LISA

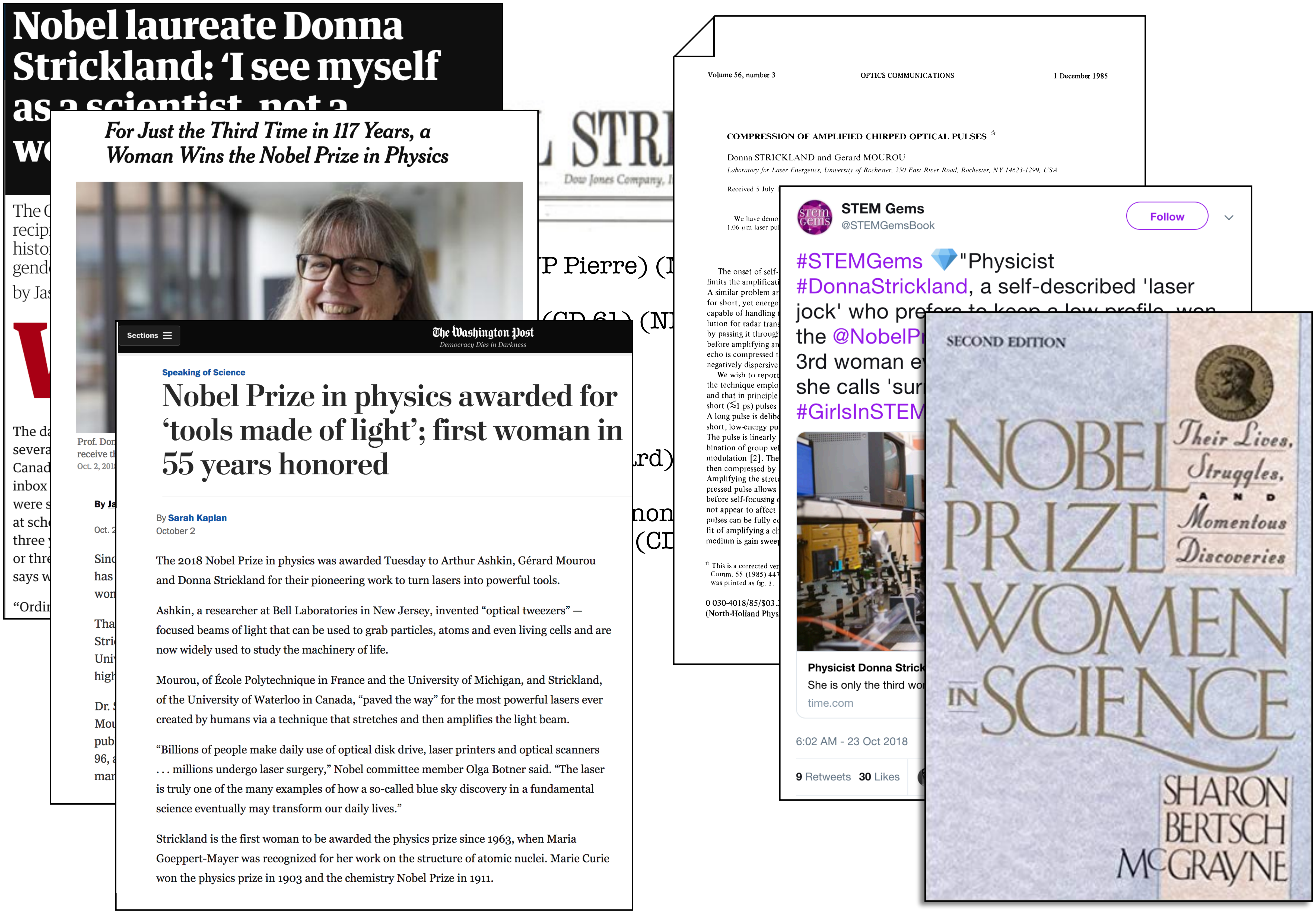
Linguistically-Informed Self-Attention for Semantic Role Labeling

Emma Strubell[1] Patrick Verga[1] Daniel Andor[2] David Weiss[2] Andrew McCallum[1]

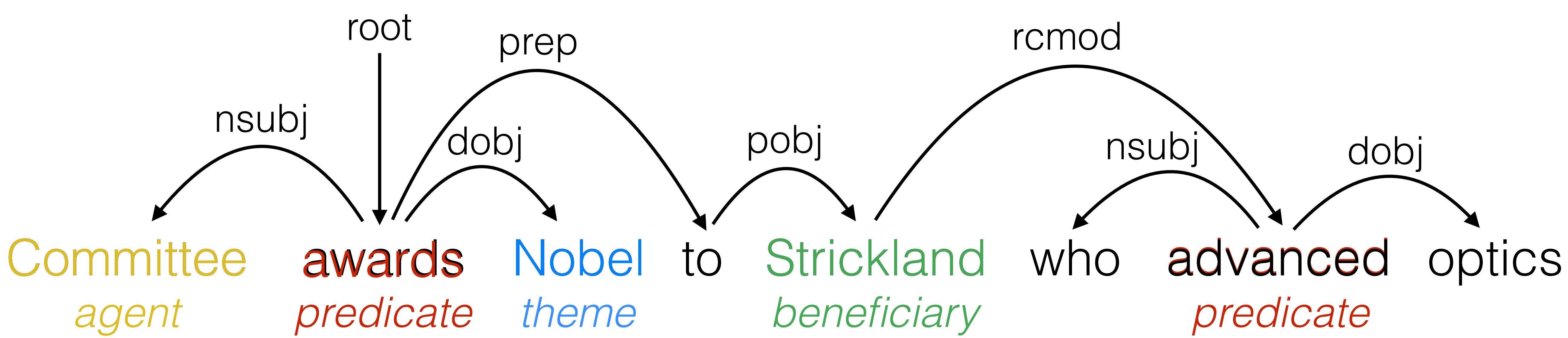
[1] UMass Amherst, College of Information and Computer Sciences

[2] Google AI

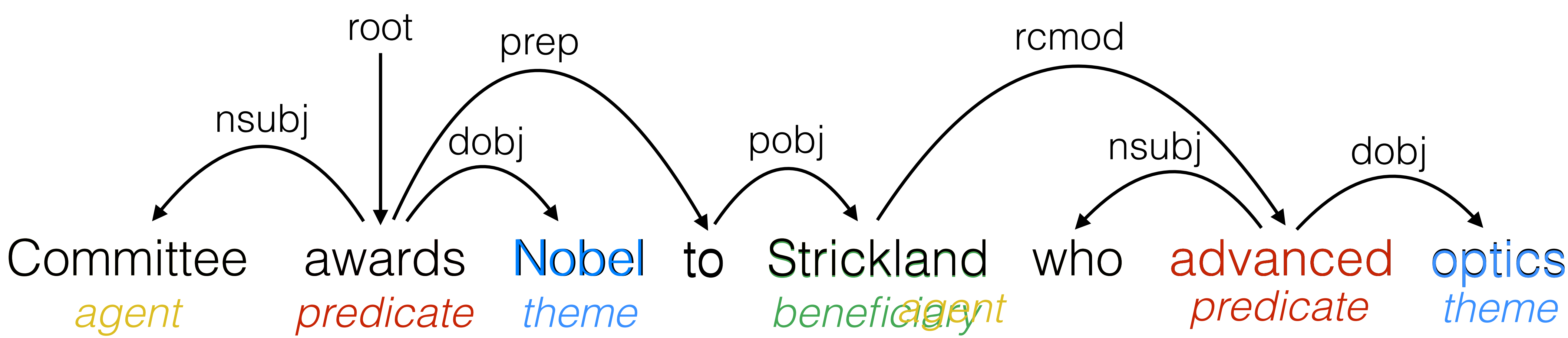
Want fast, accurate, robust NLP



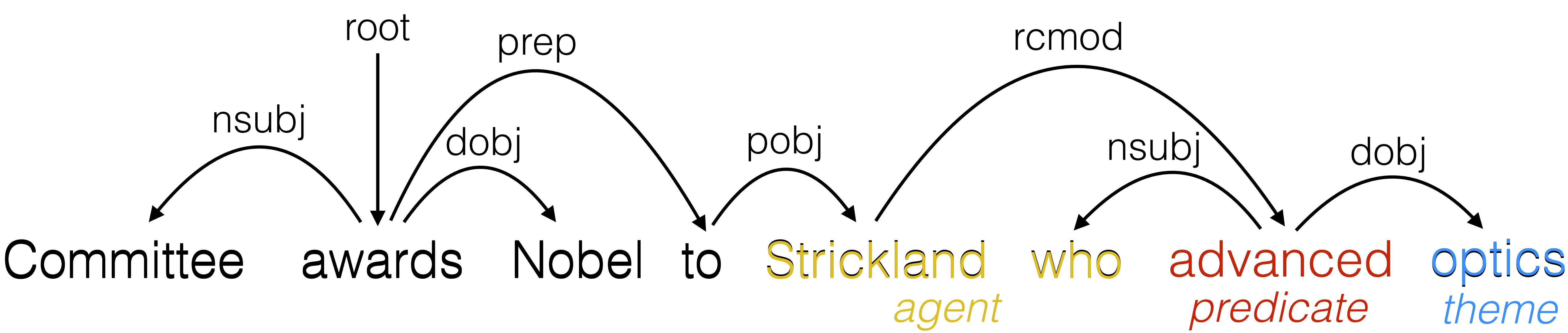
SRL: Who did what to whom?



SRL: Who did what to whom?

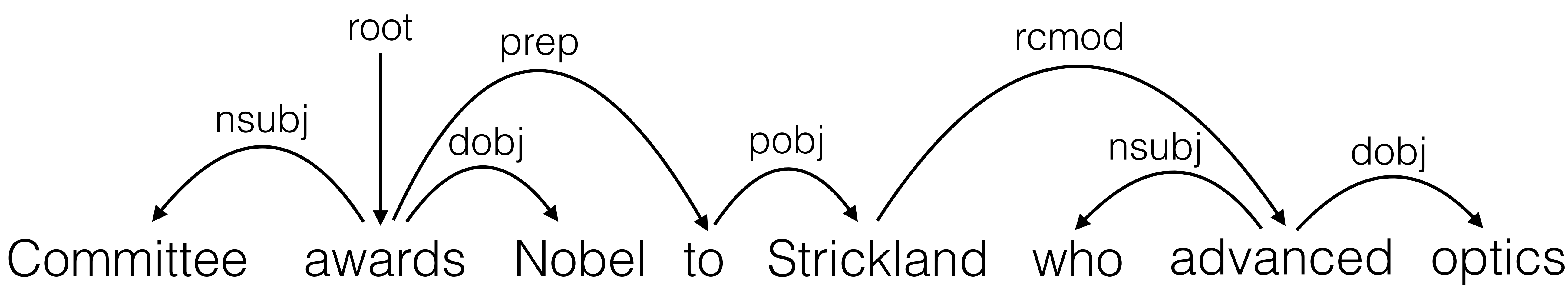


SRL: Who did what to whom?



Committee awards Nobel to Strickland who advanced optics
agent *predicate* *theme* *beneficiary*

PropBank SRL: Who did what to whom?



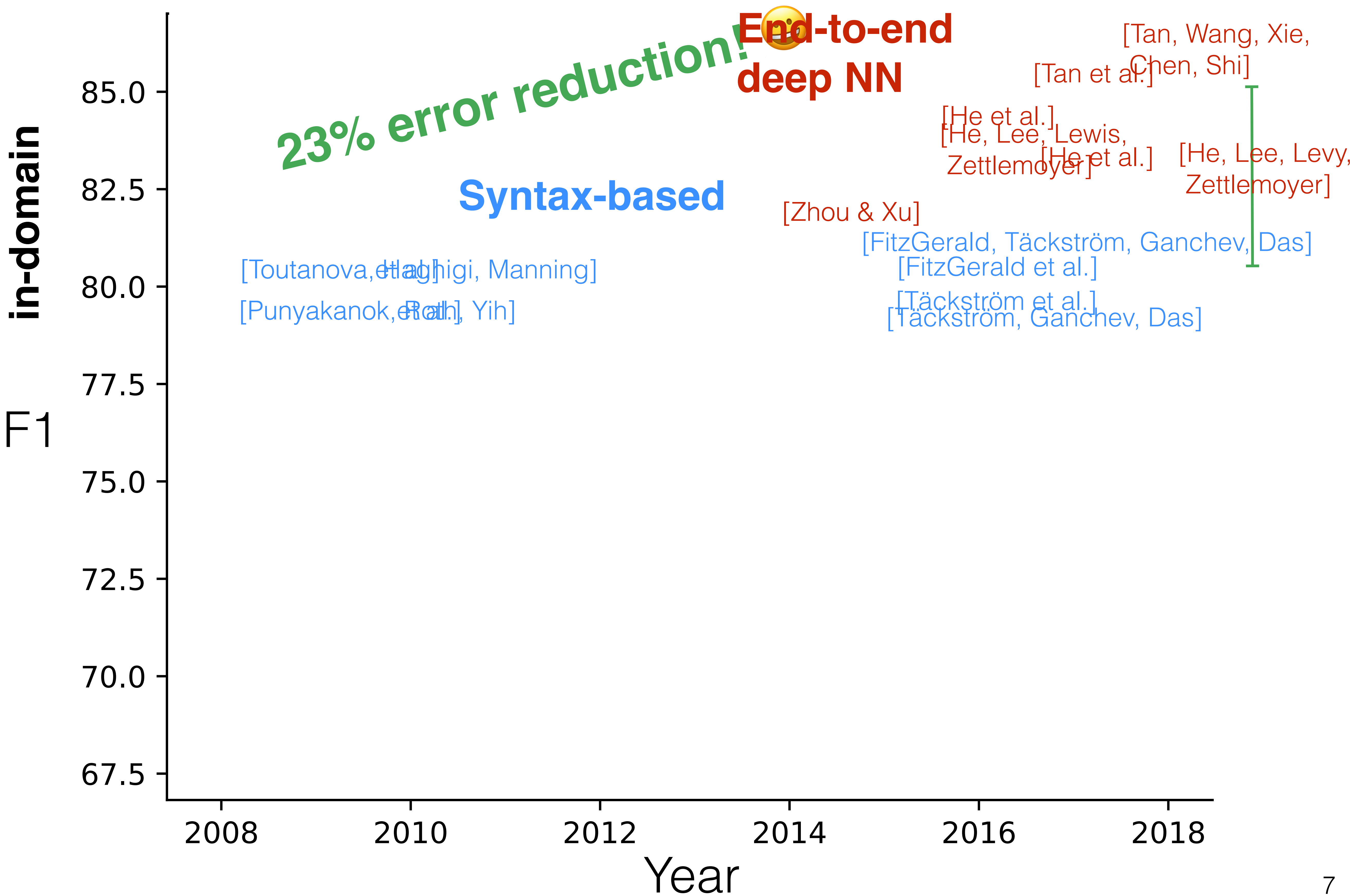
Committee awards Nobel to Strickland who advanced optics

Strickland: ARG_0 / *agent*; who: R-ARG_0; advanced: V / *predicate*; optics: ARG_1 / *theme*

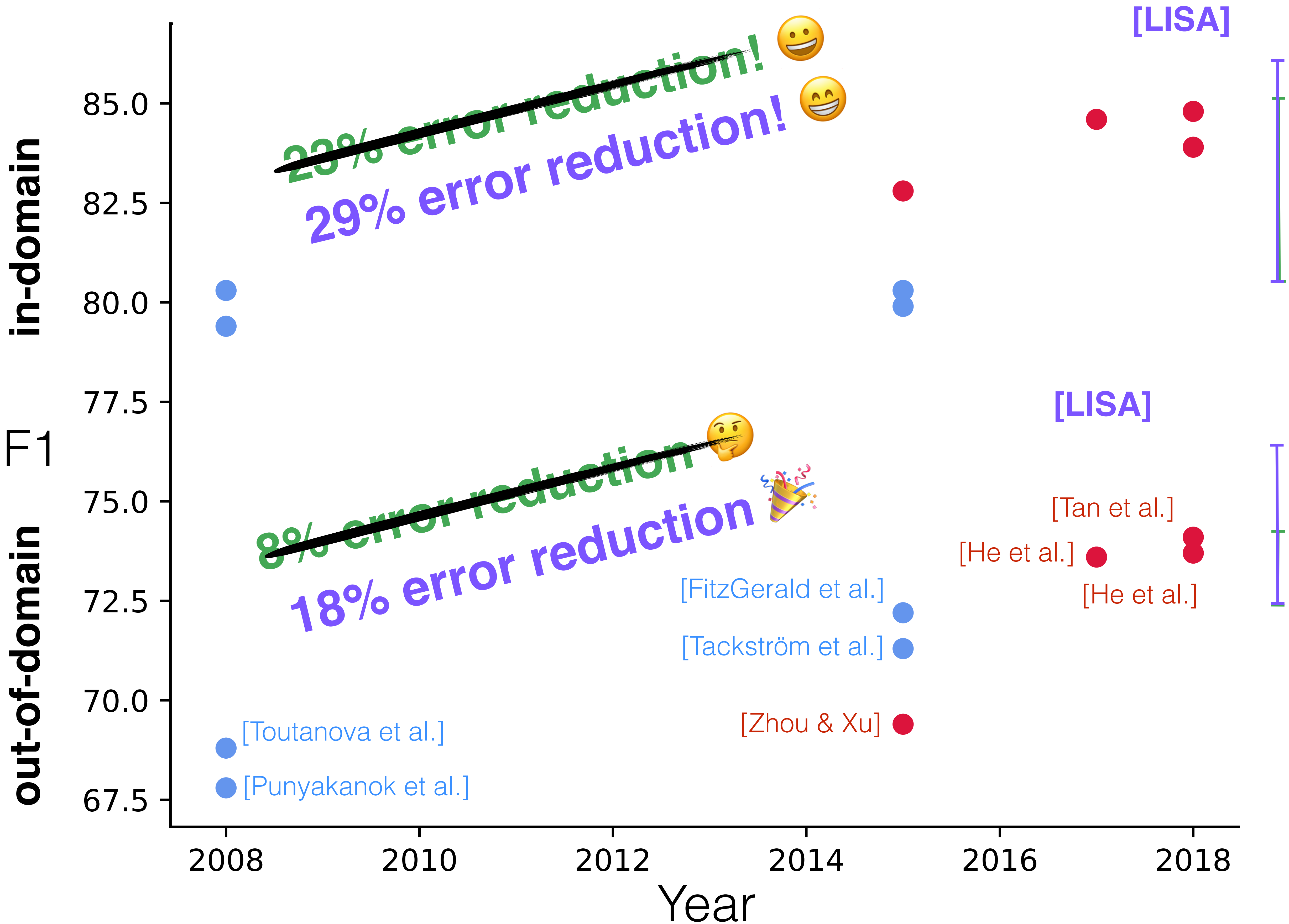
Committee awards Nobel to Strickland who advanced optics

