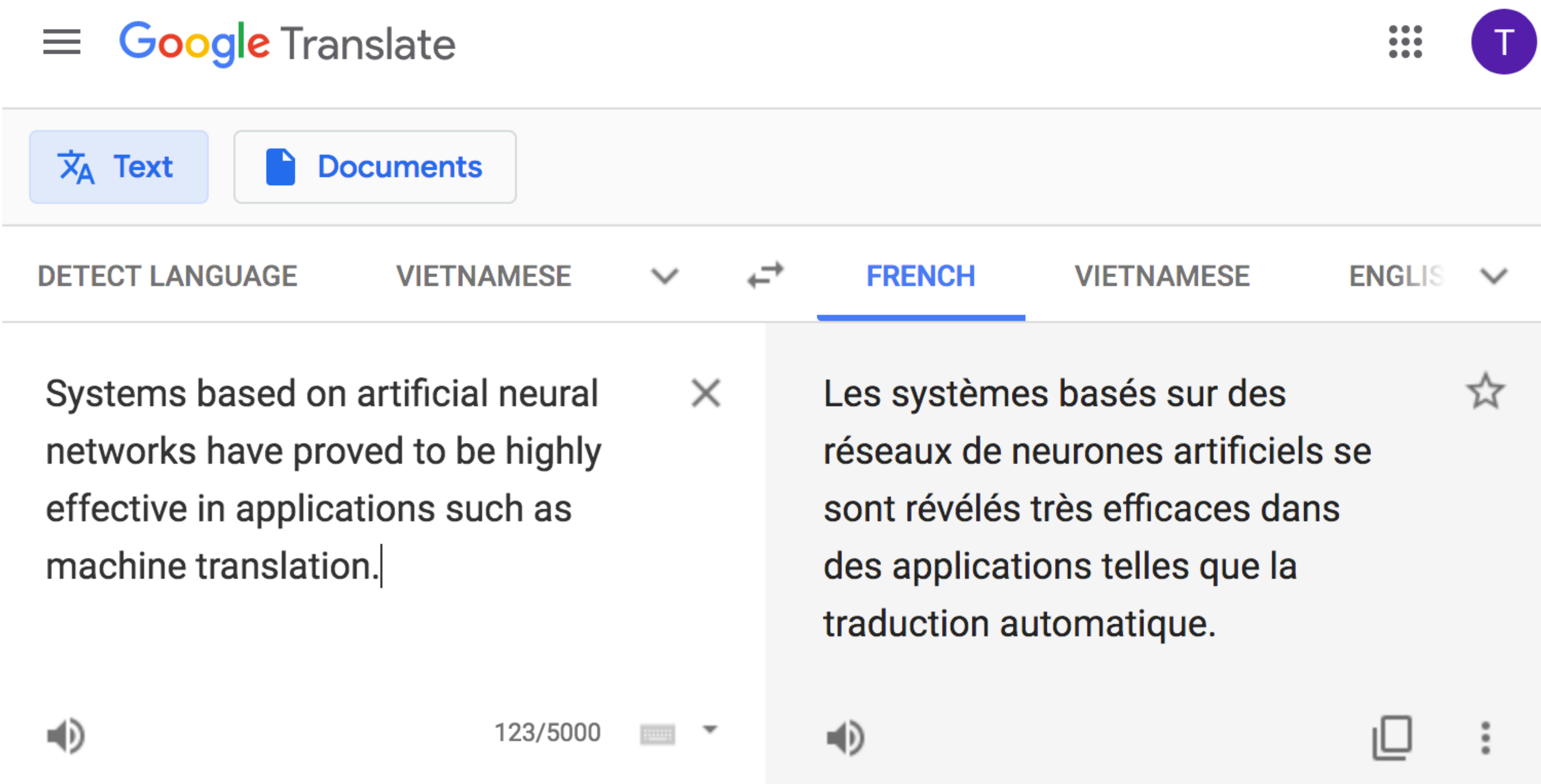
Analyzing and interpreting neural networks for NLP

Tal Linzen
Department of Cognitive Science
Johns Hopkins University

Neural networks are remarkably effective in language technologies



Language modeling

The boys went outside to _____

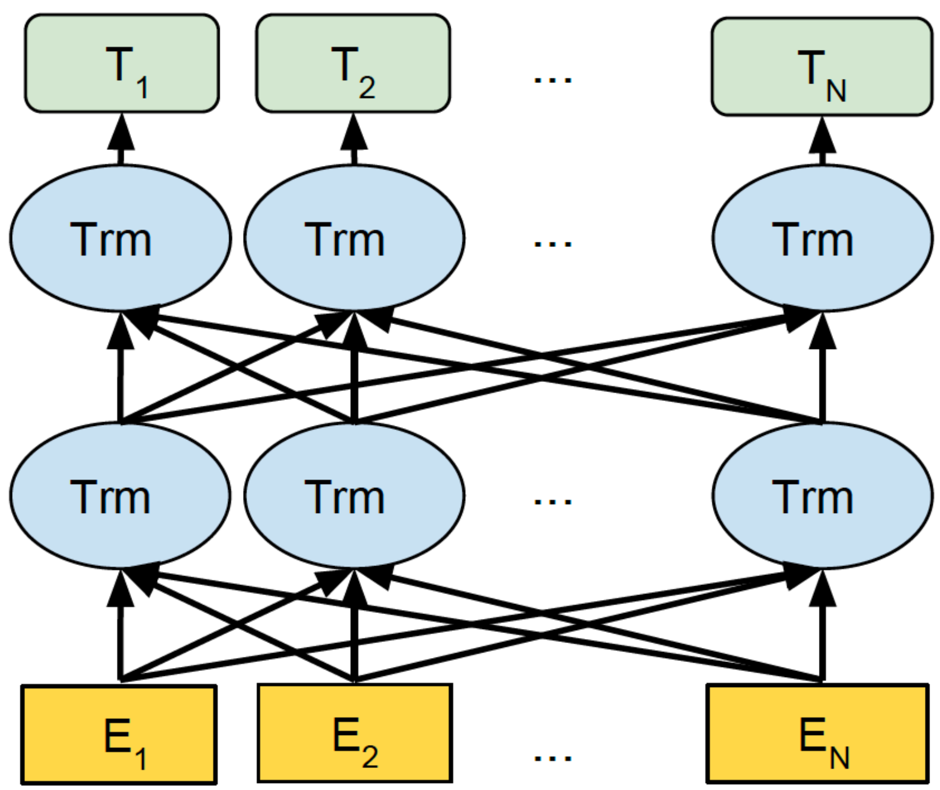
$$\hat{P}(w_n = w^k \mid w_1, \ldots, w_{n-1})$$

MODEL	TEST PERPLEXITY	NUMBER OF PARAMS [BILLIONS]
SIGMOID-RNN-2048 (JI ET AL., 2015A)	68.3	4.1
INTERPOLATED KN 5-GRAM, 1.1B N-GRAMS (CHELBA ET AL., 2013)	67.6	1.76
SPARSE NON-NEGATIVE MATRIX LM (SHAZEER ET AL., 2015)	52.9	33
RNN-1024 + MAXENT 9-GRAM FEATURES (CHELBA ET AL., 2013)	51.3	20
LSTM-512-512	54.1	0.82
LSTM-1024-512	48.2	0.82
LSTM-2048-512	43.7	0.83
LSTM-8192-2048 (NO DROPOUT)	37.9	3.3
LSTM-8192-2048 (50% DROPOUT)	32.2	3.3
2-LAYER LSTM-8192-1024 (BIG LSTM)	30.6	1.8
BIG LSTM+CNN INPUTS	**30.0**	**1.04**

(Jozefowicz et al., 2016)

The interpretability challenge

- The network doesn't follow human-designed rules
- Its internal representations are not formatted in a human-readable way
- What is the network doing, how, and why?



Why do interpretability and explainability matter?

Apple Card is accused of gender bias. Here's how that can happen

By Evelina Nedlund, CNN Business

Updated 2:04 PM ET, Tue November 12, 2019

New York (CNN Business) — Some Apple Card customers say the credit card's issuer, Goldman Sachs, is giving women far lower credit limits, even if they share assets and accounts with their spouse. But it's impossible to know if the Apple Card -- or any other credit card -- discriminates against women, because creditworthiness algorithms are notoriously opaque.

"It's such a mystery we are seeing," said Sara Rathner, travel and credit cards expert at NerdWallet. "Because we don't know exactly what those algorithms are looking for, it can be hard to say if there might be some bias built into them."

https://www.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html

Why do interpretability and explainability matter?

- We are typically uncomfortable with having a system we do not understand make decisions with significant societal and ethical consequences (or other high-stakes consequences)
- Examples: the criminal justice system, health insurance, hiring, loans
- If we don't understand why the system made a decision, we cannot judge whether it conforms to our values

Why do interpretability and explainability matter?

- Human-in-the-loop settings: cooperation between humans and ML systems
- Debugging neural networks
- Scientific understanding and cognitive science:
  - A system that performs a task well can help generate hypotheses for how humans might perform it
  - Those hypotheses would be more useful if they were interpretable to a human (the “customer” of the explanation)

Outline

- Using behavioral experiments to characterize what the network learned (“psycholinguistics on neural networks”)
- What information is encoded in intermediate vectors? (“artificial neuroscience”)
- Interpreting attention weights
- Symbolic approximations of neural networks

Outline

- **Using behavioral experiments to characterize what the network learned**
- What information is encoded in intermediate vectors? (“artificial neuroscience”)
- Interpreting attention weights
- Symbolic approximations of neural networks
- Interpretable models

Linguistically targeted evaluation

- Average metrics (such as perplexity) are primarily affected by frequent phenomena: those are often very simple

