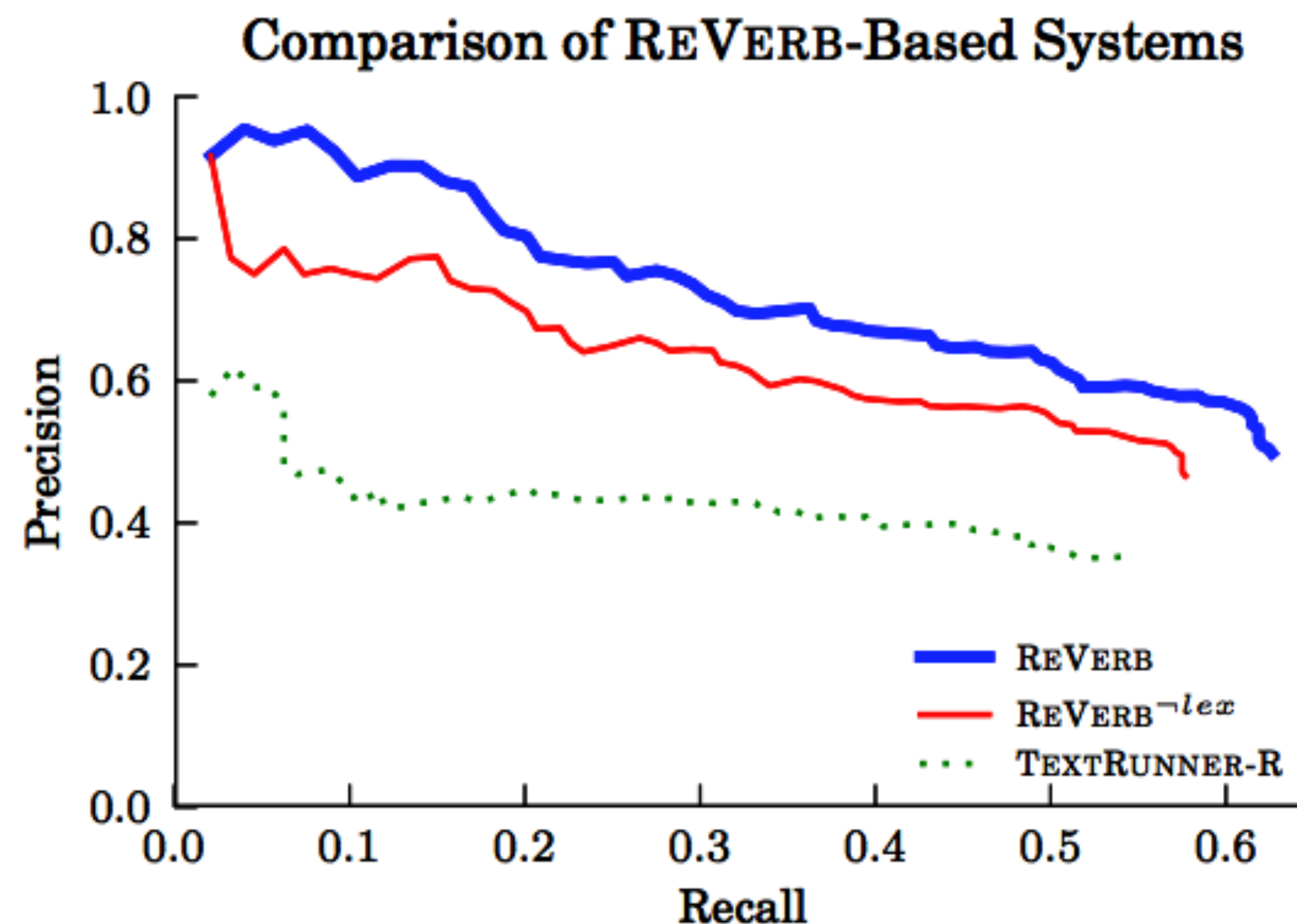
ReVerb

- ▶ More constraints: open relations have to begin with verb, end with preposition, be contiguous (e.g., *was born on*)
- ▶ Extract more meaningful relations, particularly with light verbs

is	is an album by, is the author of, is a city in
has	has a population of, has a Ph.D. in, has a cameo in
made	made a deal with, made a promise to
took	took place in, took control over, took advantage of
gave	gave birth to, gave a talk at, gave new meaning to
got	got tickets to, got a deal on, got funding from

ReVerb

- ▶ For each verb, identify the longest sequence of words following the verb that satisfy a POS regex ($V \cdot^* P$) and which satisfy heuristic lexical constraints on specificity
- ▶ Find the nearest arguments on either side of the relation
- ▶ Annotators labeled relations in 500 documents to assess recall



Takeaways

- ▶ SRL: handle a bunch of phenomena, but more or less like syntax++ in terms of what they represent
- ▶ Relation extraction: can collect data with distant supervision, use this to expand knowledge bases
- ▶ Slot filling: tied to a specific ontology, but gives fine-grained information
- ▶ Open IE: extracts lots of things, but hard to know how good or useful they are
 - ▶ Can combine with standard question answering
 - ▶ Add new facts to knowledge bases
- ▶ Many, many applications and techniques