Committee: ARG_0 / *agent*; awards: V / *predicate*; Nobel: ARG_1 / *theme*; Strickland: ARG_2 / *beneficiary*

10 years of PropBank SRL

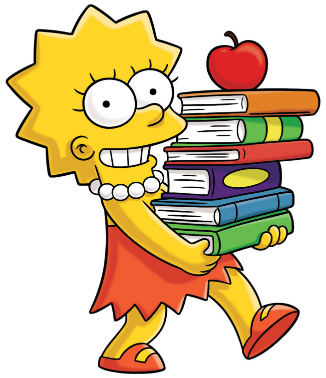


10 years of PropBank SRL



Linguistically-Informed Self-Attention

- **Multi-task learning**
  - Part-of-speech tagging
  - Labeled dependency parsing
  - Predicate detection
  - Semantic role spans & labeling
- **Syntactically-informed self-attention**
  - Multi-head self-attention supervised by syntax

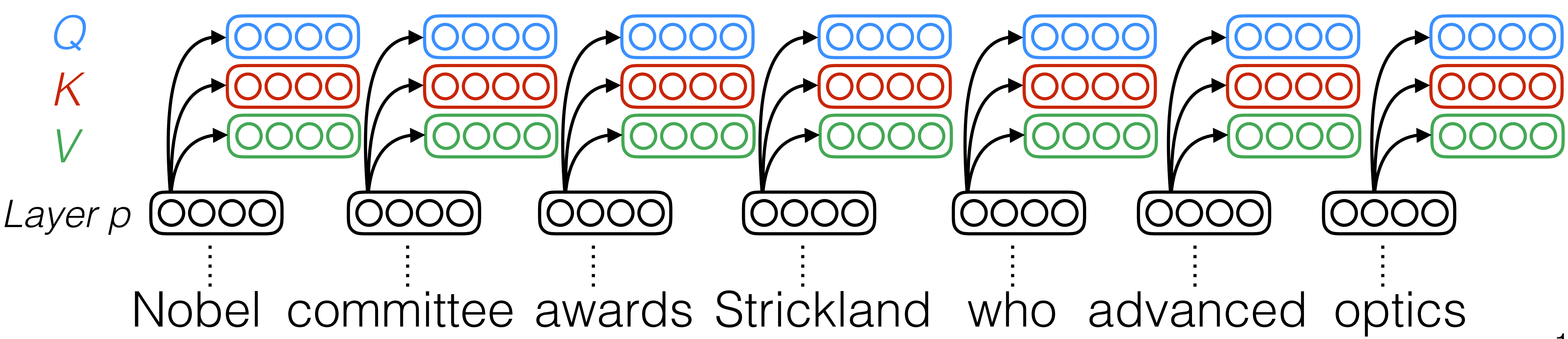


Outline

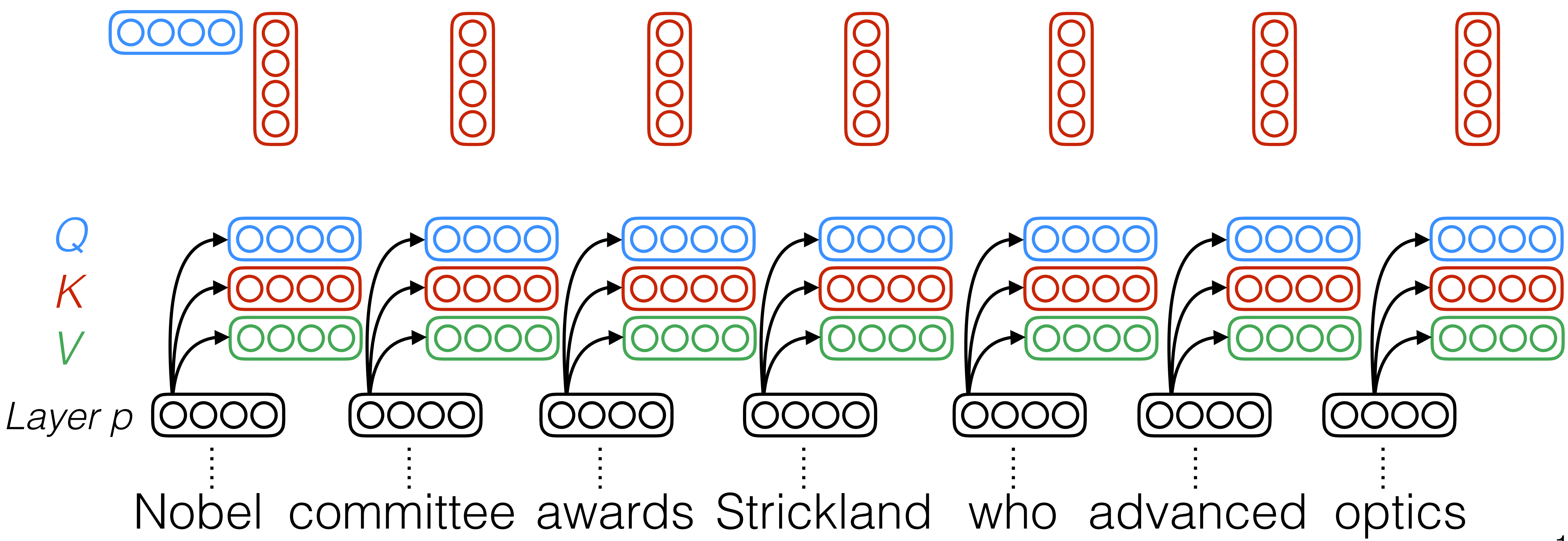
- Want fast, accurate, robust NLU
- PropBank SRL: Who did what to whom?
- 10 years of PropBank SRL
- LISA: Linguistically-informed self attention
    - Multi-head self-attention [Vaswani et al. 2017]
    - Syntactically-informed self-attention
    - Multi-task learning, single-pass inference
    - Training
- Experimental results & error analysis

[Vaswani et al. 2017]

Self-attention

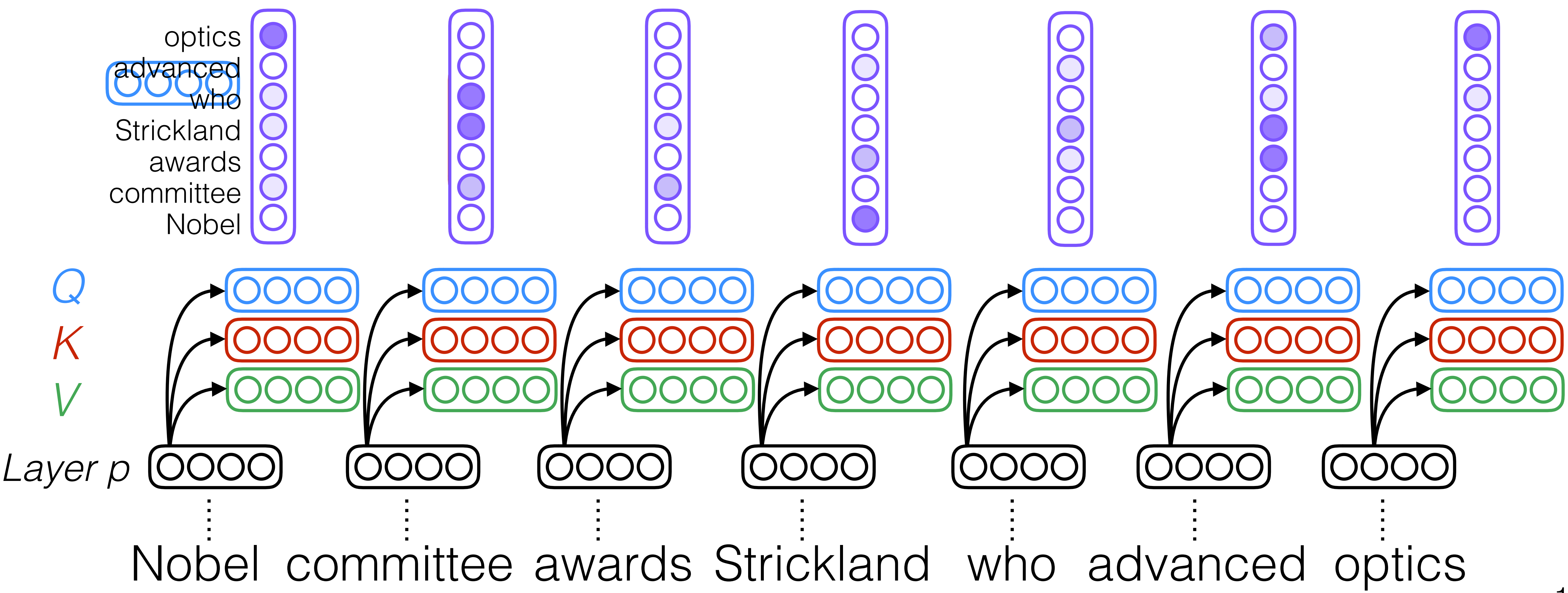


Self-attention



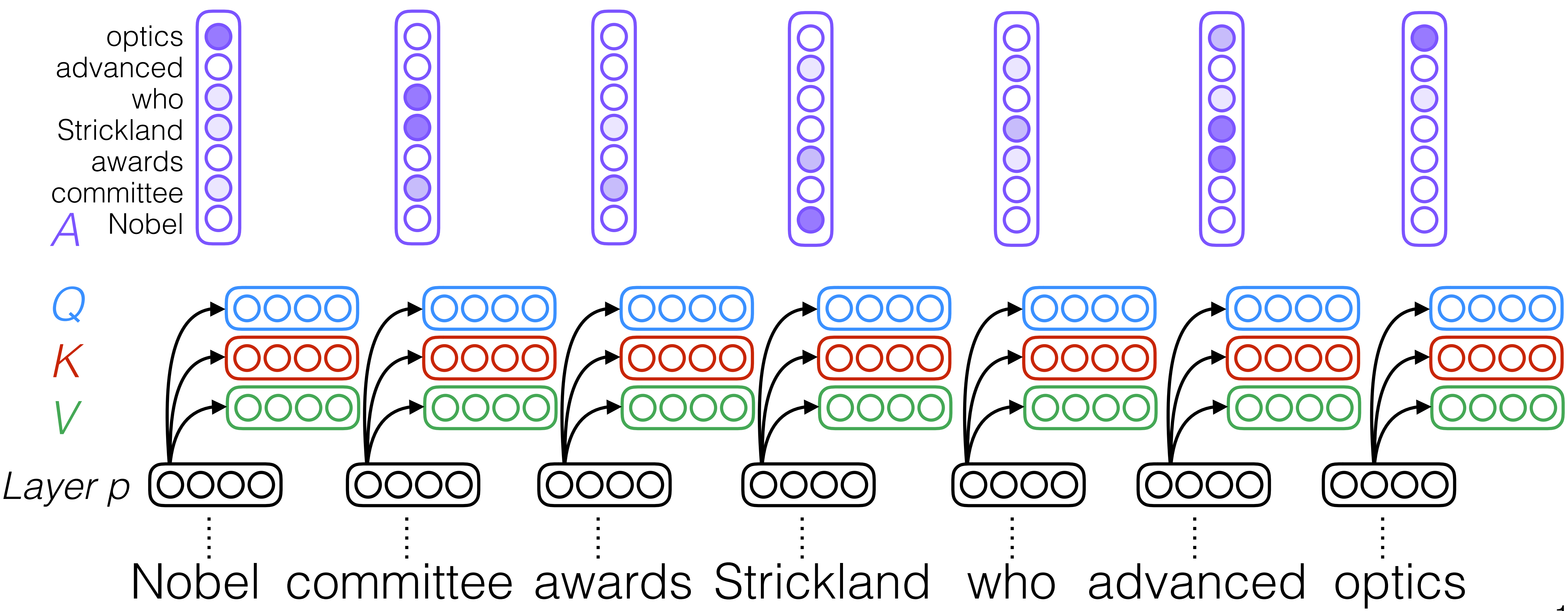
Self-attention

[Vaswani et al. 2017]



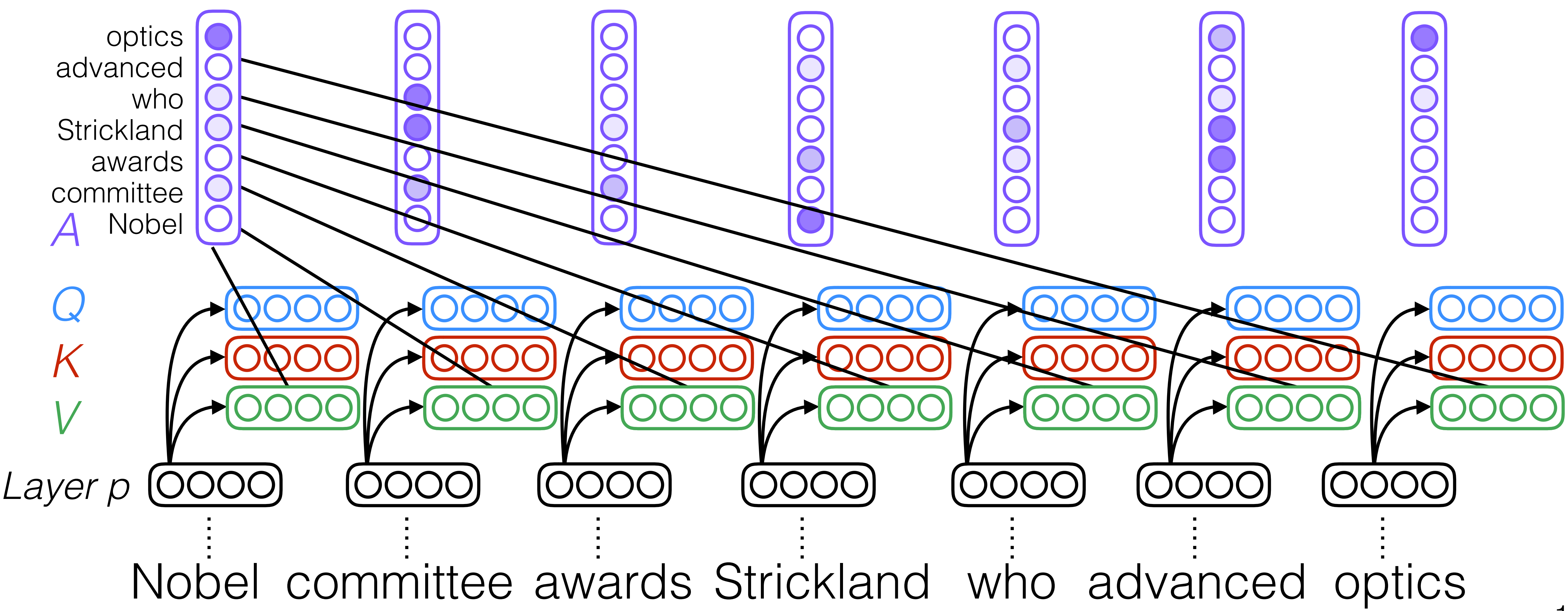
Self-attention

[Vaswani et al. 2017]