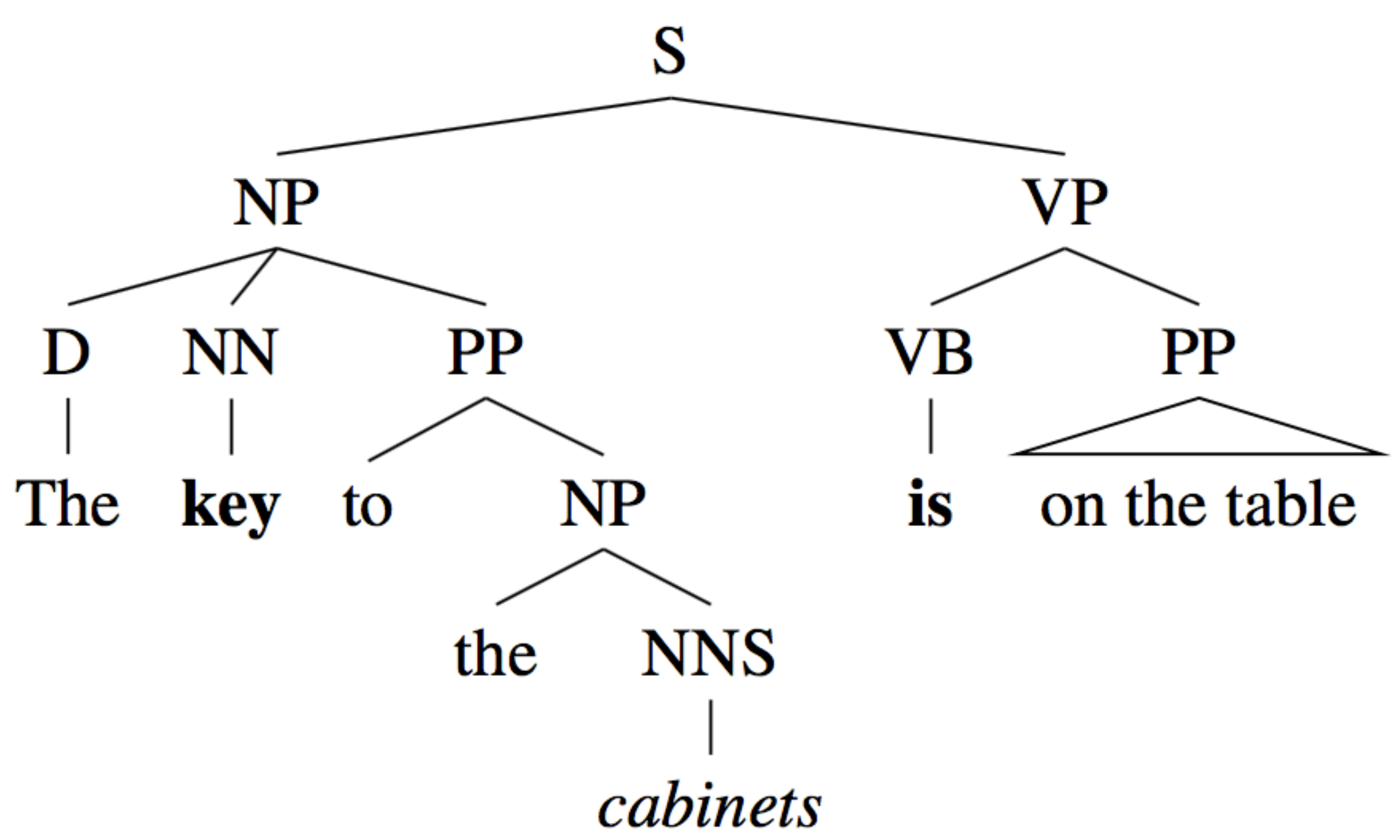
- Effective word prediction on the average case can be due to collocations, semantics, syntax… Is the model capturing all of these?

- How does the model **generalize** to (potentially infrequent) cases that probe a particular linguistic ability?

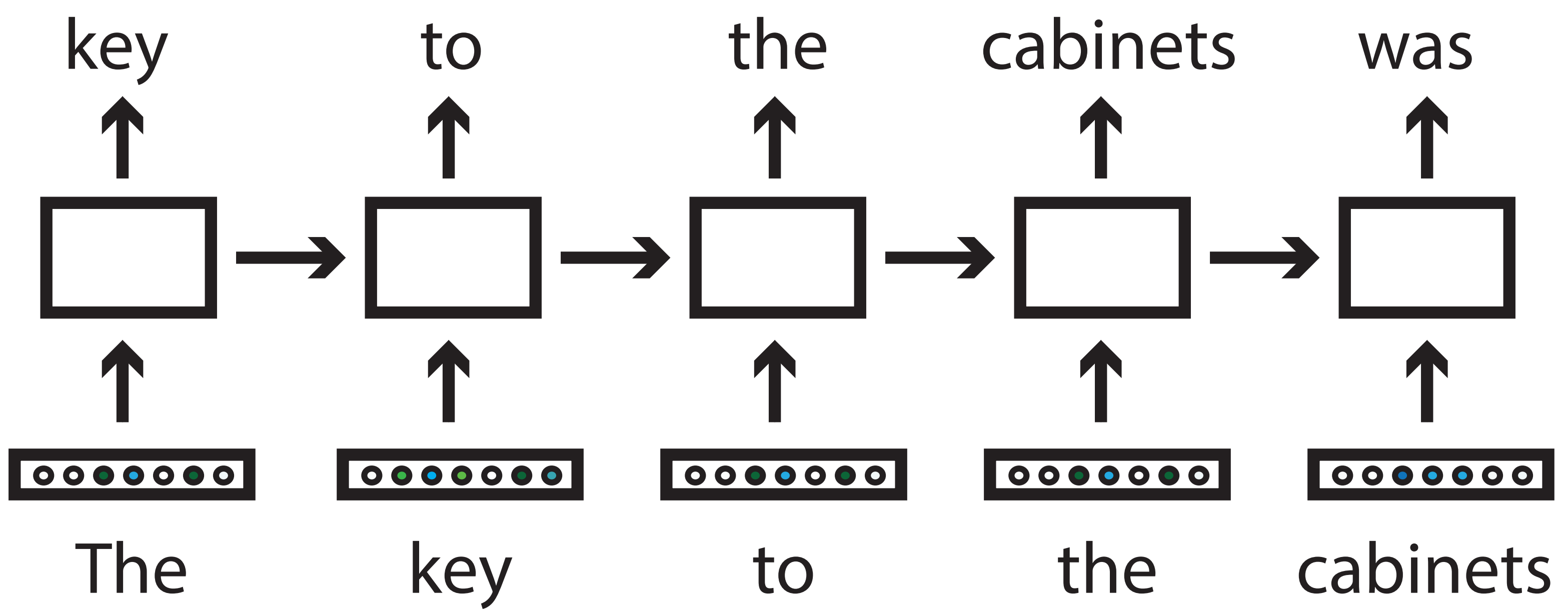
- Behavioral evaluation of a system as a whole rather than of individual vector representations

Syntactic evaluation with subject-verb agreement

The **key** to the **cabinets** **is** on the table.



Evaluating syntactic predictions in a language model



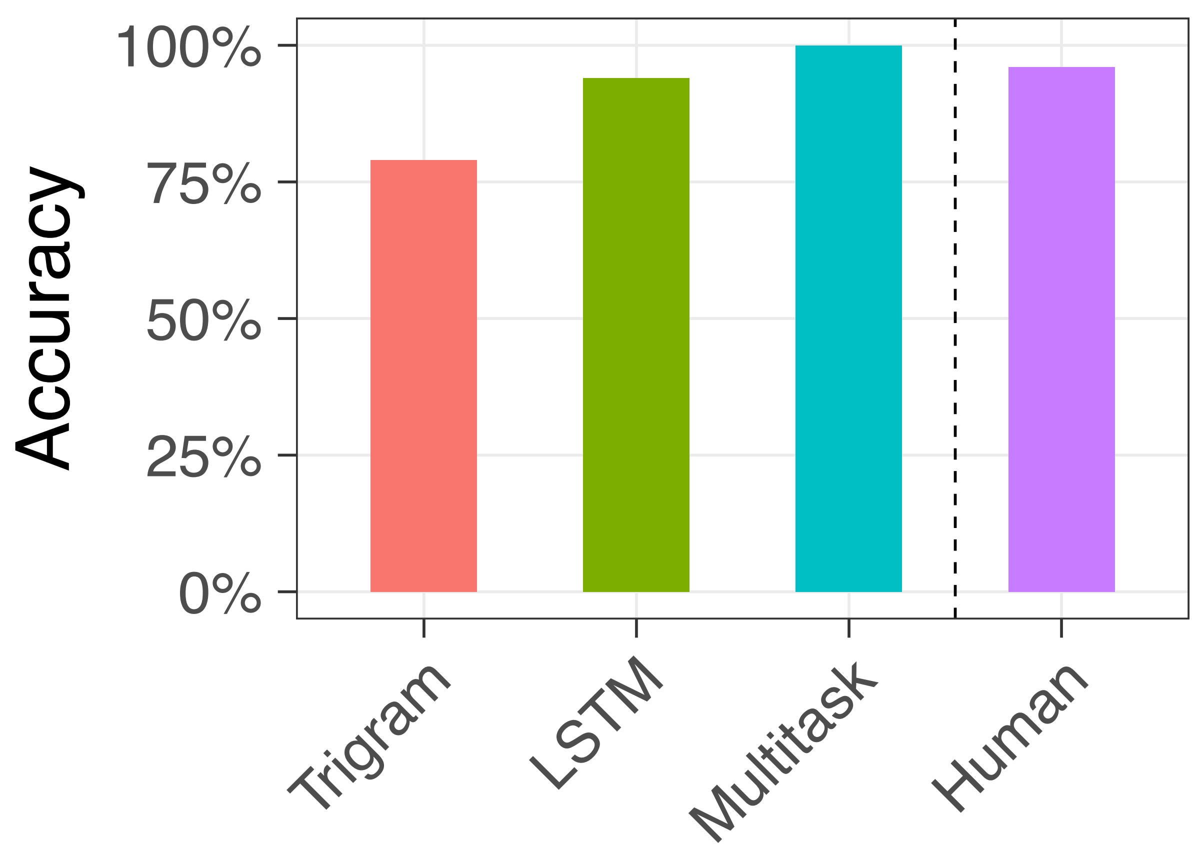
- *The key to the cabinets….* P(*was*) > P(*were*)?

(Linzen, Dupoux & Goldberg, 2016, *TACL*)

Agreement in a simple sentence

The **author laughs**.

*The **author laugh**.