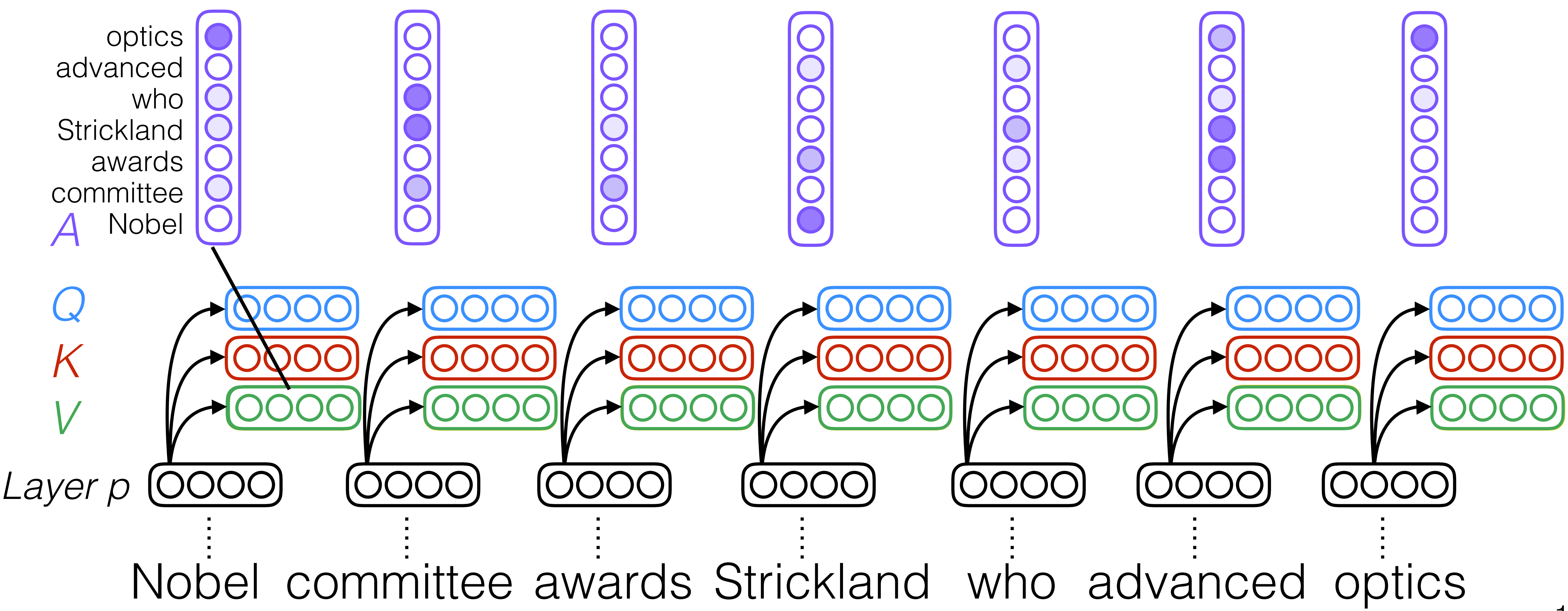
[Vaswani et al. 2017]

Self-attention



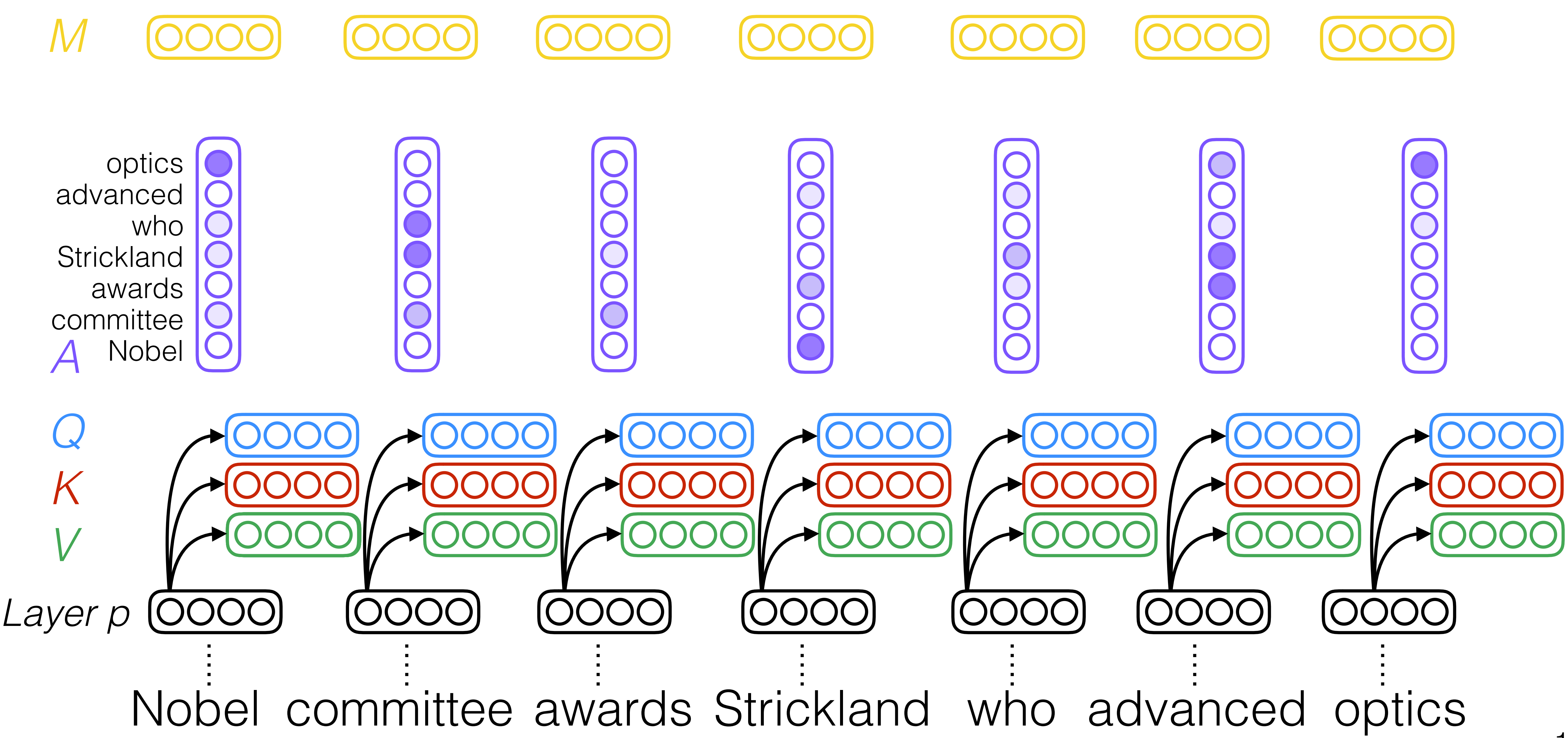
Self-attention

[Vaswani et al. 2017]



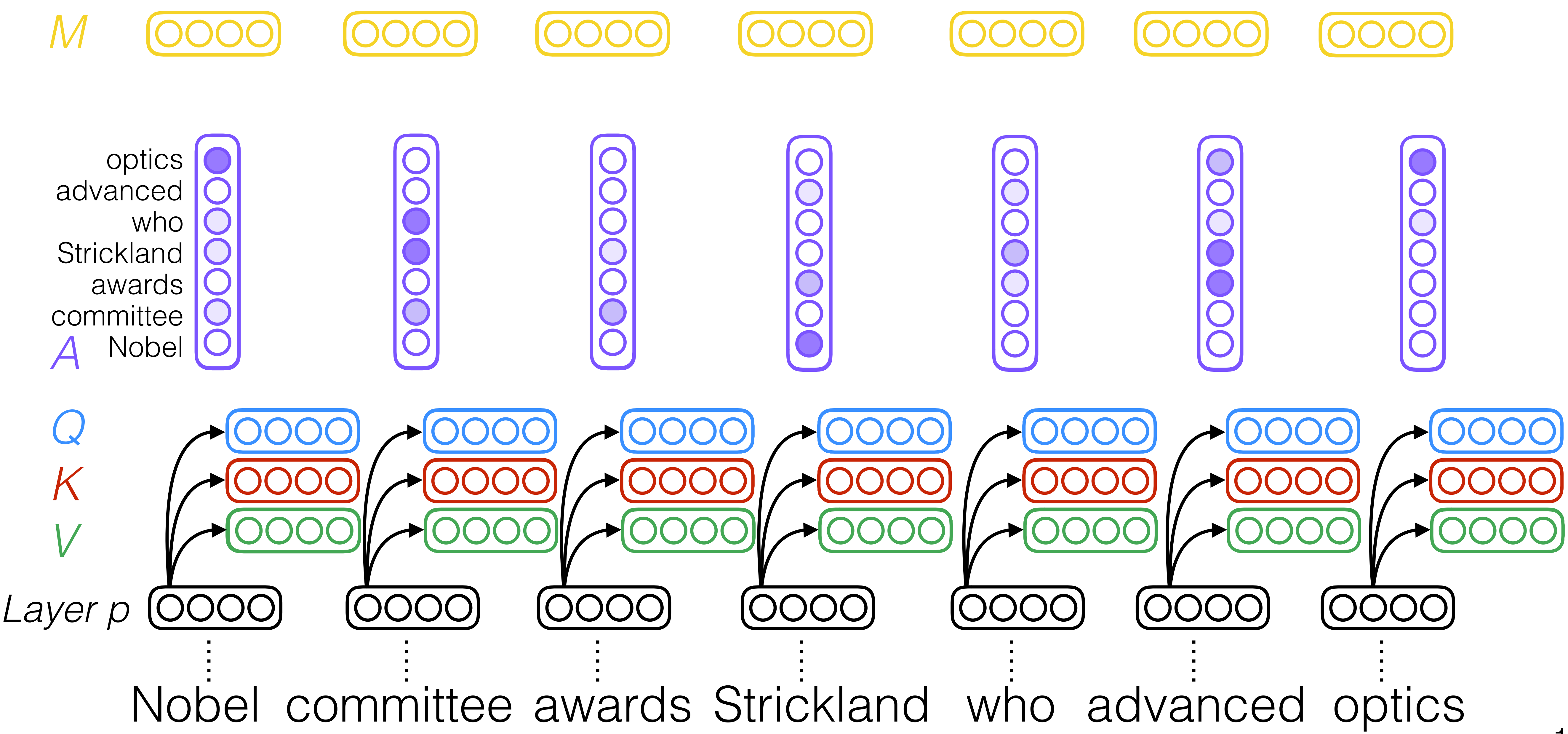
[Vaswani et al. 2017]

Self-attention

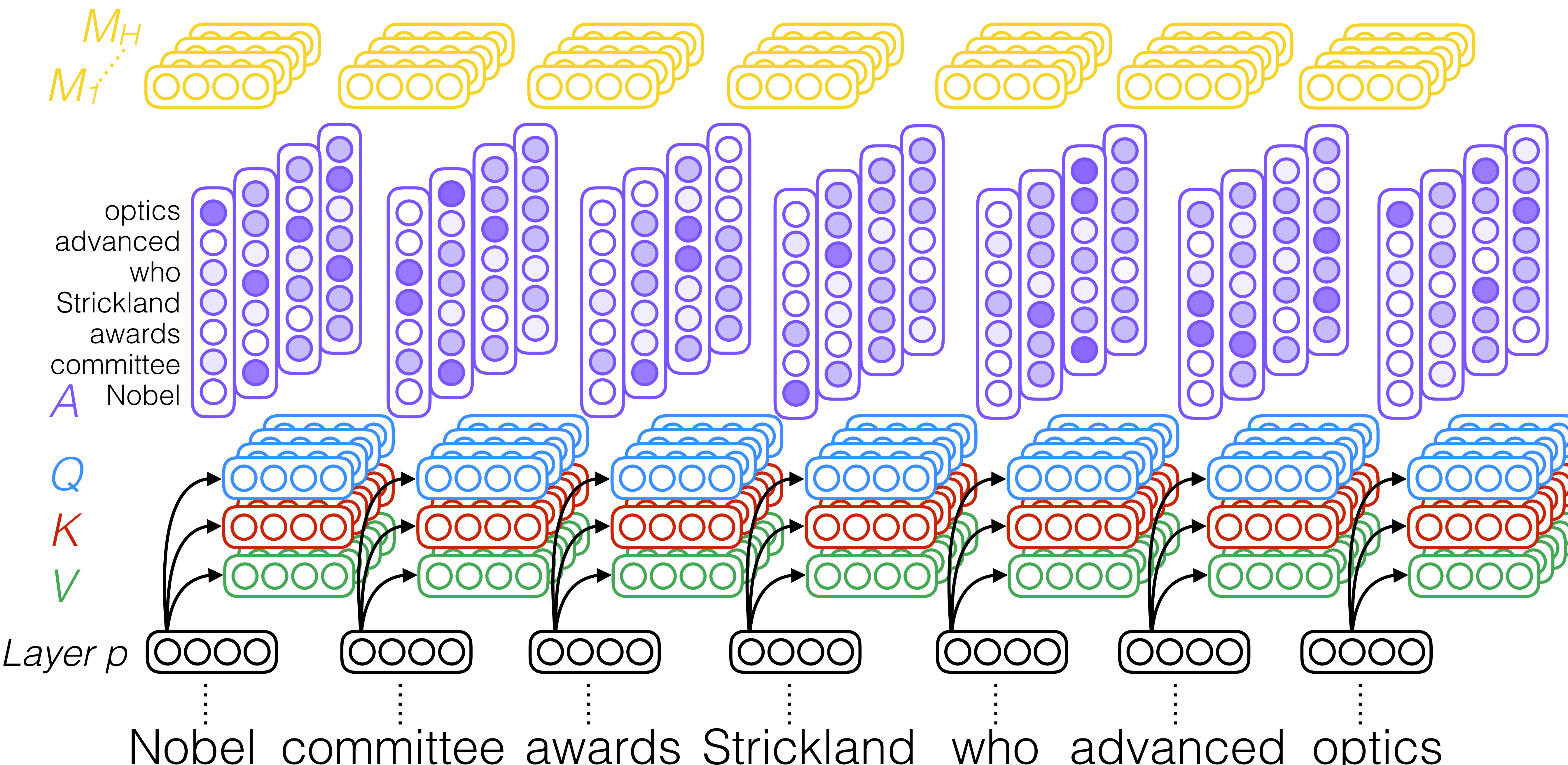


[Vaswani et al. 2017]

Self-attention

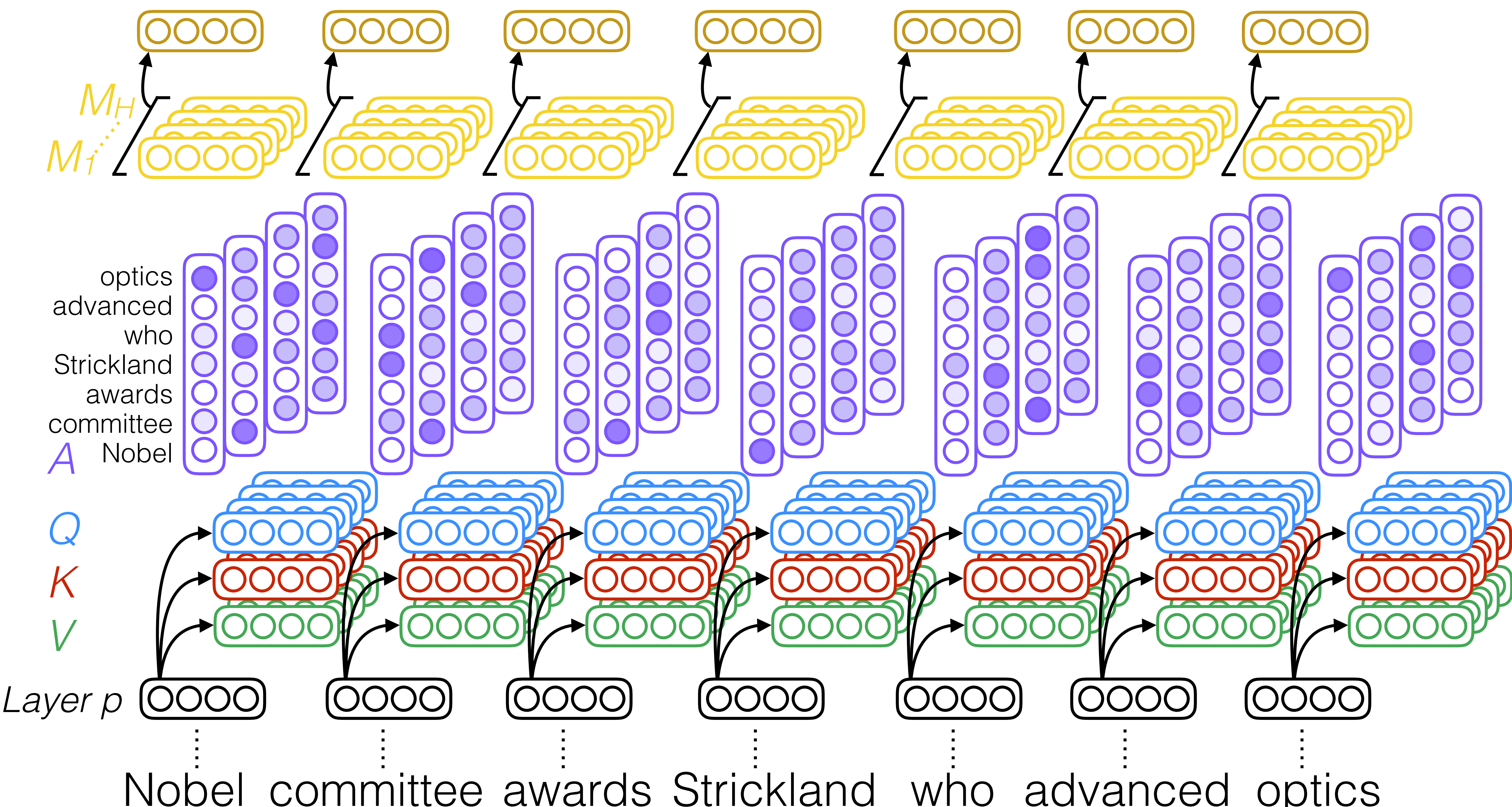


Multi-head self-attention



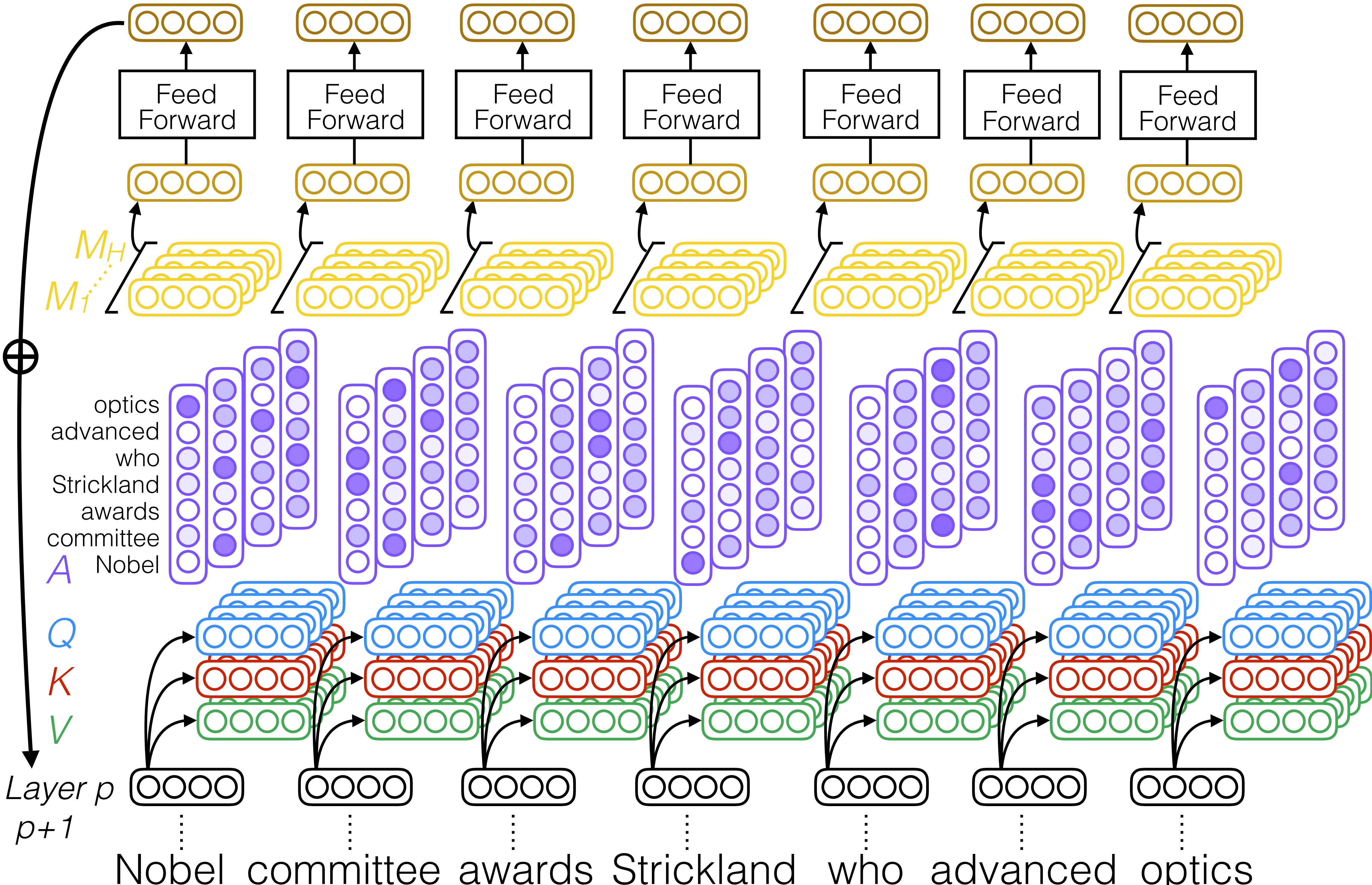
[Vaswani et al. 2017]

Multi-head self-attention

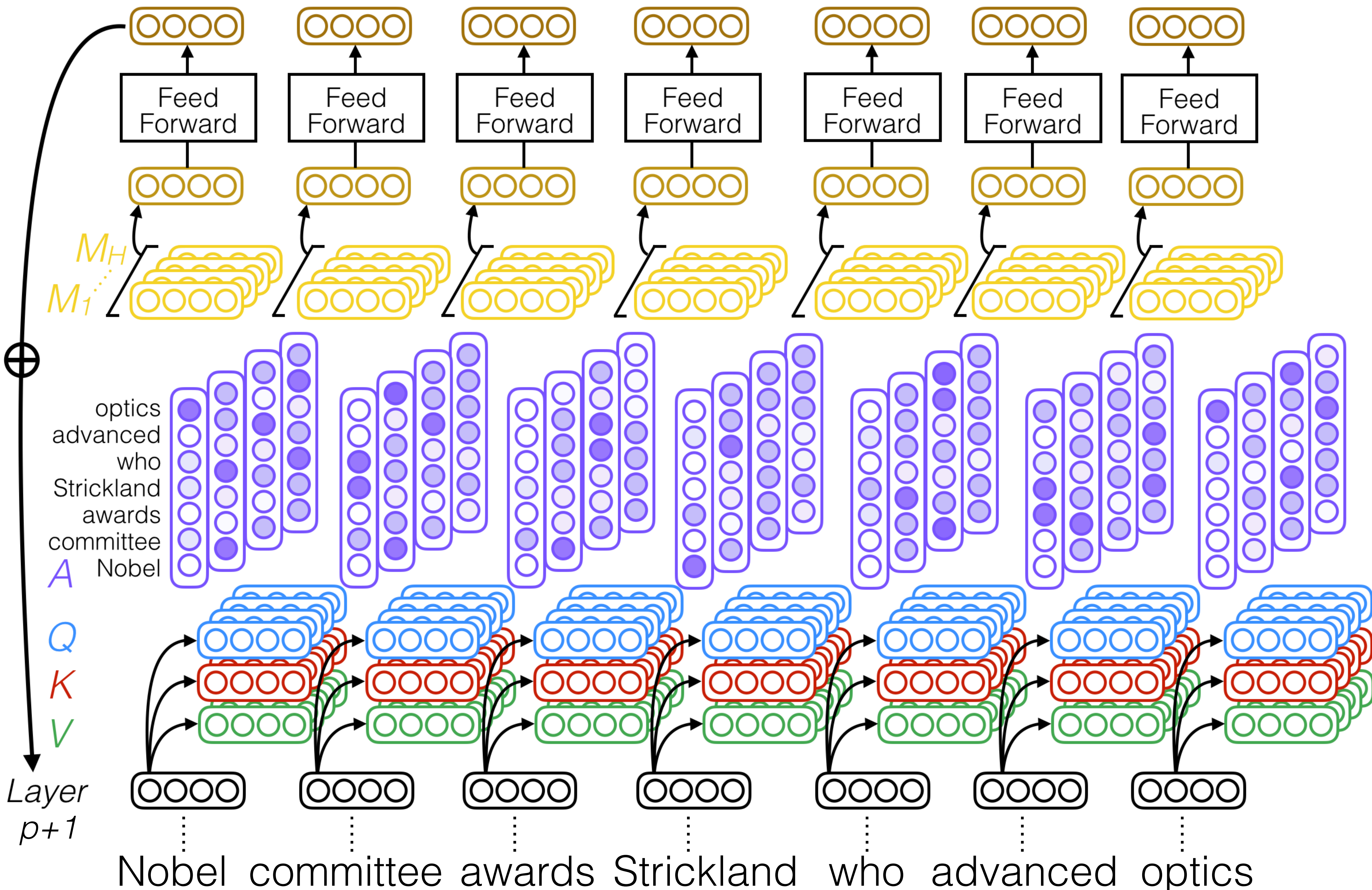


[Vaswani et al. 2017]

Multi-head self-attention



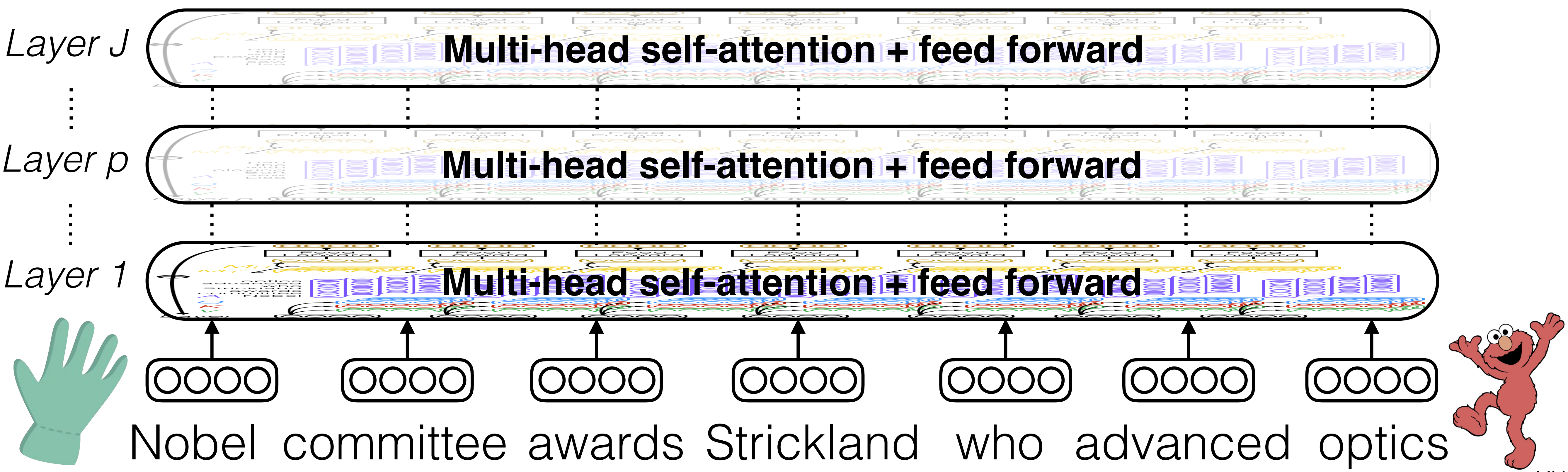
Multi-head self-attention



[Vaswani et al. 2017]

Multi-head self-attention

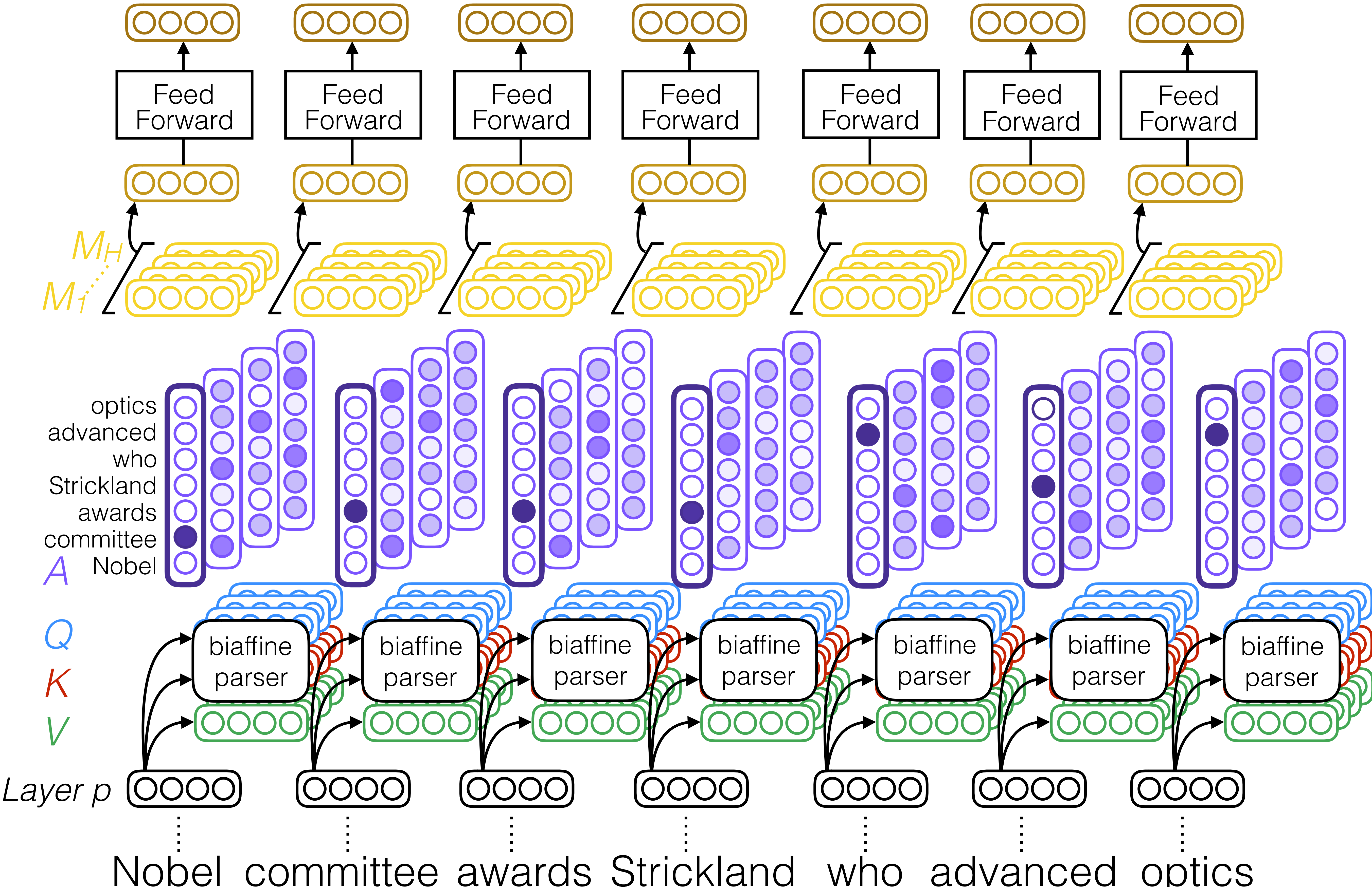
$$s_t^{(j)} = LN(s_t^{(j-1)} + T^{(j)}(s_t^{(j-1)}))$$



Outline

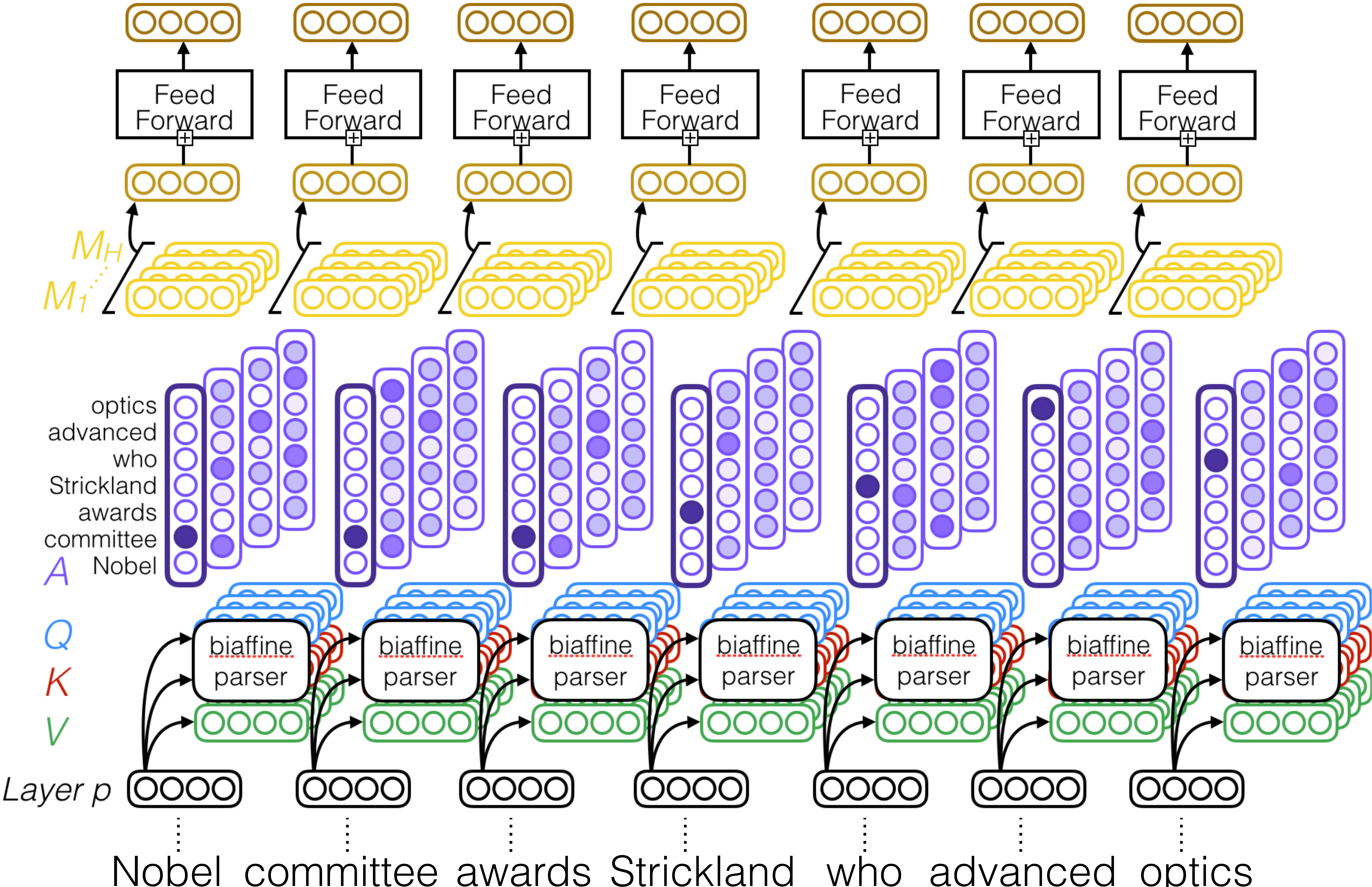
- Want fast, accurate, robust NLU
- PropBank SRL: Who did what to whom?
- 10 years of PropBank SRL
- LISA: Linguistically-informed self attention
    - Multi-head self-attention [Vaswani et al. 2017]
    - Syntactically-informed self-attention
    - Multi-task learning, single-pass inference
    - Training
- Experimental results & error analysis

Syntactically-informed self-attention



[Dozat and Manning 2017]