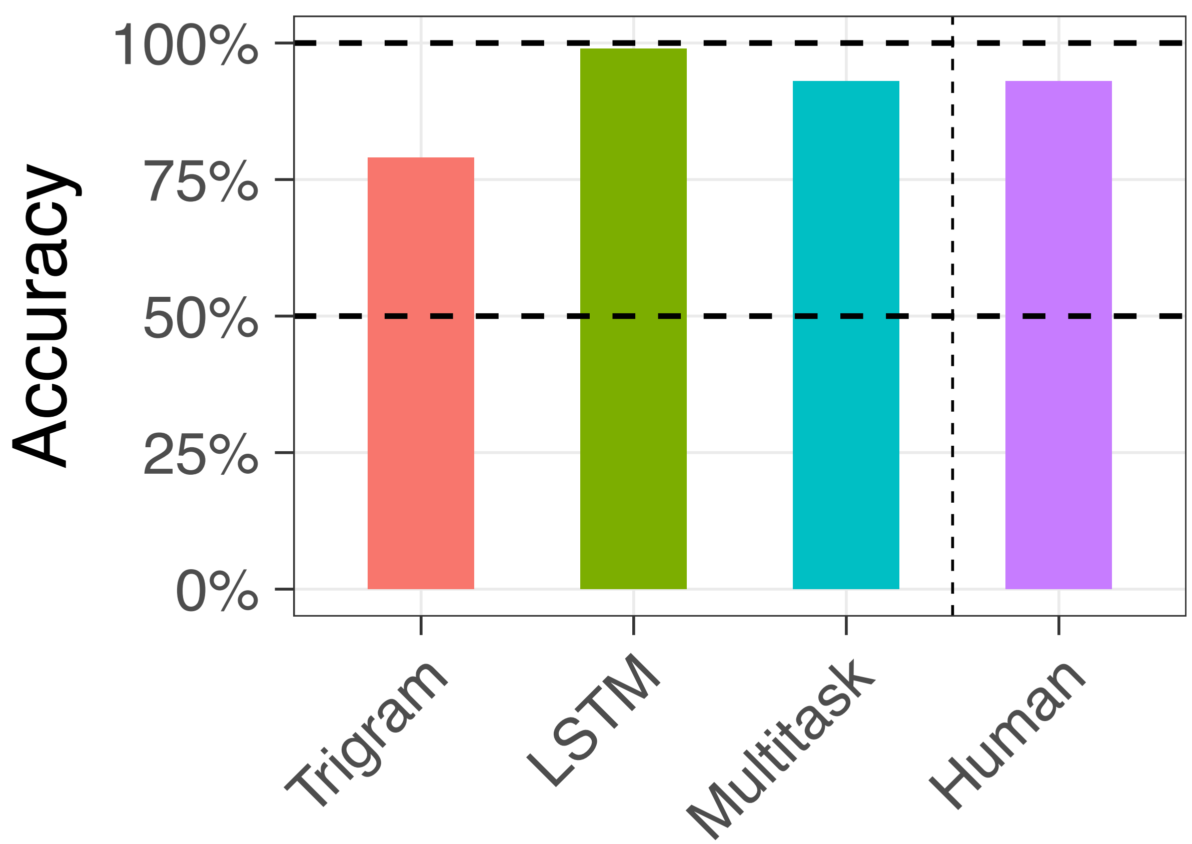


(Marvin & Linzen, 2018, EMNLP)

Agreement in a sentential complement

The **mechanics** said the security **guard laughs**.

*The **mechanics** said the security **guard laugh**.



No interference from sentence-initial noun

(Marvin & Linzen, 2018, EMNLP)

Most sentences are simple; focus on dependencies with attractors

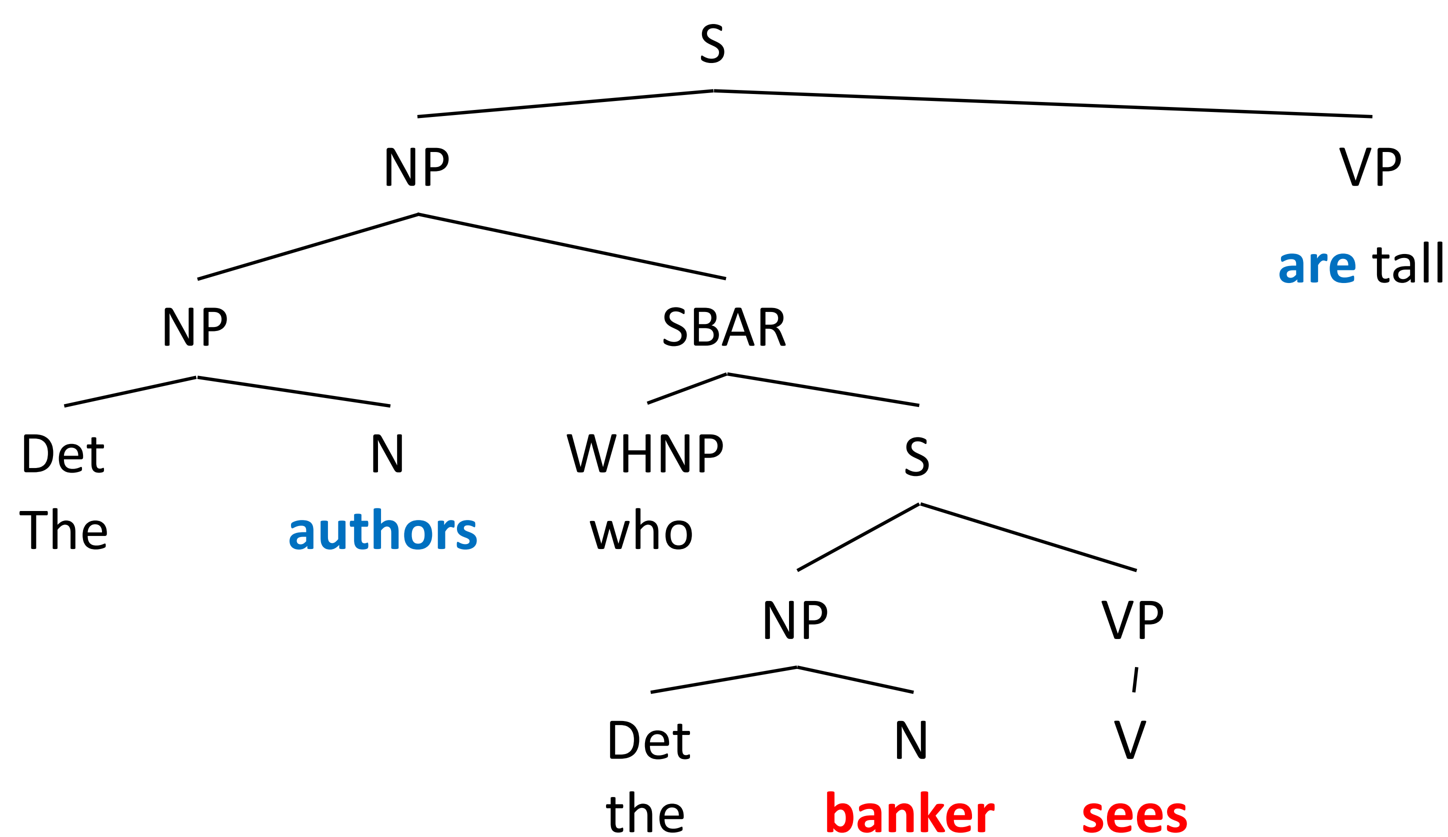
- The **keys are** rusty.
- The **keys** to the cabinet **are** rusty.
- The **ratio** of men to women **is** not clear.
- The **ratio** of men to women and children **is** not clear.
- ~~The **keys** to the cabinets **are** rusty.~~
- ~~The **keys** to the door and the cabinets **are** rusty.~~
- **Evaluation only: the model is still trained on all sentences!**

RNNs' inductive bias favors short dependencies (recency)!
(Ravfogel, Goldberg & Linzen, 2019, *NAACL*)

Agreement across an object relative clause

The **authors** who the **banker** sees **are** tall.

*The **authors** who the **banker** sees **is** tall.

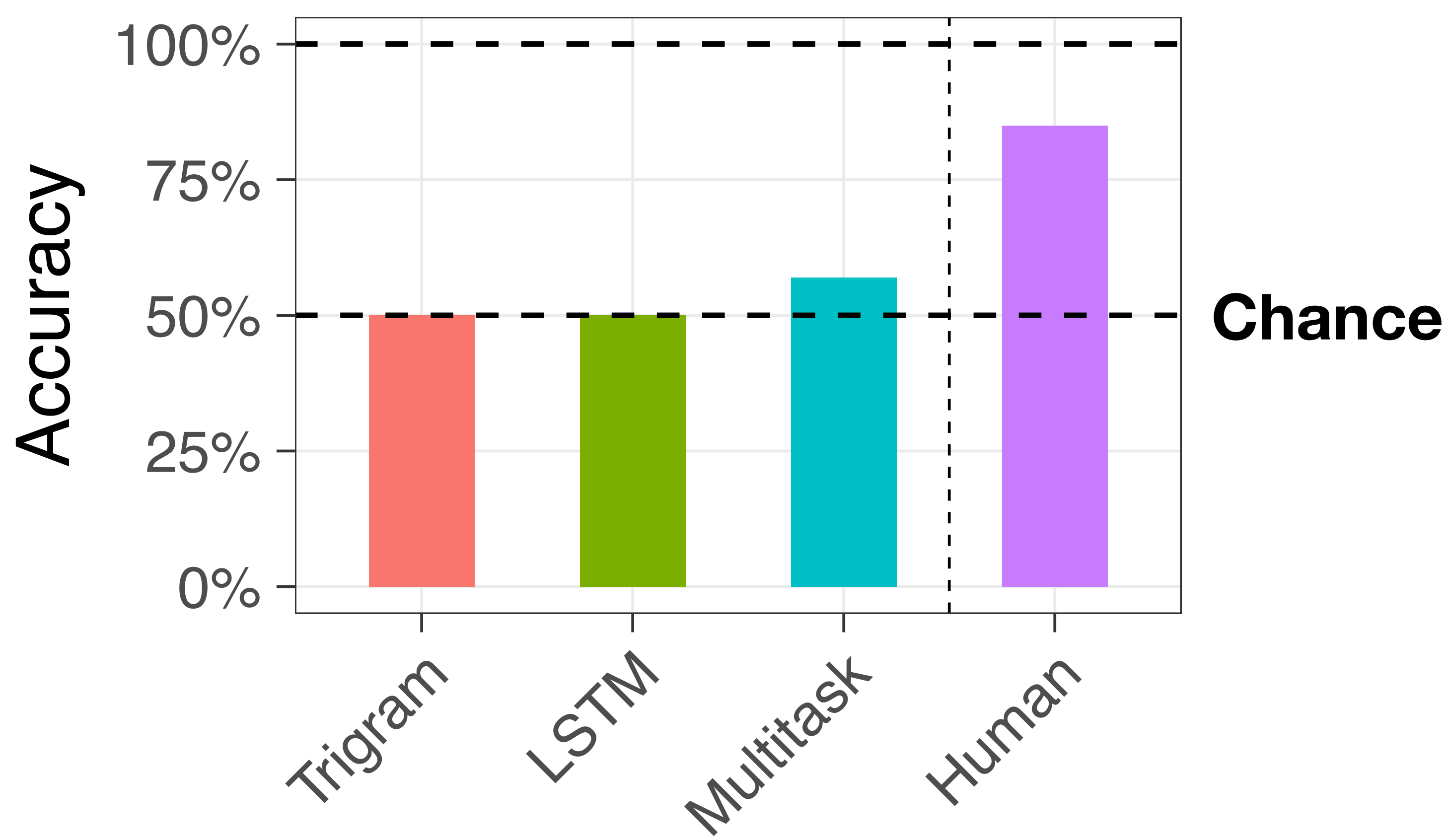


Agreement across an object relative clause

The **authors** who the **banker** sees **are** tall.