Syntactically-informed self-attention



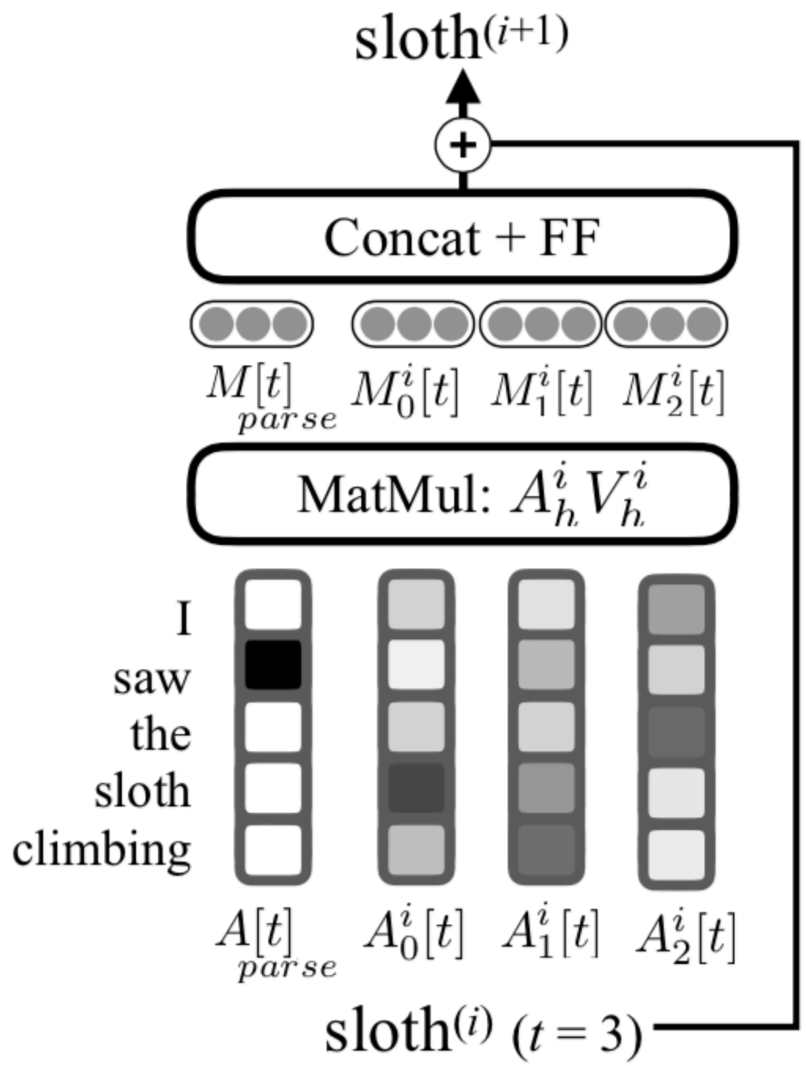


Figure 2: Syntactically-informed self-attention fo the query word *sloth*. Attention weights A_{parse} heavily weight the token's syntactic governor, *saw*, in a weighted average over the token values V_{parse}. The other attention heads act as usual, and the attended representations from all heads are concatenated and projected through a feed-forward layer to produce the syntactically-informed representation for *sloth*.

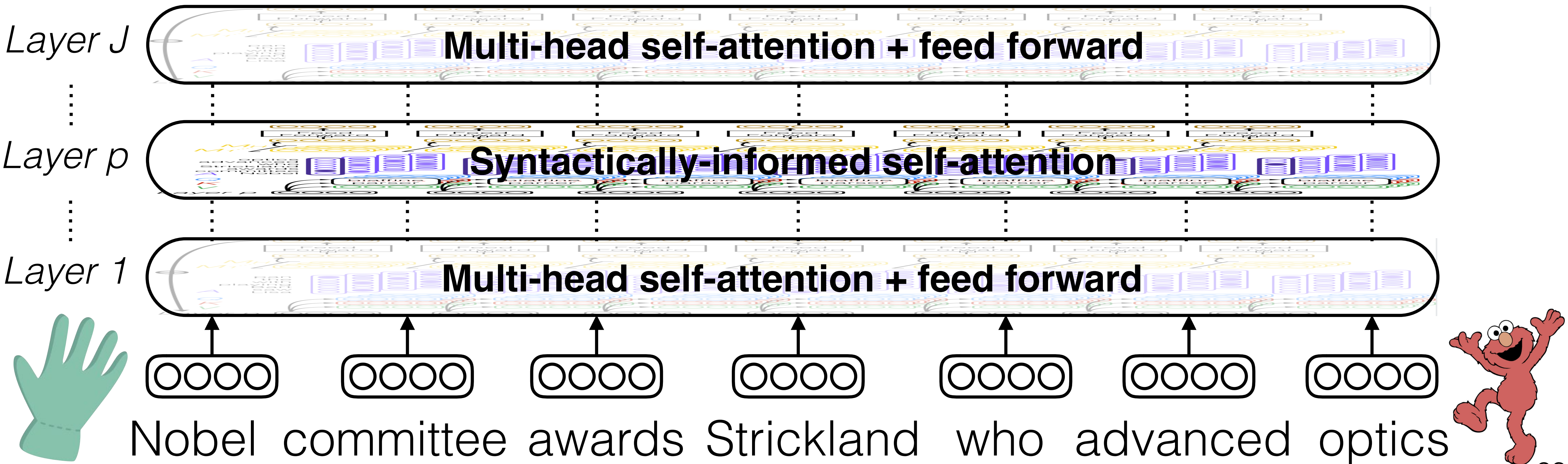
$$A_h^{(j)} = \mathrm{softmax}(d_k^{-0.5} Q_h^{(j)} {K_h^{(j)}}^T)$$

$$M_h^{(j)} = A_h^{(j)} V_h^{(j)}$$

$$A_{parse} = \mathrm{softmax}(Q_{parse} U_{heads} K_{parse}^T)$$

$$P(q = \mathrm{head}(t) \mid \mathcal{X}) = A_{parse}[t, q]$$

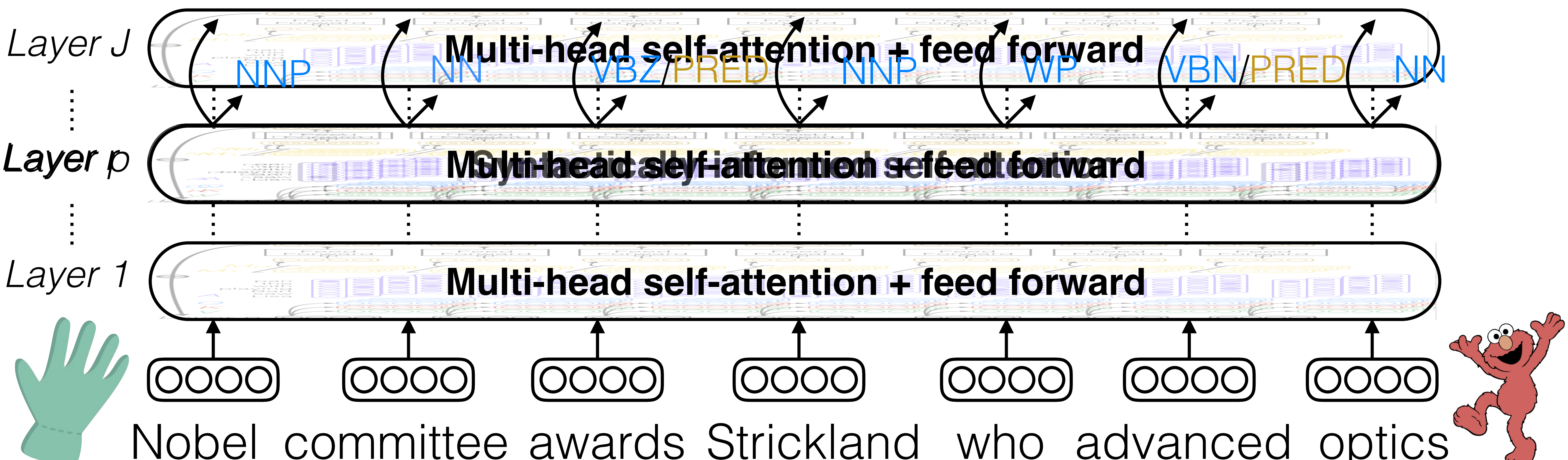
Syntactically-informed self-attention



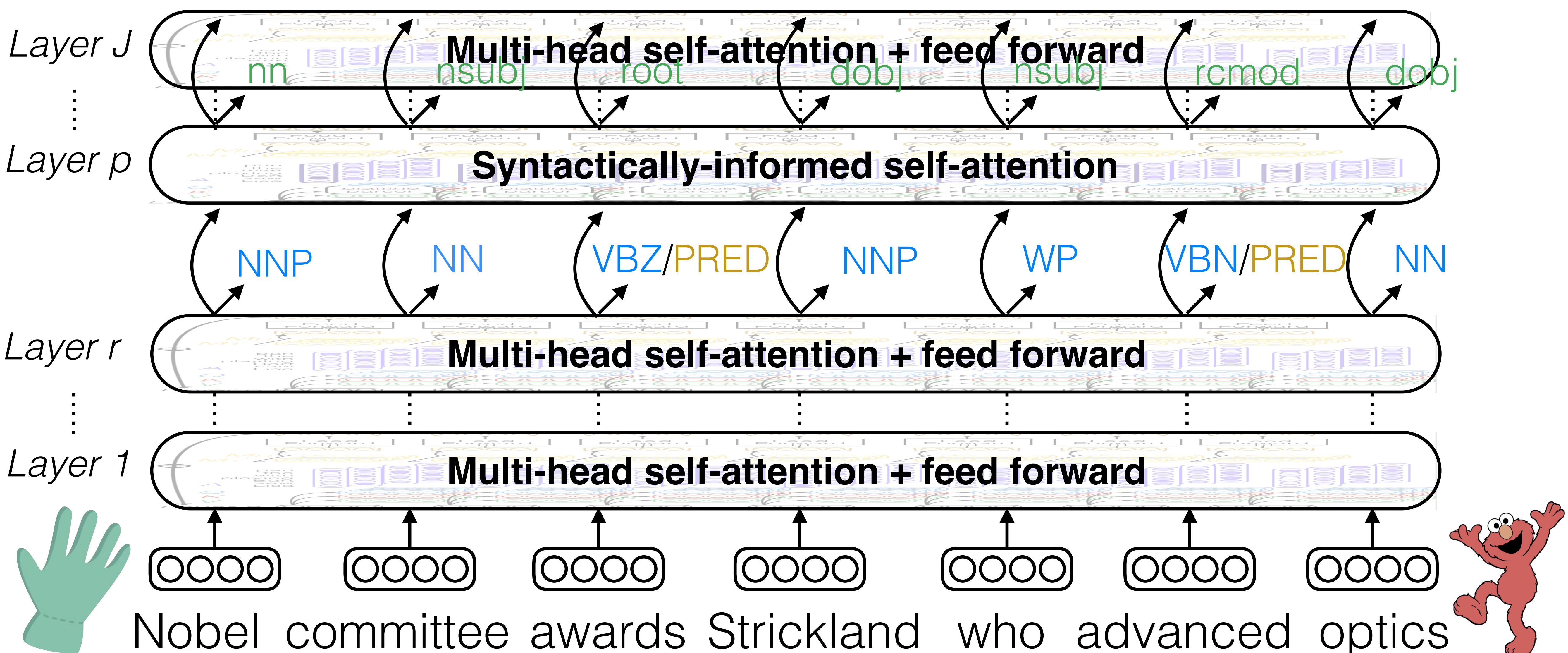
Outline

- Want fast, accurate, robust NLU
- PropBank SRL: Who did what to whom?
- 10 years of PropBank SRL
- LISA: Linguistically-informed self attention
    - Multi-head self-attention
    - Syntactically-informed self-attention
    - Multi-task learning, single-pass inference
    - Training
- Experimental results & error analysis

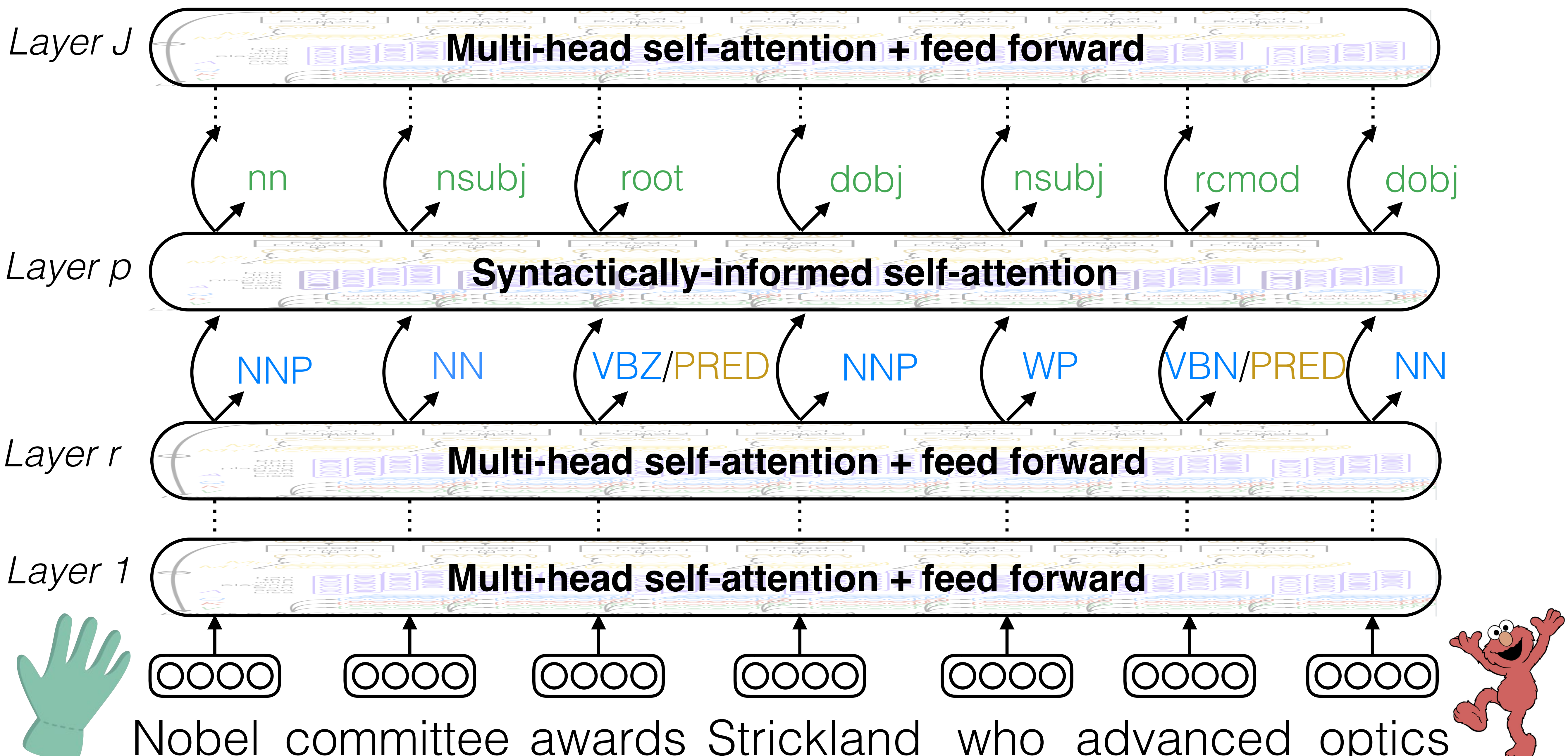
LISA: Linguistically-Informed Self-Attention



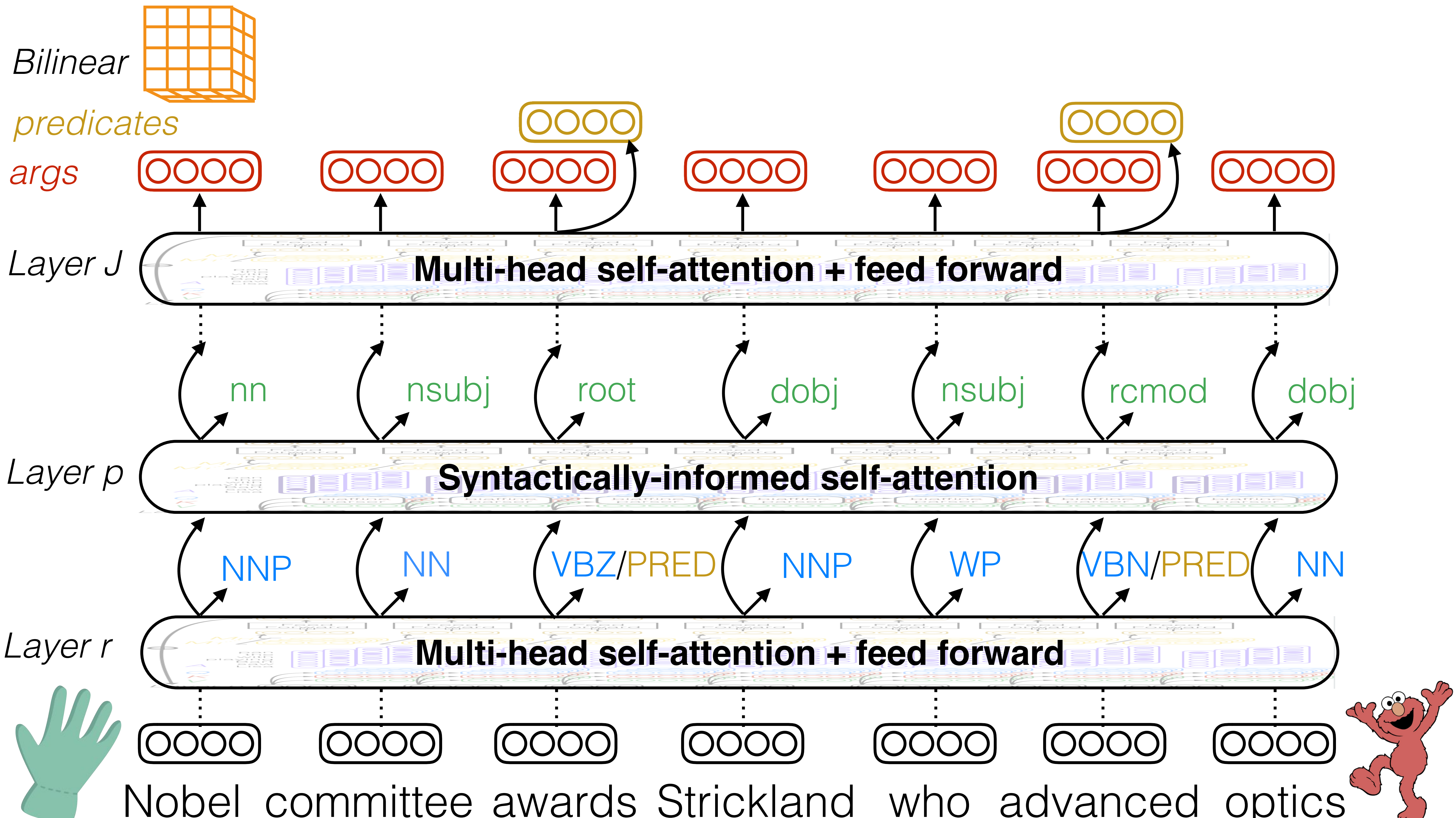
LISA: Linguistically-Informed Self-Attention



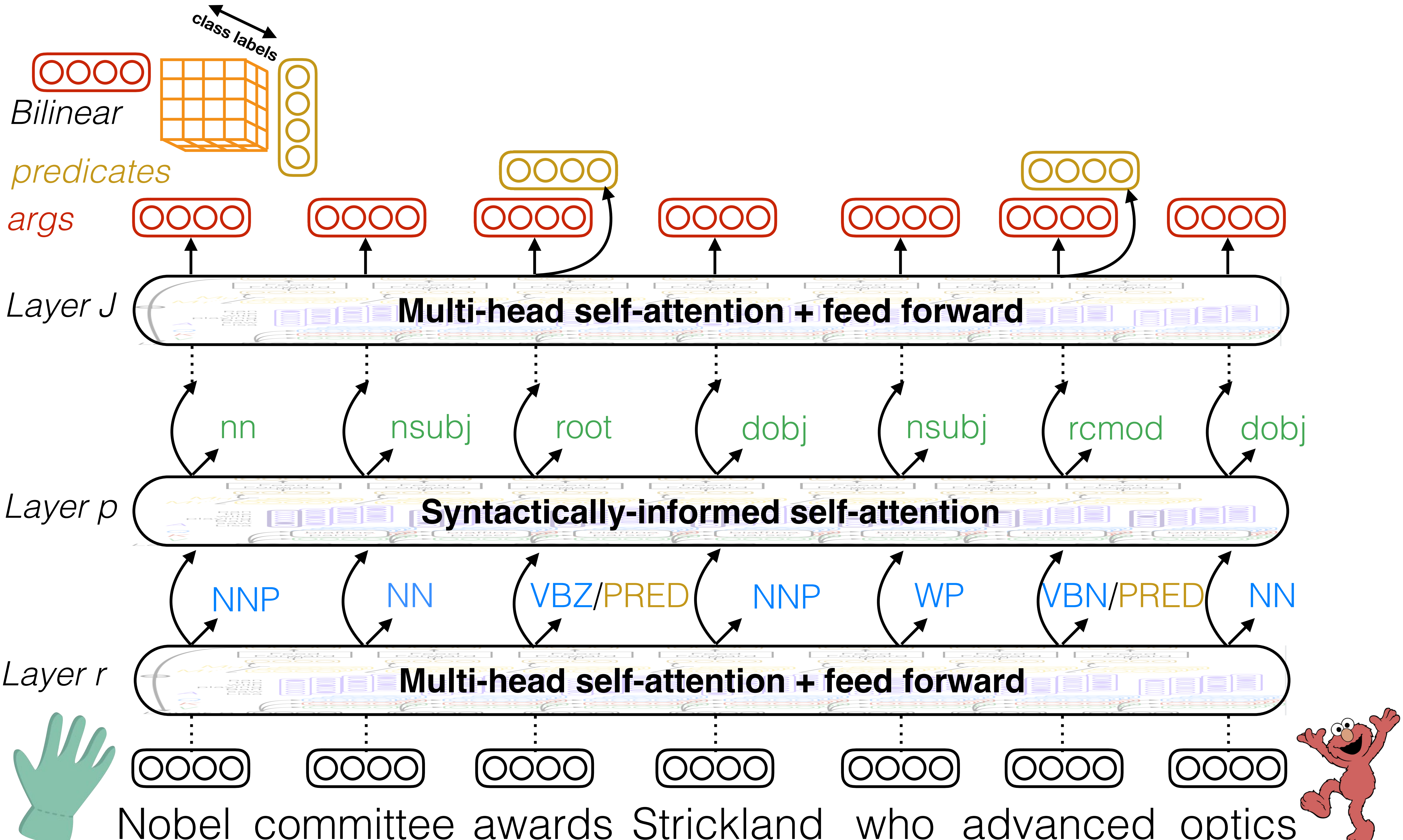
LISA: Linguistically-Informed Self-Attention



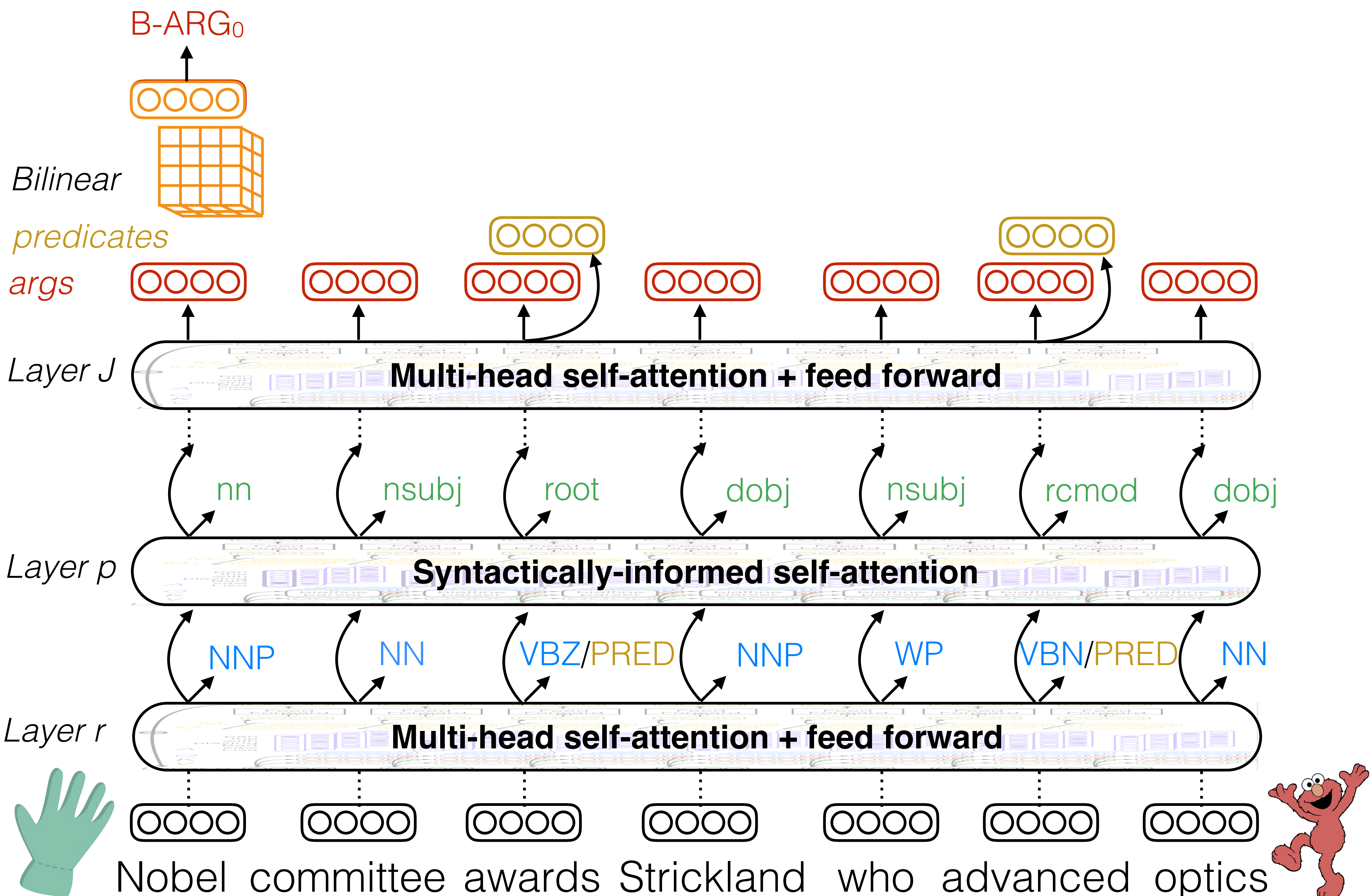
LISA: Linguistically-Informed Self-Attention



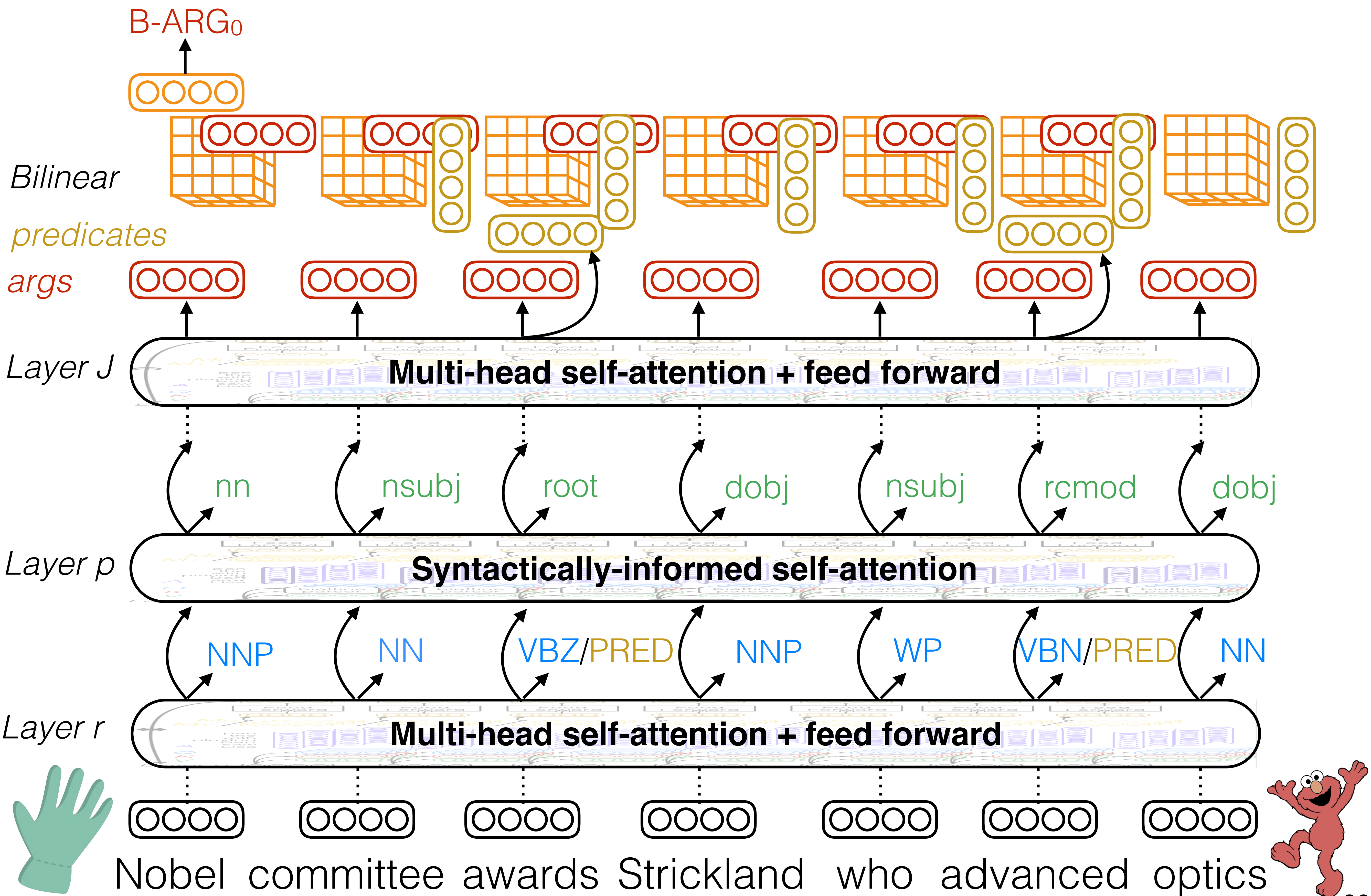
LISA: Linguistically-Informed Self-Attention



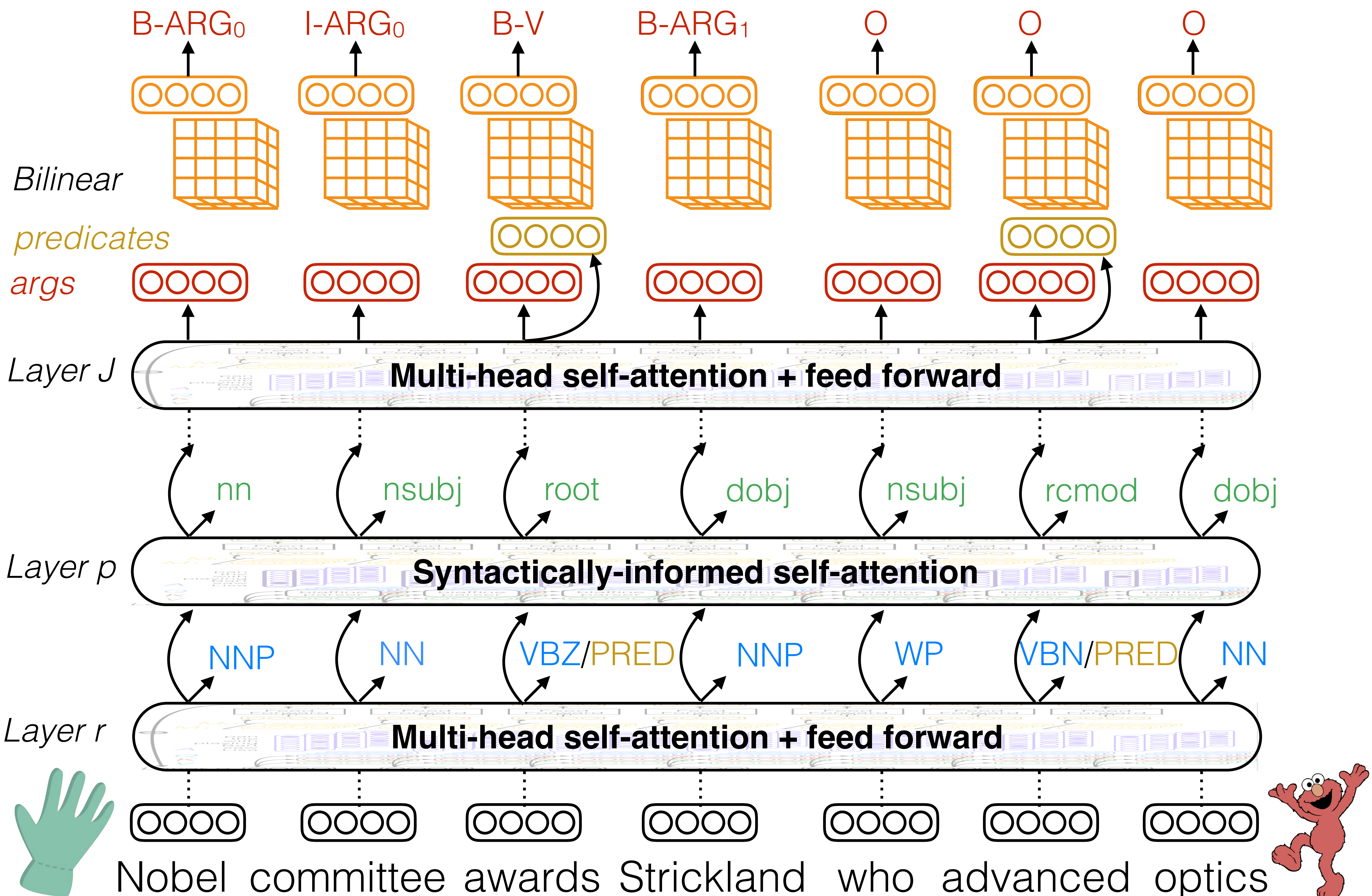
LISA: Linguistically-Informed Self-Attention