*The **authors** who the **banker** sees **is** tall.

Multitask learning with syntax barely helps…



(Marvin & Linzen, 2018, EMNLP)

Adversarial examples

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

(Jia and Liang, 2017, EMNLP)

Adversarial examples indicate that the model is sensitive to factors that are not the ones we think it should be sensitive to

Adversarial examples

Prepending a single word to SNLI hypotheses:

Ground Truth	Trigger	ESIM	DA	DA-ELMo
		89.49	89.46	90.88
	nobody	0.03	0.15	0.50
	never	0.50	1.07	0.15
Entailment	sad	1.51	0.50	0.71
	scared	1.13	0.74	1.01
	championship	0.83	0.06	0.77
	Avg. Δ	-88.69	-88.96	-90.25

Triggers transfer across models! (Likely because they reflect dataset bias and neural models are very good at latching onto that)

(Wallace et al., 2019, EMNLP)

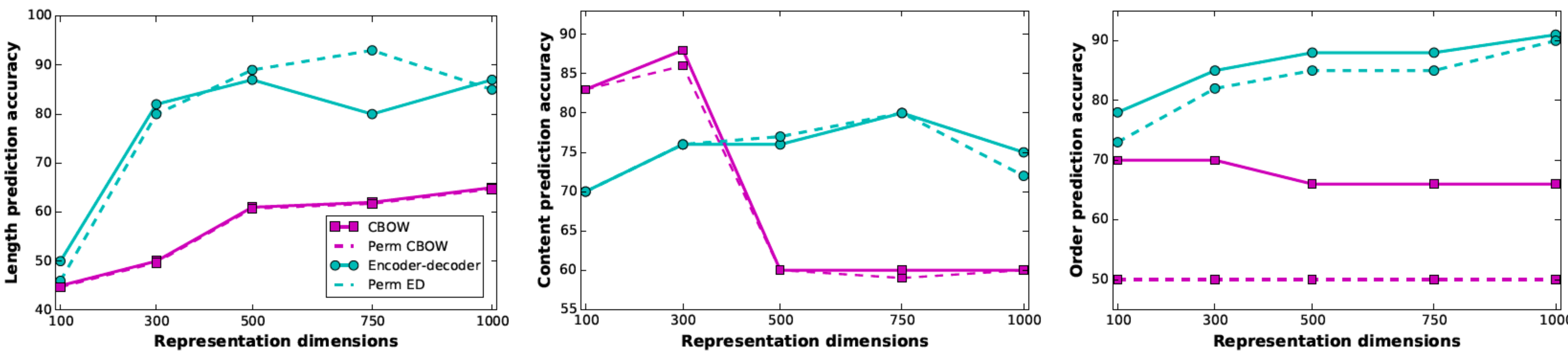
Outline

- Using behavioral experiments to characterize what the network learned ("psycholinguistics on neural networks")
- **What information is encoded in intermediate vectors? ("artificial neuroscience")**
- Interpreting attention heads
- Symbolic approximations of neural networks

Diagnostic classifier

- Train classifier to predict a property of a sentence embedding (supervised!)
- Test it on new sentences

(Adi et al., 2017, ICLR)

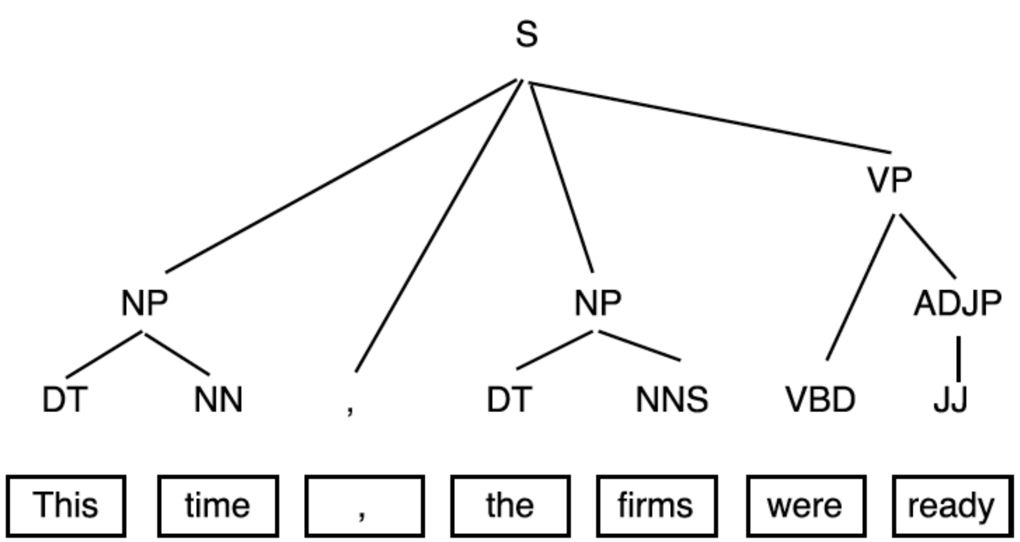


(a) Length test. (b) Content test. (c) Order test.

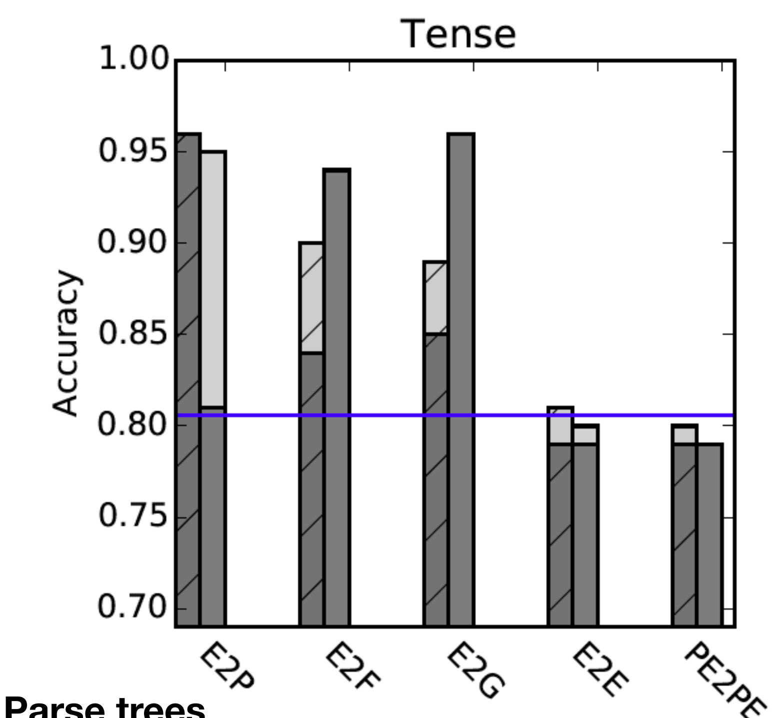
(Eight length bins) **(Does w appear in s?)** **(Does w_1 appear before w_2?)**

Diagnostic classifier

Hidden state of a 2-layer LSTM NMT system

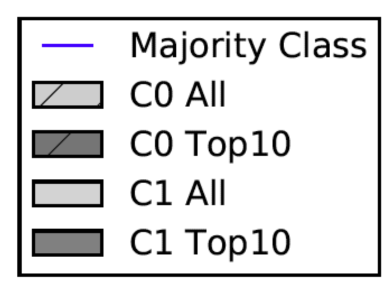


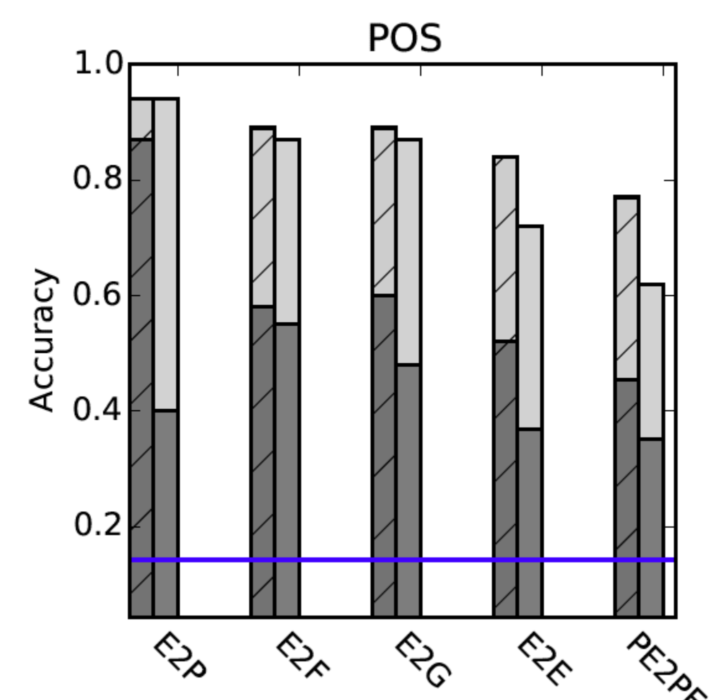
POS	DT	NN	,	DT	NNS	VBD	JJ
Smallest Phrase Constituent	NP	NP	-	NP	NP	VP	ADJP
Top-level Syntactic Sequence	(NP, NP, VP)						
Voice	Active						
Tense	Past						



Parse trees

French **German**





<u>**(Shi, Padhi & Knight, 2016, EMNLP)**</u>

Effect of power of probing model

Probing Model	NER	GED	Conj	GGParent
Linear	82.85	29.37	38.72	67.50
MLP (1024d)	87.19	47.45	55.09	78.80
LSTM (200d) + Linear	**88.08**	**48.90**	**78.21**	**84.96**
BiLSTM (512d) + MLP (1024d)	90.05	48.34	87.07	90.38

(Liu et al., 2019, NAACL)

(All models trained on top of ELMo;
GED = Grammatical error detection,
Conj = conjunct identification,
GGParent = label of great-grandparent in constituency tree)

What does it mean for something to be represented?