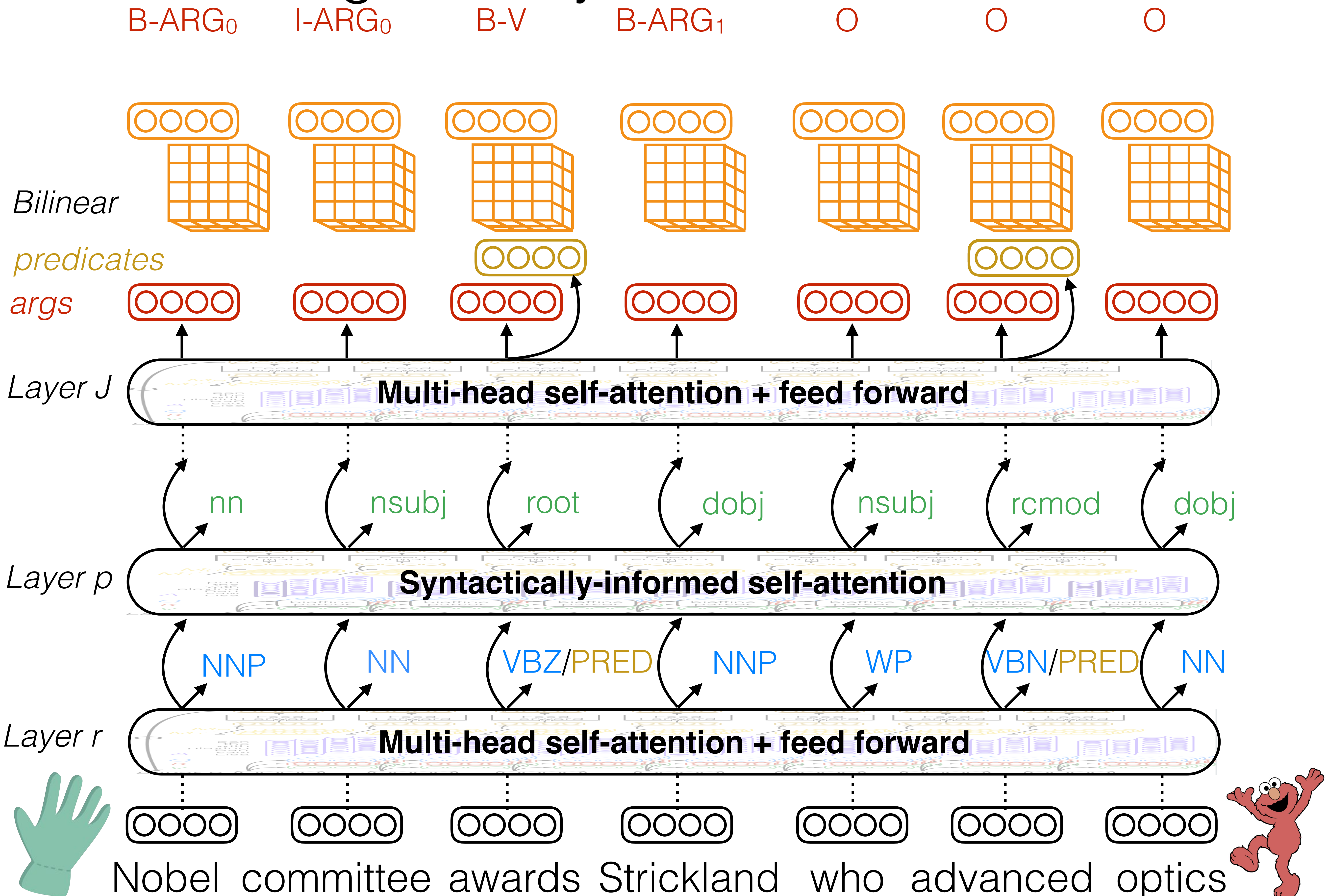
LISA: Linguistically-Informed Self-Attention



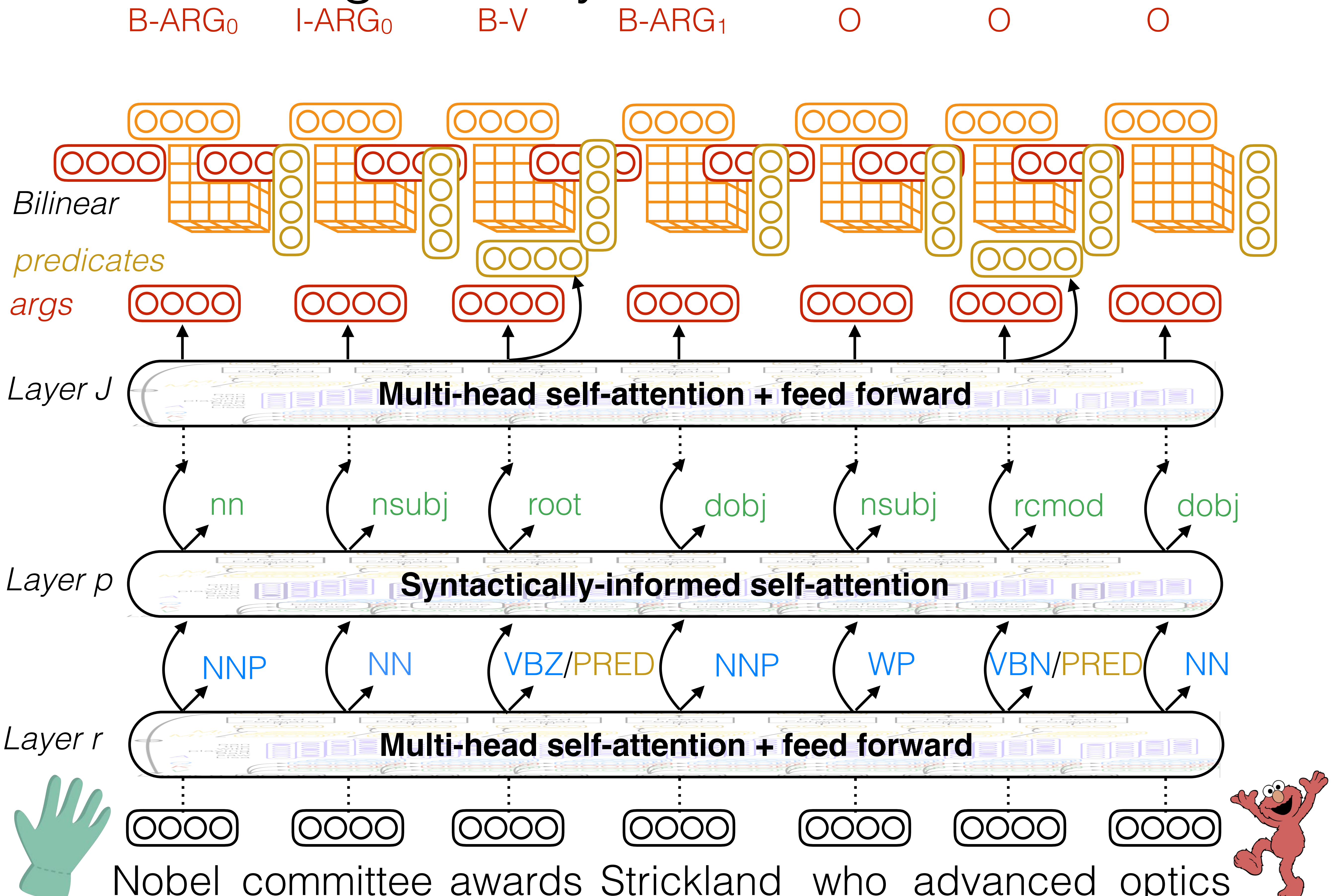
LISA: Linguistically-Informed Self-Attention



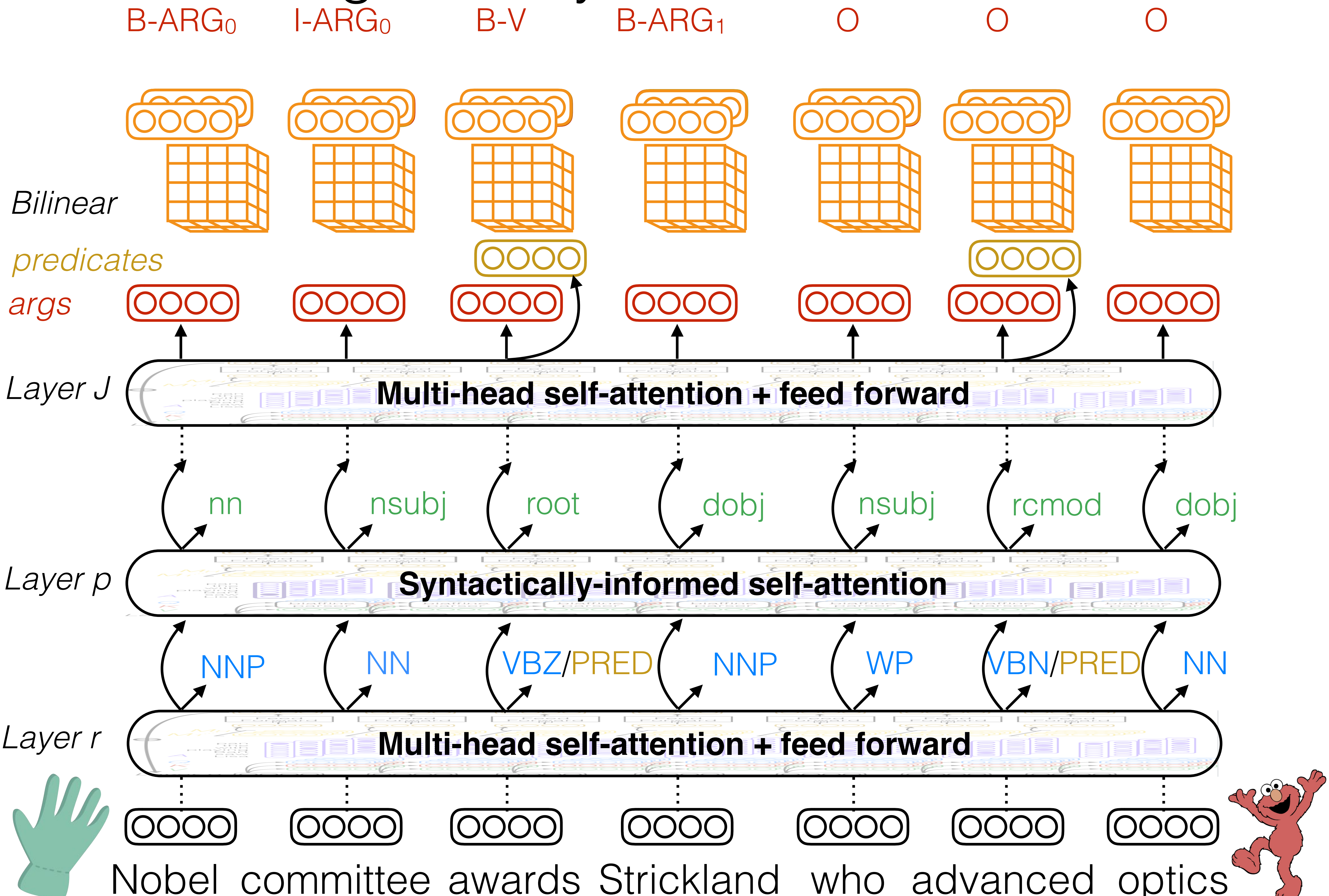
LISA: Linguistically-Informed Self-Attention



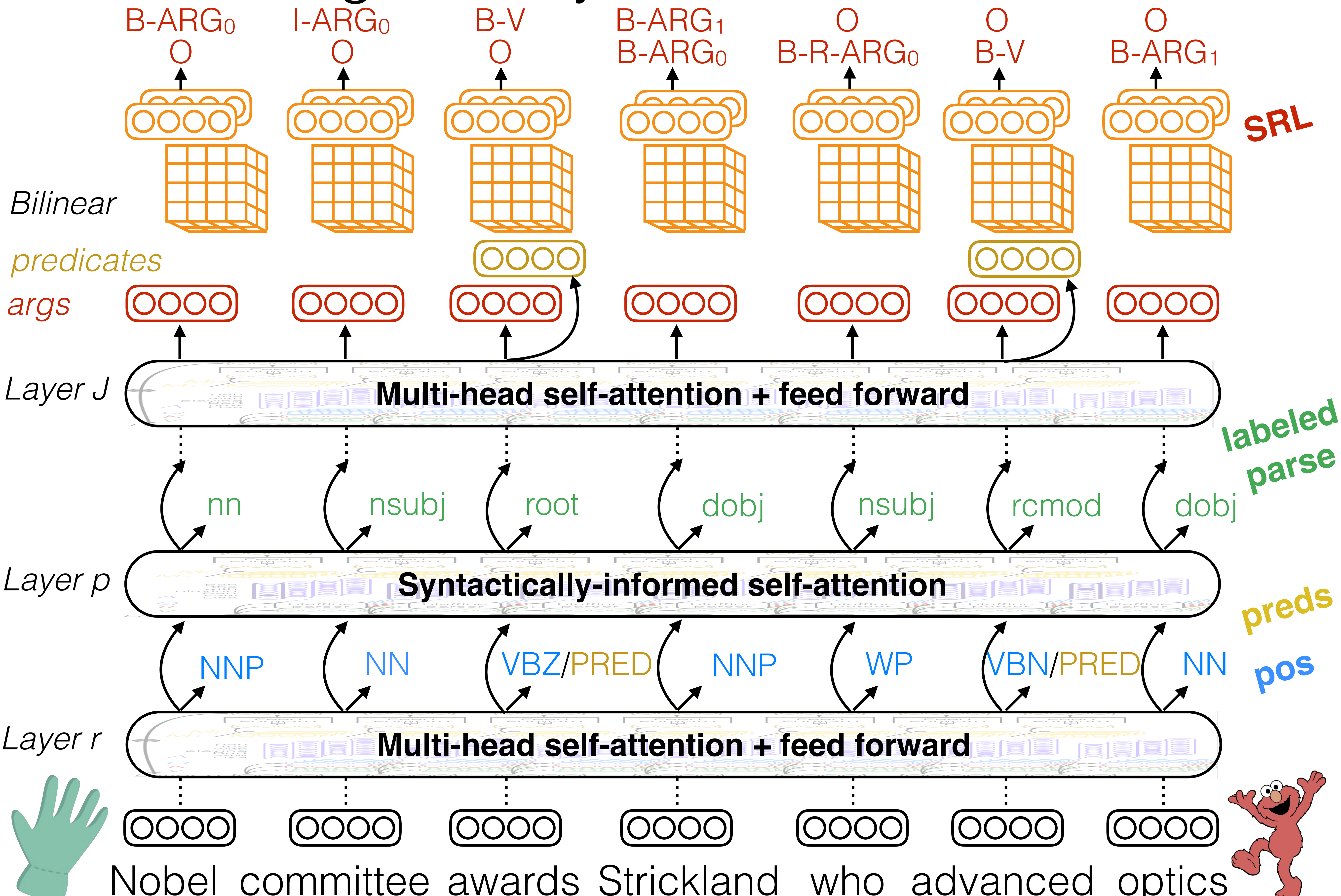
LISA: Linguistically-Informed Self-Attention



LISA: Linguistically-Informed Self-Attention

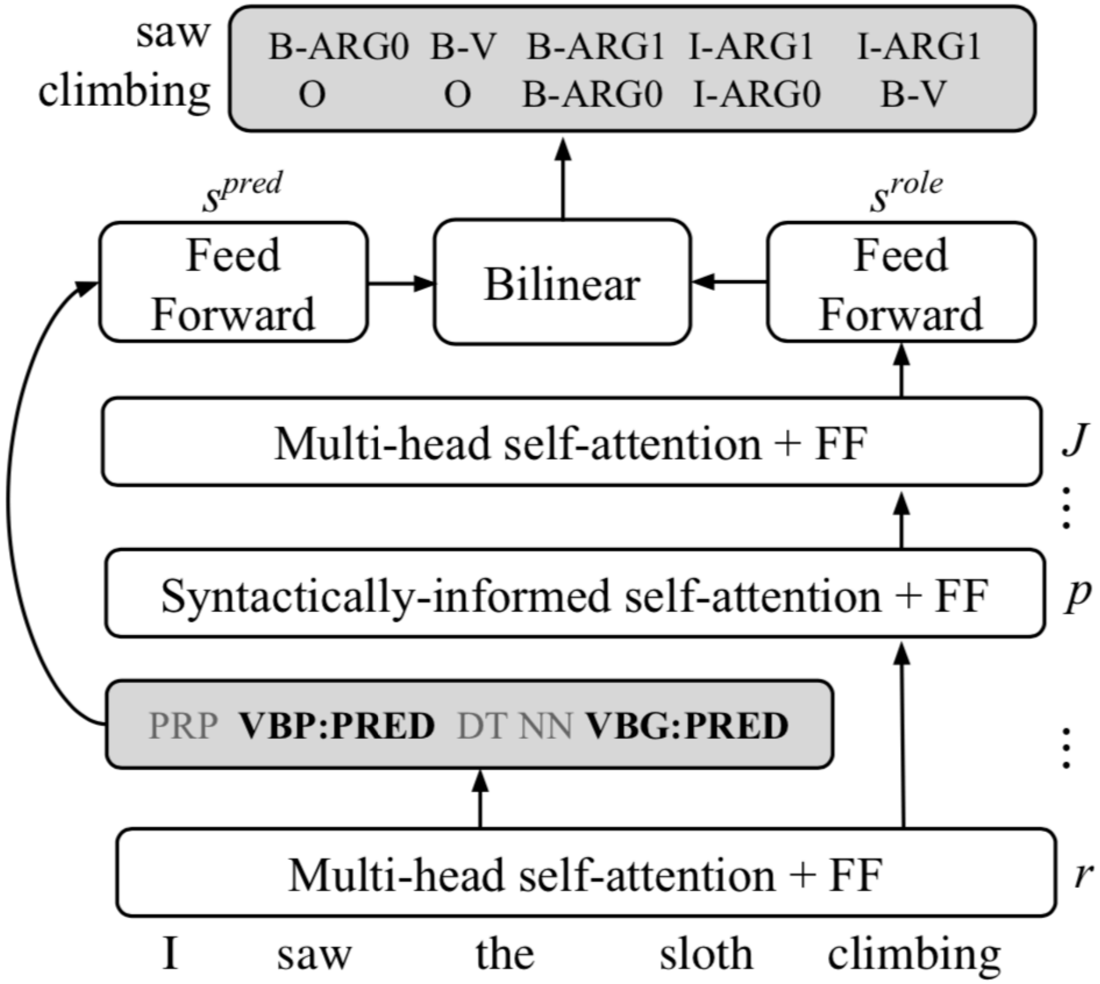


LISA: Linguistically-Informed Self-Attention



Outline

- Want fast, accurate, robust NLU
- PropBank SRL: Who did what to whom?
- 10 years of PropBank SRL
- LISA: Linguistically-informed self attention
  - Multi-head self-attention
  - Syntactically-informed self-attention
  - Multi-task learning, single-pass inference
  - Training
- Experimental results & error analysis



$$\frac{1}{T}\sum_{t=1}^{T}\Big[\sum_{f=1}^{F}\log P(y_{ft}^{role}\mid \mathcal{P}_G,\mathcal{V}_G,\mathcal{X}) + \log P(y_t^{prp}\mid\mathcal{X}) + \lambda_1 \log P(\mathrm{head}(t)\mid\mathcal{X}) + \lambda_2\log P(y_t^{dep}\mid\mathcal{P}_G,\mathcal{X})\Big]$$

y_{ft}^{role}: SRL Label

P_G: Gold Parse

$head(t)$, y_t^{dep}: Dependency Parsing Annotation (Edge Existence, Arc-label)