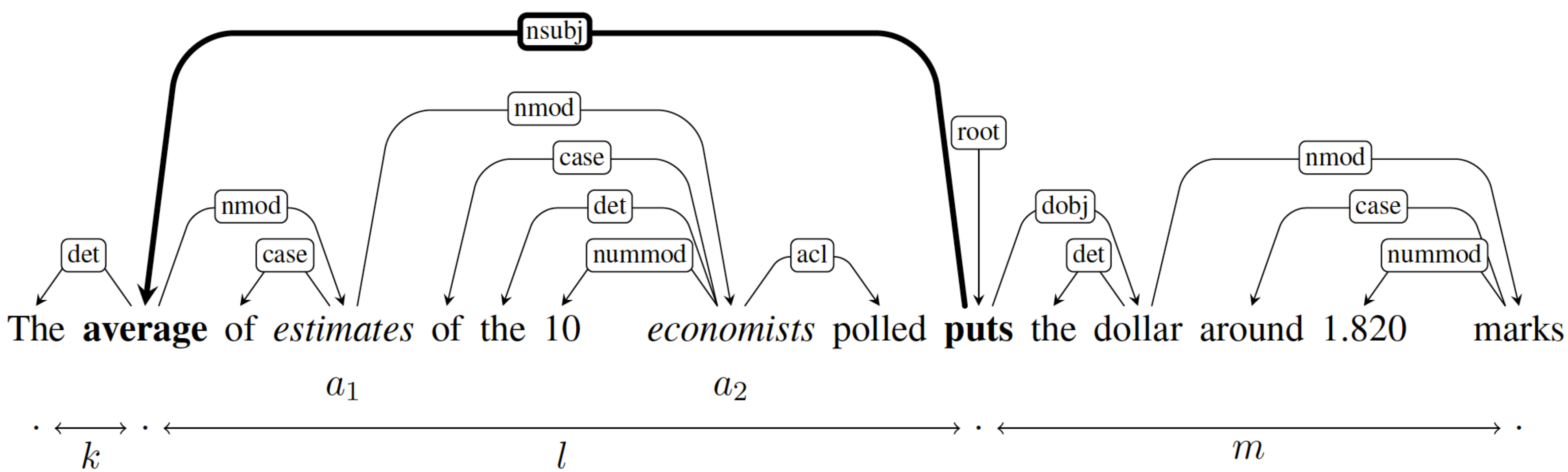
- The information can be recovered from the intermediate encoding

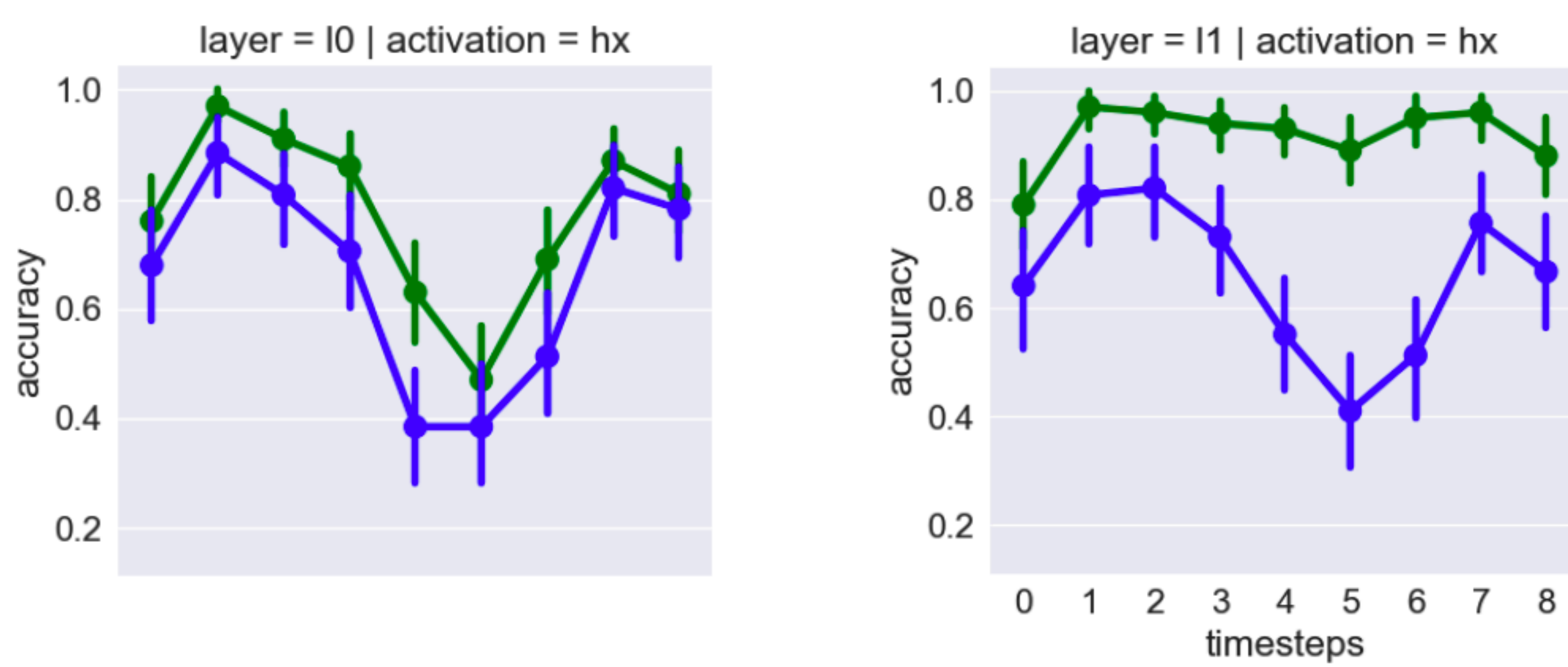
- The information can be recovered using a “simple” classifier (simple architecture, or perhaps trained on a small number of examples)

- The information can be recovered by the downstream process (e.g., linear readout)

- The information is in fact used by the downstream process

Diagnostic classifier

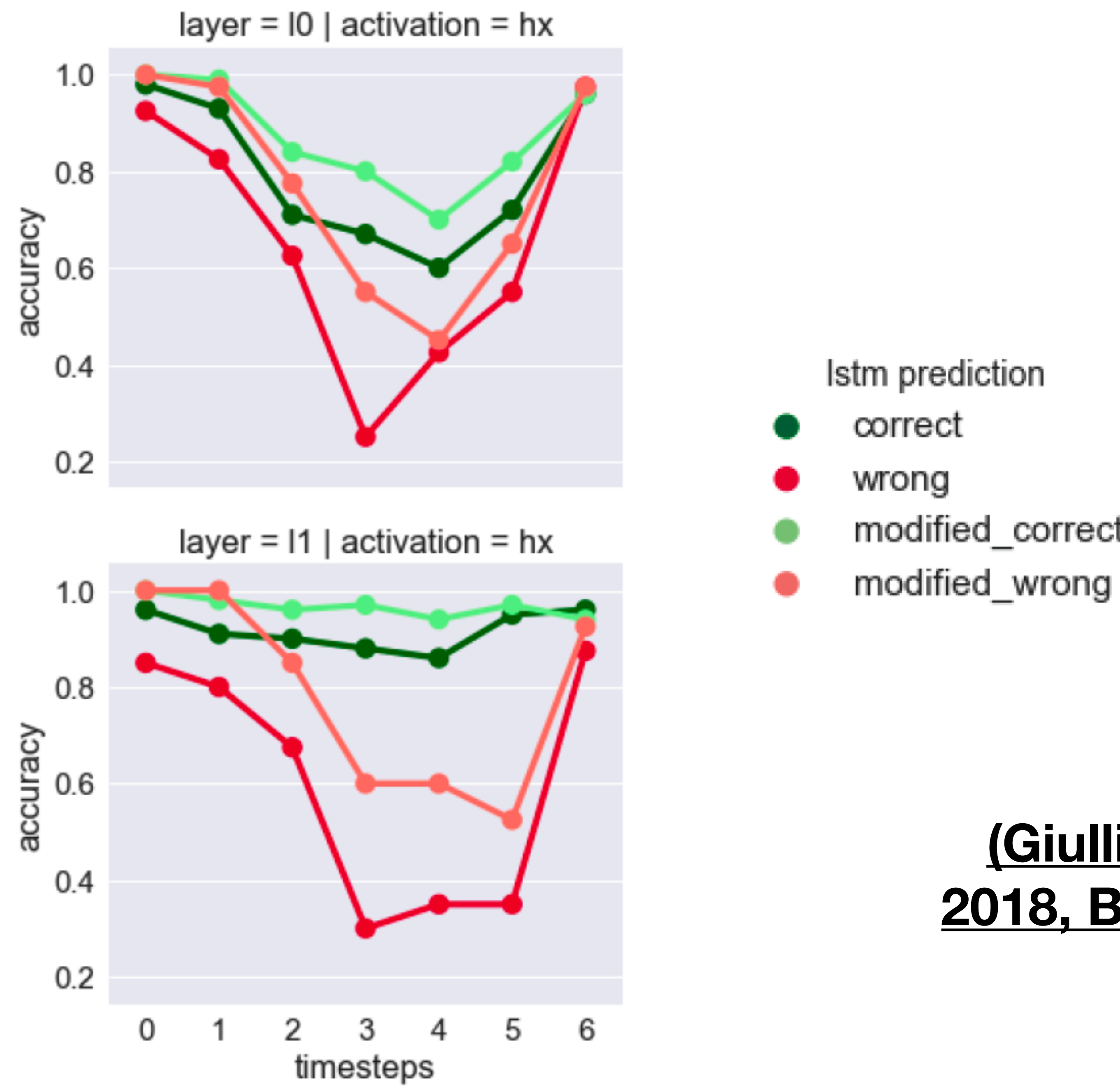




(Giullianeli et al., 2018, BlackboxNLP)

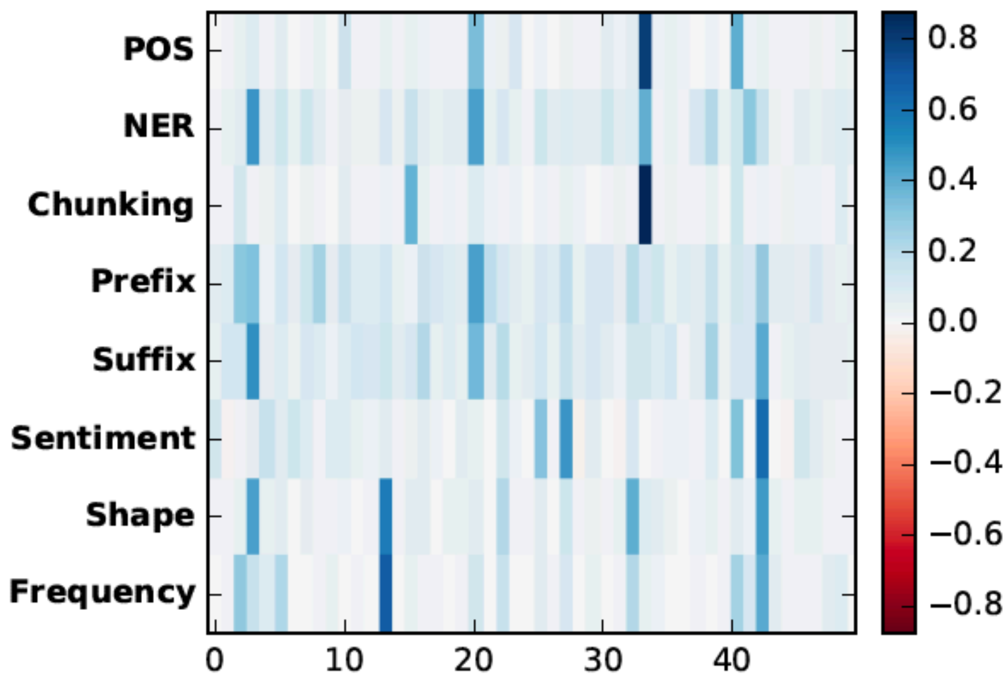
(Blue: correct prediction; green: incorrect)

Diagnostic classifier

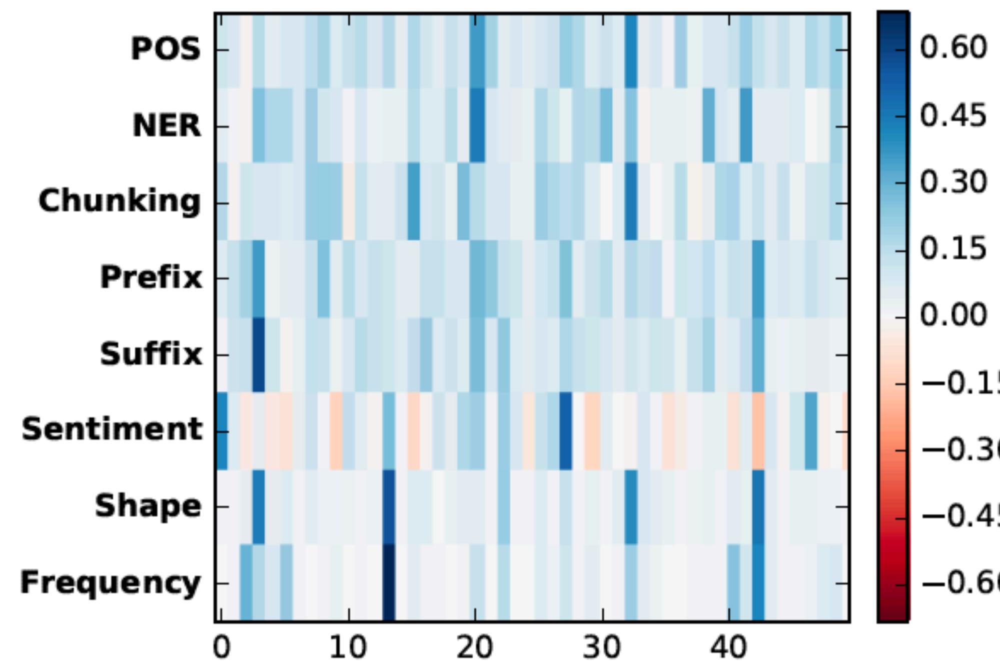


(Giullianeli et al., 2018, BlackboxNLP)

Erasure: how much does the classifier's prediction change if an input dimension is set to 0?



(a) Word2vec, no dropout.



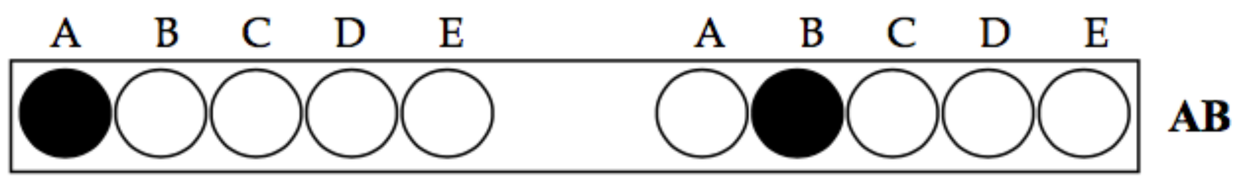
(b) Word2vec, with dropout.

(Related to ablation of a hidden unit!)

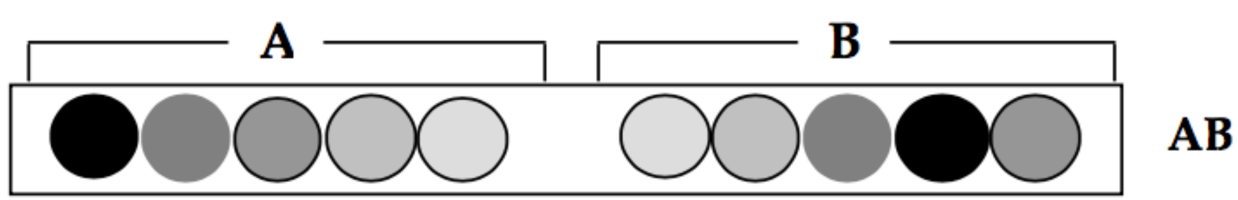
(Li et al., 2016, arXiv)

How do we represent discrete inputs and outputs in a network?

Localist ("one hot") representation: each unit represents an item (e.g., a word)



Distributed representation: each item is represented by multiple units, and each unit participates in representing multiple items



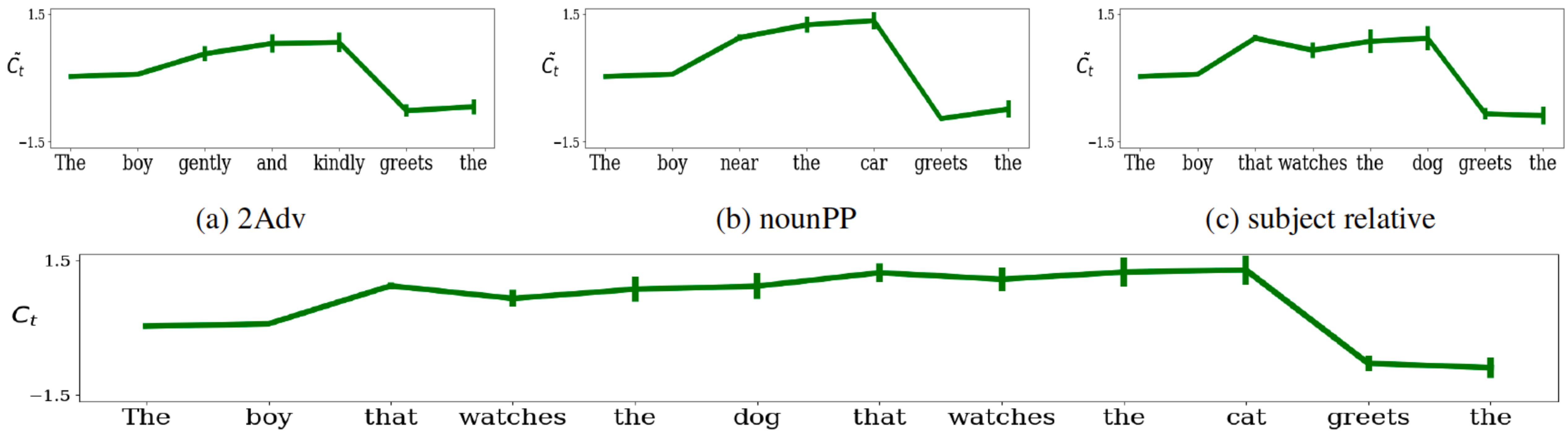
How localist are LSTM LM representations? (Ablation study)

Simple	the **boy greets** the guy
Adv	the **boy** probably **greets** the guy
2Adv	the **boy** most probably **greets** the guy
CoAdv	the **boy** openly and deliberately **greets** the guy
NamePP	the **boy** near Pat **greets** the guy
NounPP	the **boy** near the car **greets** the guy
NounPPAdv	the **boy** near the car kindly **greets** the guy

NA task	**C**	**Ablated**		**Full**
		776	**988**	
Simple	S	-	-	100
Adv	S	-	-	100
2Adv	S	-	-	99.9
CoAdv	S	-	82	98.7
namePP	SS	-	-	99.3
nounPP	SS	-	-	99.2
nounPP	SP	-	54.2	87.2
nounPPAdv	SS	-	-	99.5
nounPPAdv	SP	-	54.0	91.2
Simple	P	-	-	100
Adv	P	-	-	99.6
2Adv	P	-	-	99.3
CoAdv	P	79.2	-	99.3
namePP	PS	39.9	-	68.9
nounPP	PS	48.0	-	92.0
nounPP	PP	78.3	-	99.0
nounPPAdv	PS	63.7	-	99.2
nounPPAdv	PP	-	-	99.8
Linzen	**-**	75.3	-	93.9

(Lakretz et al., 2019, NAACL)

How localist are LSTM LM representations? (Single-unit recording)