V_G: Gold Predicates

X: Input Sequence

y_t^{prp}: POS+Predicate Joint Label

Outline

- Want fast, accurate, robust NLU
- PropBank SRL: Who did what to whom?
- 10 years of PropBank SRL
- LISA: Linguistically-informed self attention
  - Multi-head self-attention
  - Syntactically-informed self-attention
  - Multi-task learning, single-pass inference
  - Training
- Experimental results & error analysis

Experimental results

	CoNLL-2005	**CoNLL-2012**
domains	Train, dev: news Test: news, novels	Train, dev, test: 7 domains (news, telephone, bible, ...)
word embeddings	GloVe [Pennington et al. 2014] ELMo [Peters et al. 2018]	GloVe [Pennington et al. 2014] ELMo [Peters et al. 2018]
predicates	predicted; gold	predicted
baselines	He et al. 2017 He et al. 2018 Tan et al. 2018	He et al. 2018
our models	SA LISA LISA+D&M, +Gold	SA LISA LISA+D&M, +Gold

Experimental results: CoNLL-2005

	GloVe		ELMo	
	in-domain	**out-of-domain**	**in-domain**	**out-of-domain**
He et al. 2017	82.7	70.1	---	---
He et al. 2018	82.5	70.8	86.0	76.1
SA	83.72	71.51	86.09	76.35
LISA	83.61	71.91	86.55	78.05
+D&M	84.99 94.9 UAS	74.66 90.3 UAS	86.90 96.3 UAS	78.25 93.4 UAS
	+2.49 F1	+3.86 F1	+0.9 F1 ?	+2.15 F1

Experimental results: CoNLL-2005

	GloVe	ELMo
	in-domain (dev)	in-domain (dev)

96.5 UAS!

Experimental results: CoNLL-2005

Gold predicates; GloVe embeddings

WSJ Test (in-domain):

	Precision	Recall	F1
He et al. 2018	84.2	83.7	83.9
Tan et al. 2018	81.2	83.9	84.8
SA	84.7	84.24	84.47
LISA	84.72	84.57	84.64
+D&M	86.02	86.05	86.04

Brown Test (out-of-domain):

	Precision	Recall	F1
He et al. 2018	74.2	73.1	73.7
Tan et al. 2018	73.5	74.6	74.1
SA	73.89	72.39	73.13
LISA	74.77	74.32	74.55
+D&M	76.65	76.44	76.54

Experimental results: CoNLL-2012

Predicted predicates

	Precision	**Recall**	**F1**
He et al. 2018	79.4	80.1	79.8
SA	82.55	80.02	81.26
LISA	81.86	79.56	80.70
+D&M	83.3	81.38	82.33

	Precision	**Recall**	**F1**
He et al. 2018	81.9	84.0	82.9
SA	84.39	82.21	83.28
LISA	83.97	82.29	83.12
+D&M	84.14	82.64	83.38

Experimental results: POS&Parsing

Data	Model	POS	UAS	LAS
WSJ	D&M$_E$	—	96.48	94.40
	LISA$_G$	96.92	94.92	91.87
	LISA$_E$	97.80	96.28	93.65
Brown	D&M$_E$	—	92.56	88.52
	LISA$_G$	94.26	90.31	85.82
	LISA$_E$	95.77	93.36	88.75
CoNLL-12	D&M$_E$	—	94.99	92.59
	LISA$_G$	96.81	93.35	90.42
	LISA$_E$	98.11	94.84	92.23

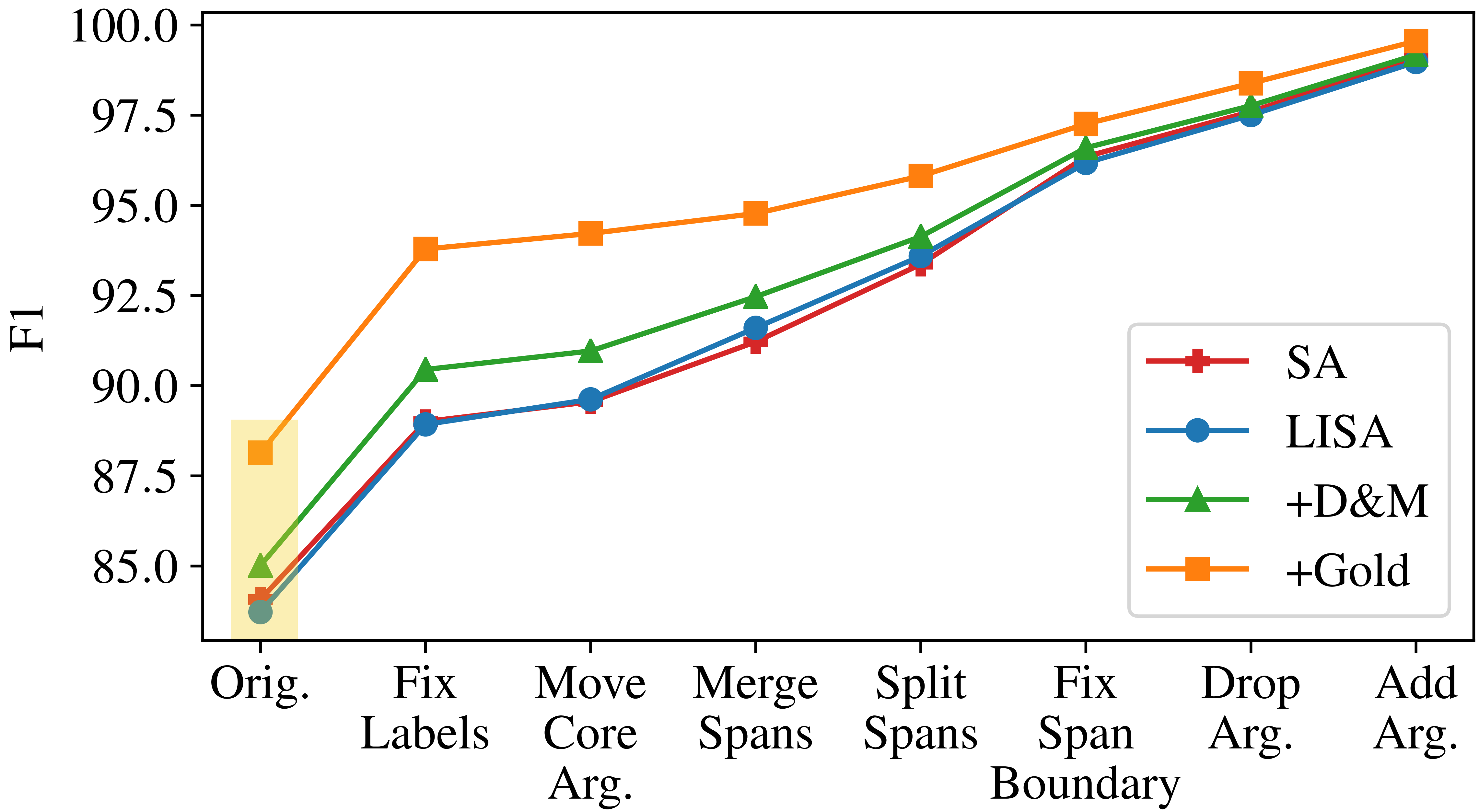
Table 4: Parsing (labeled and unlabeled attachment) and POS accuracies attained by the models used in SRL experiments on test datasets. Subscript G denotes GloVe and E ELMo embeddings.

Experimental results: Predicate Prediction

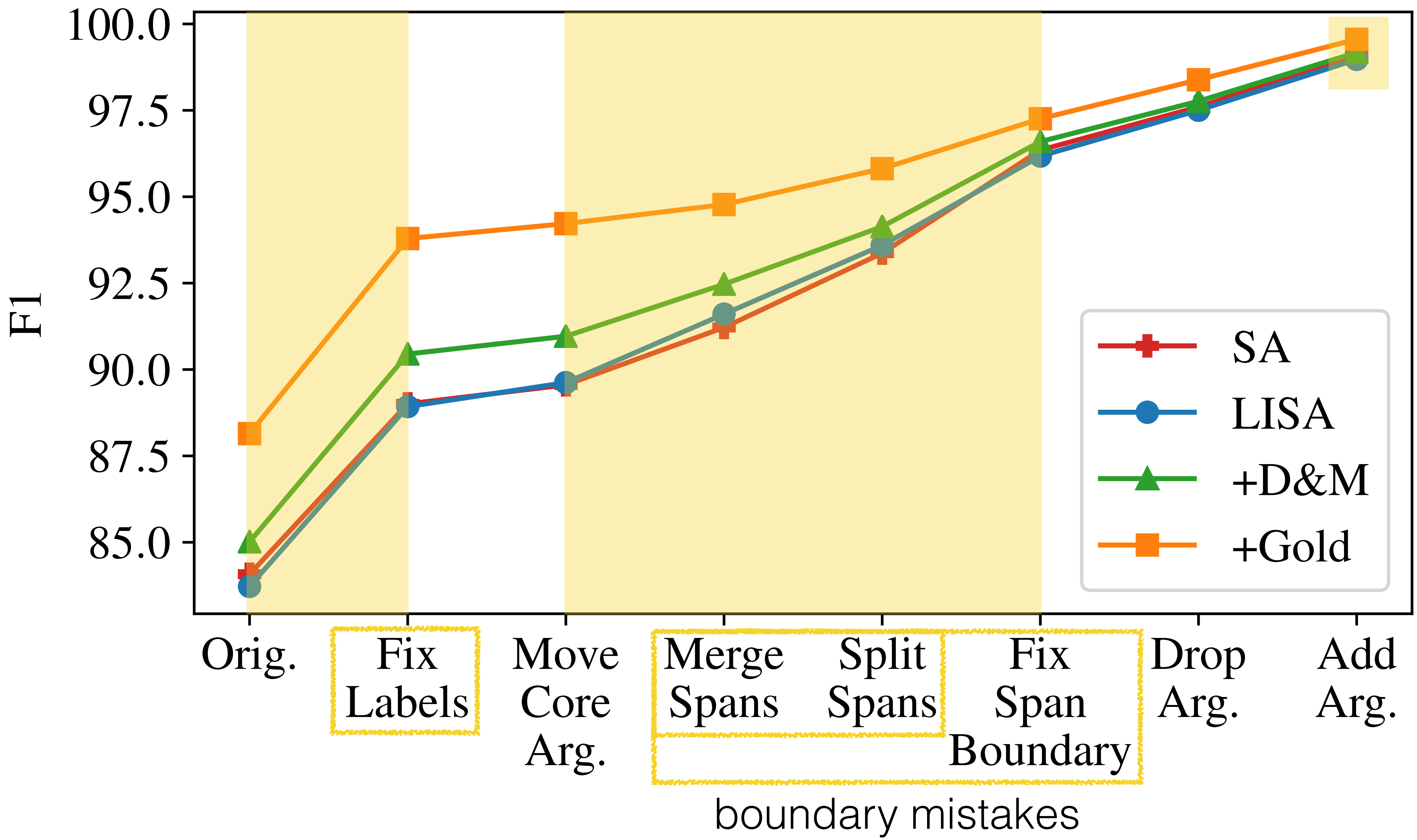
	Model	P	R	F1
WSJ	He et al. (2017)	94.5	98.5	96.4
	LISA	98.9	97.9	98.4
Brown	He et al. (2017)	89.3	95.7	92.4
	LISA	95.5	91.9	93.7
CoNLL-12	LISA	99.8	94.7	97.2

Table 5: Predicate detection precision, recall and F1 on CoNLL-2005 and CoNLL-2012 test sets.

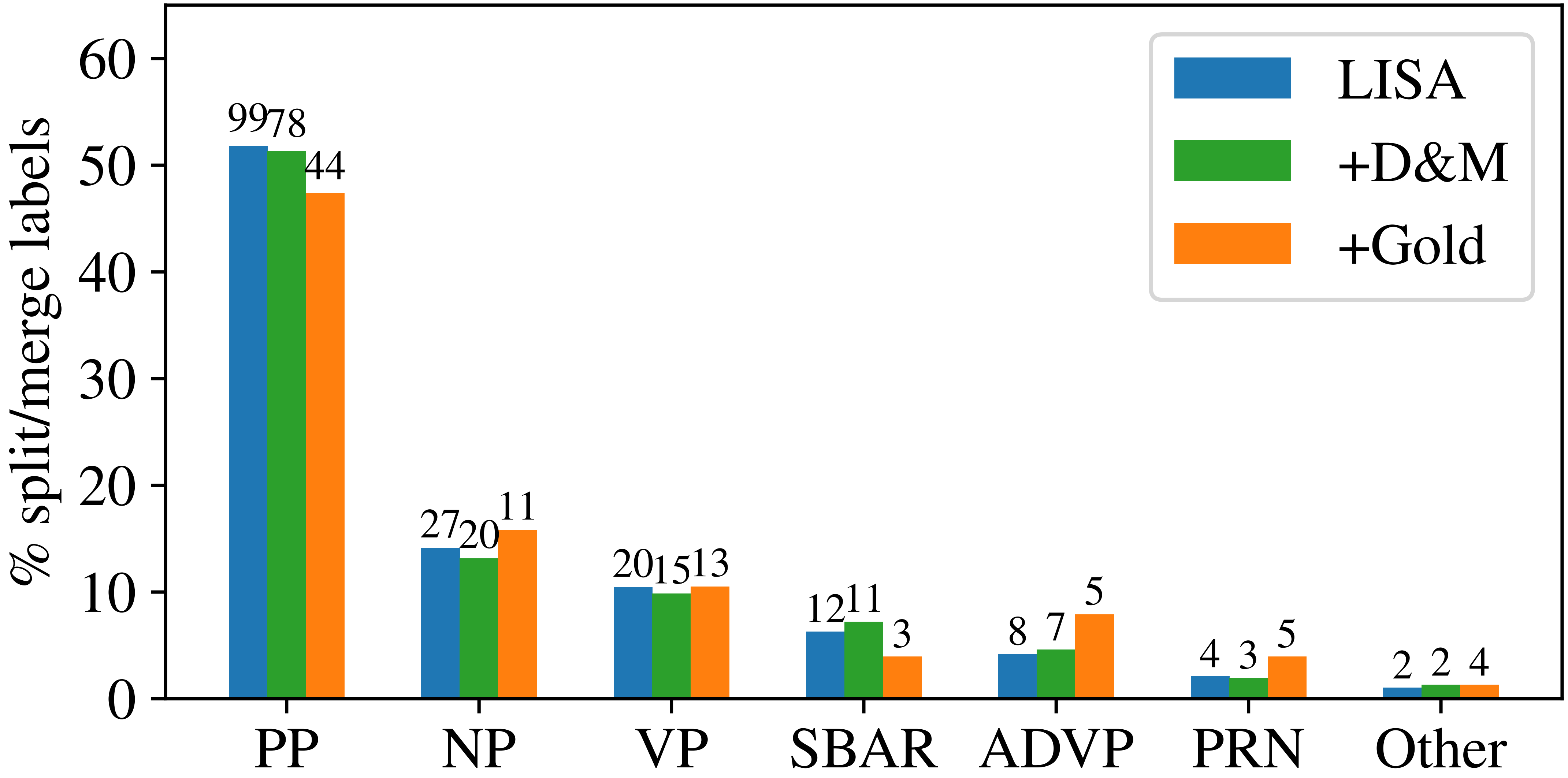
Experimental results: Analysis



Experimental results: Analysis



Experimental results: Analysis



Summary

Thank you!

- **LISA**: Multi-task learning + multi-head self attention trained to attend to syntactic parents
  - Achieves state-of-the-art F1 on PropBank SRL
  - Linguistic structure improves generalization
  - Fast: encodes sequence *only once* to predict predicates, parts-of-speech, labeled dependency parse, SRL

- Everyone wants to run NLP on the entire web:
  - **accuracy**, **robustness**, **computational efficiency**.

Models & Code: `https://github.com/strubell/LISA`