(a) 2Adv

(b) nounPP

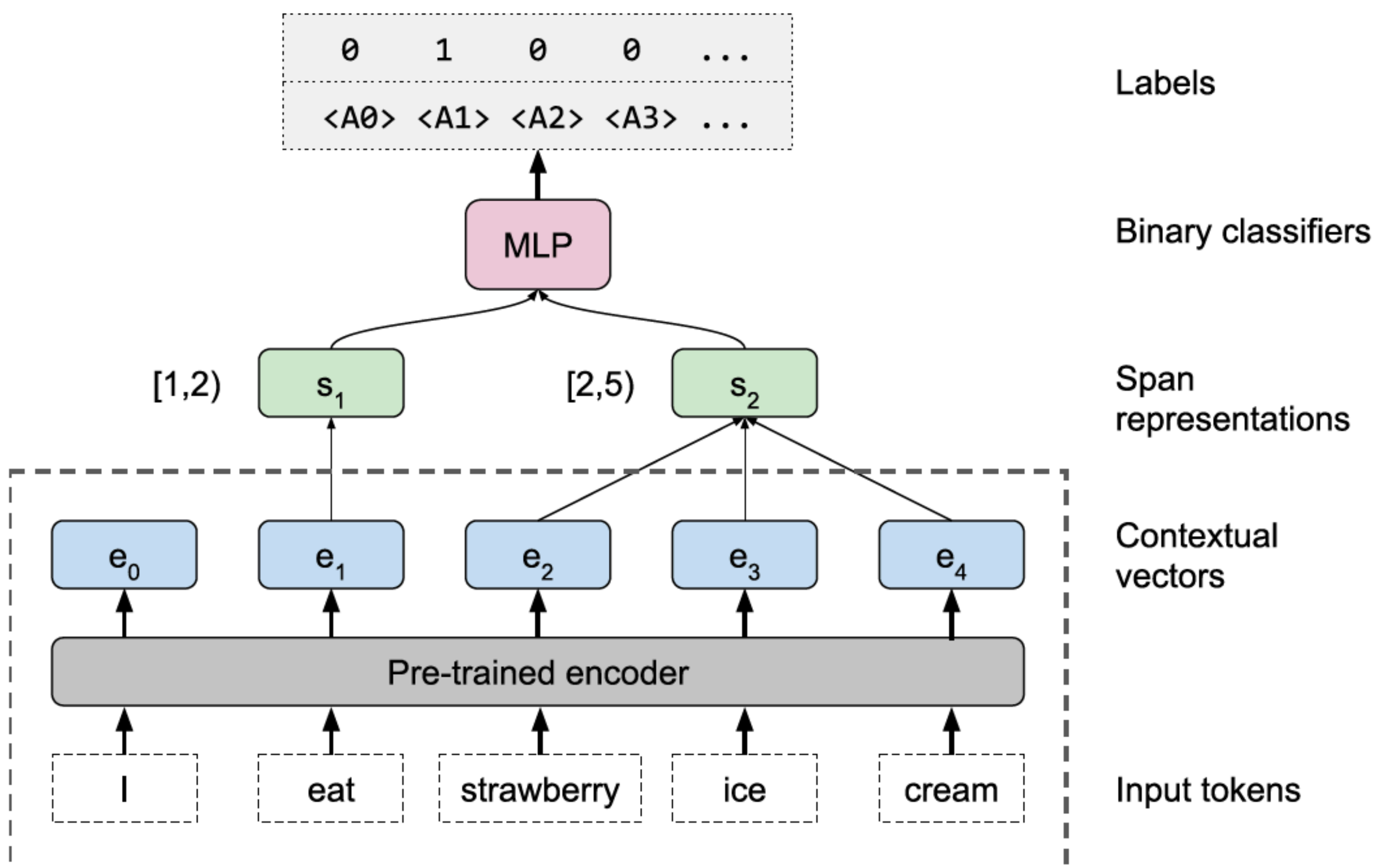
(c) subject relative

(d) Two embeddings with subject relatives

(Lakretz et al., 2019, NAACL)

Edge probing

Constit.	The important thing about Disney is that it [is a global brand]$_1$. → VP (Verb Phrase)
Depend.	[Atmosphere]$_1$ is always [fun]$_2$ → nsubj (nominal subject)
Entities	The important thing about [Disney]$_1$ is that it is a global brand. → Organization
SRL	[The important thing about Disney]$_2$ [is]$_1$ that it is a global brand. → Arg1 (Agent)



(Tenney et al., 2019, ICLR)

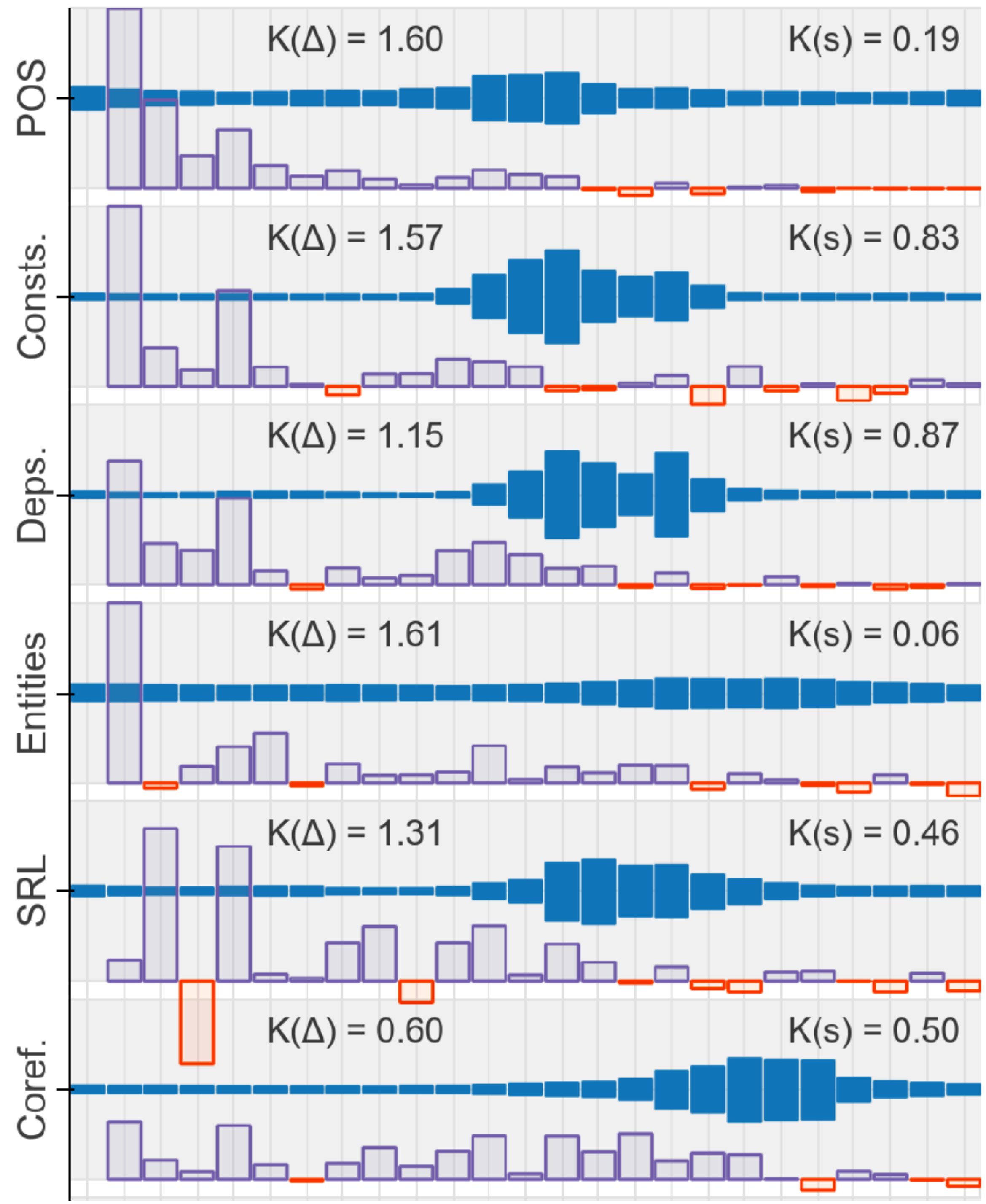
Edge probing

	Lex.	CNN1	CNN2	Ortho.	Full
Part-of-Speech	90	96	96	91	97
Constituents	69	84	85	72	85
Dependencies	80	91	92	85	94
Entities	92	94	94	93	96
SRL (all)	74	85	86	78	90
SRL (core)	74	87	89	79	93
SRL (non-core)	75	80	81	77	84
OntoNotes Coref.	75	80	80	80	84
SPR1	80	81	81	81	85
SPR2	82	82	82	83	83
Winograd Coref.	55	53	52	59	52

ELMo edge probing improves over baselines in syntactic tasks, not so much in semantic tasks

(Tenney et al., 2019, ICLR)

Layer-incremental edge probing on BERT

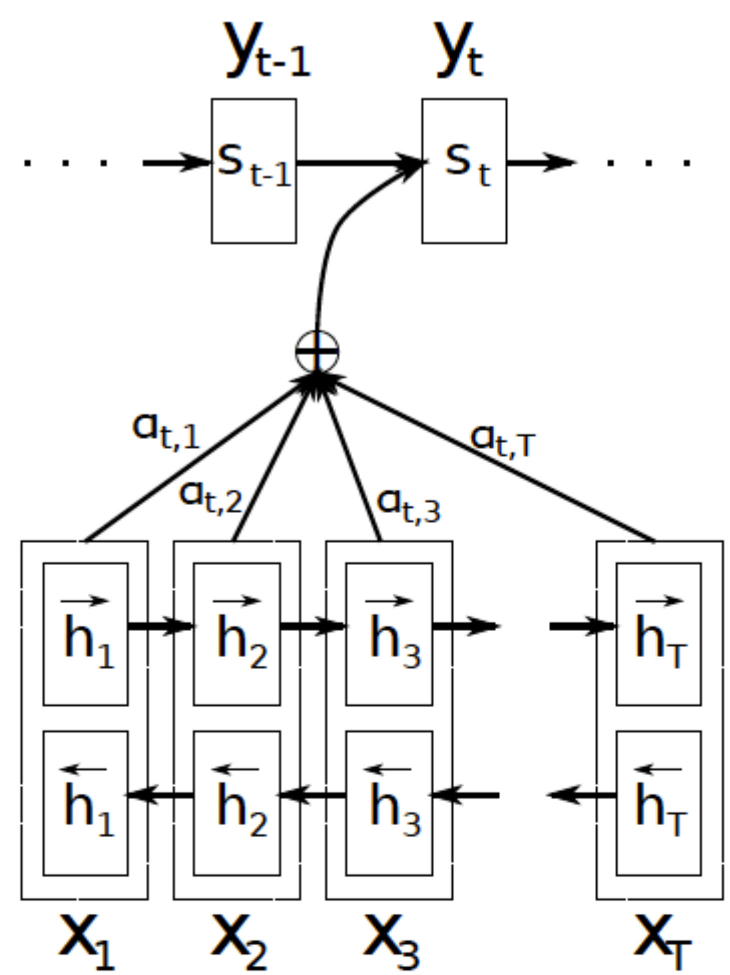


(Tenney et al., 2019, ACL)

Outline

- Characterizing what the network learned using behavioral experiments (“psycholinguistics on neural networks”)
- What information is encoded in intermediate vectors? (“artificial neuroscience”)
- **Interpreting attention heads**
- Symbolic approximations of neural networks

"Attention"



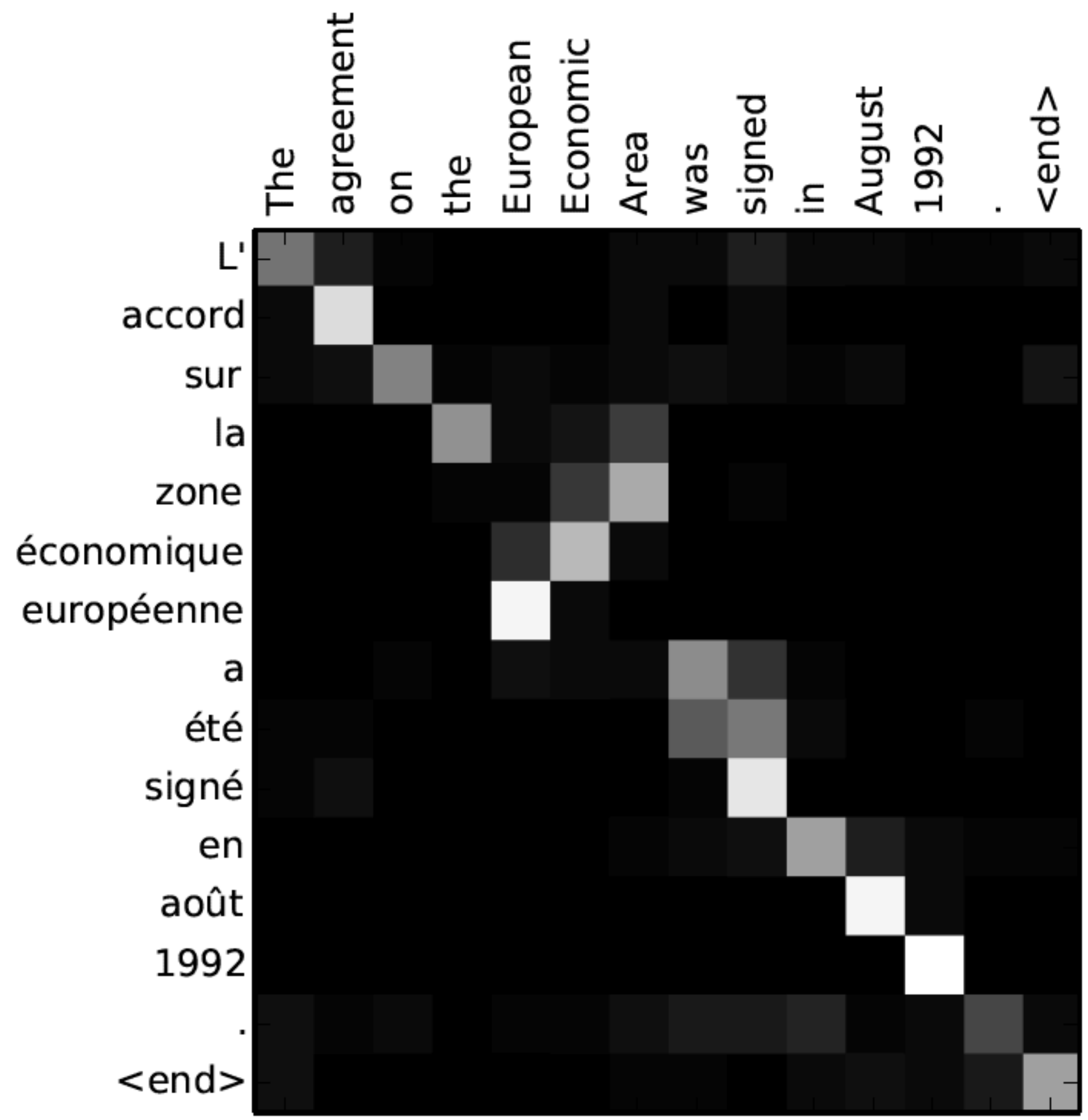
(Bahdanau et al., 2015, ICLR)

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

$$e_{ij} = a(s_{i-1}, h_j)$$

Can we use the attention weights to determine which n-th layer representation the model cares about in layer n+1?

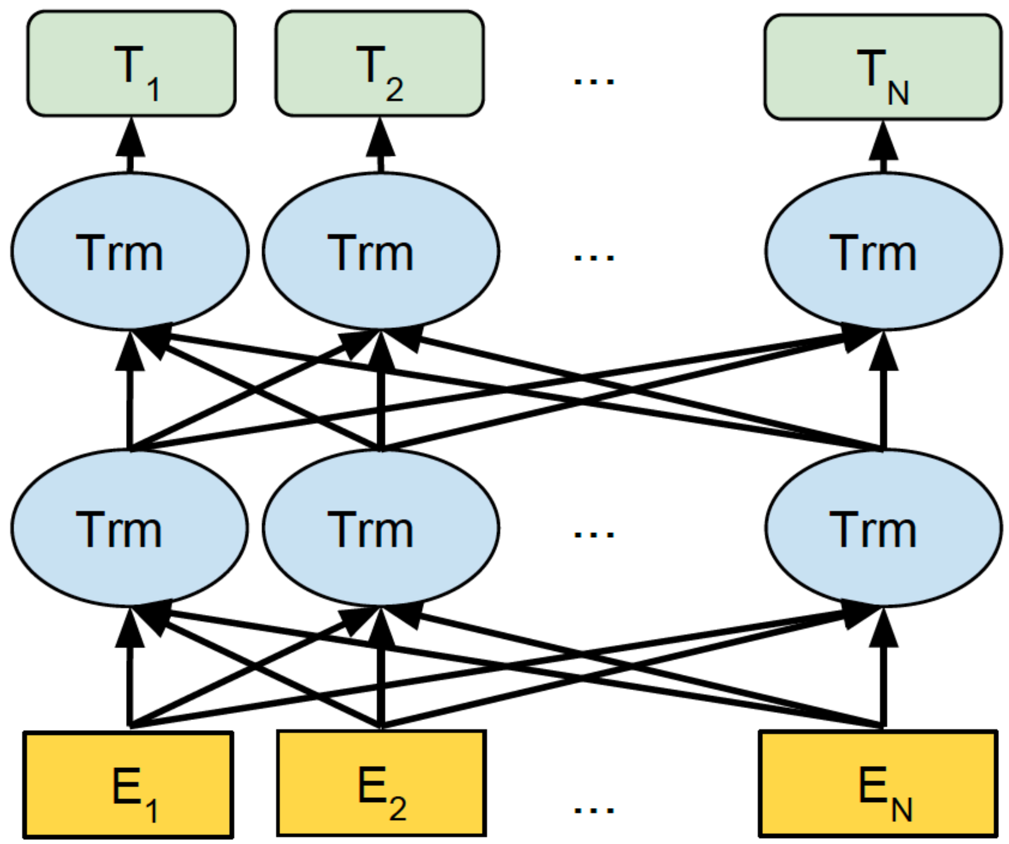
Attention as MT alignment



Caveat: an RNN's n-th hidden state is a compressed representation of the first n-1 words

<u>**(Bahdanau et al., 2015, ICLR)**</u>

Self-attention (e.g. BERT)

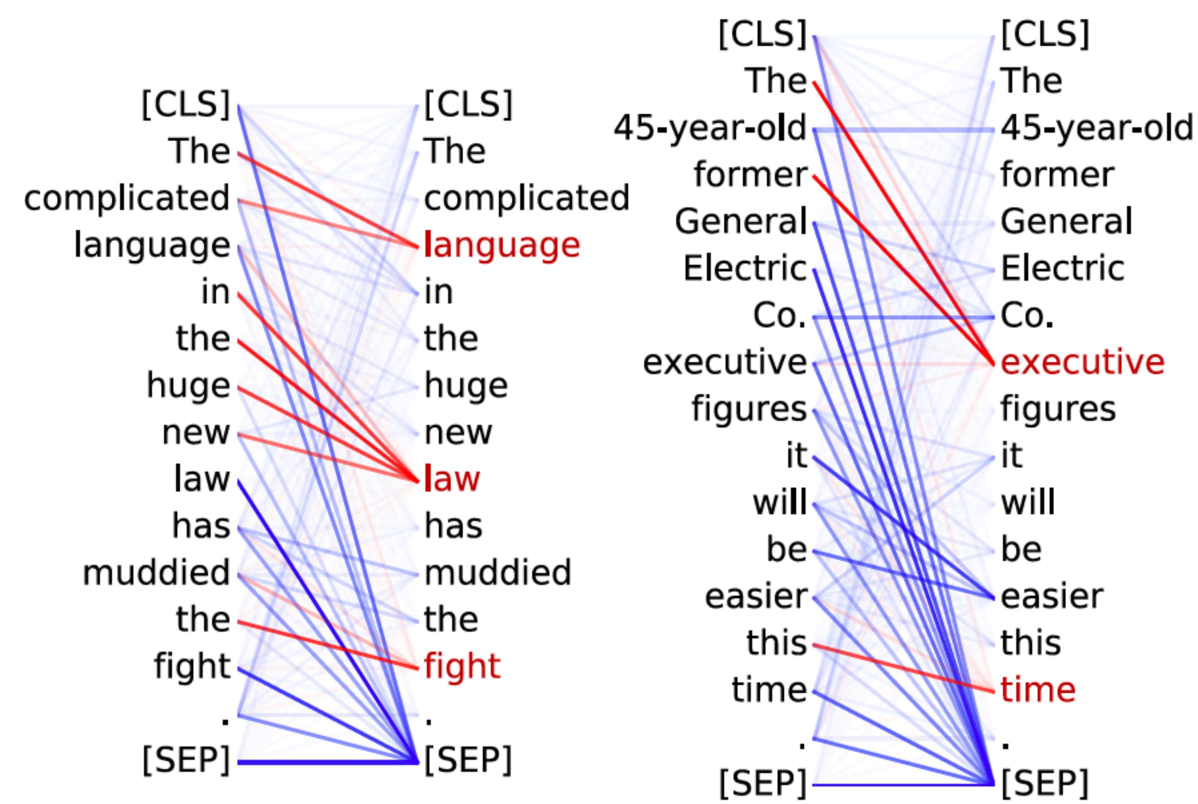


Syntactically interpretable self-attention heads (in BERT)

Relation	Head	Accuracy	Baseline
All	7-6	34.5	26.3 (1)
prep	7-4	66.7	61.8 (-1)
pobj	9-6	**76.3**	34.6 (-2)
det	8-11	**94.3**	51.7 (1)
nn	4-10	70.4	70.2 (1)
nsubj	8-2	58.5	45.5 (1)
amod	4-10	75.6	68.3 (1)
dobj	8-10	**86.8**	40.0 (-2)
advmod	7-6	48.8	40.2 (1)
aux	4-10	81.1	71.5 (1)

Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the `det` relation



(Clark et al., 2019, BlackboxNLP)

Is attention explanation?

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Attention correlates only weakly with other importance metrics (feature erasure, gradients)!

https://www.aclweb.org/anthology/N19-1357/

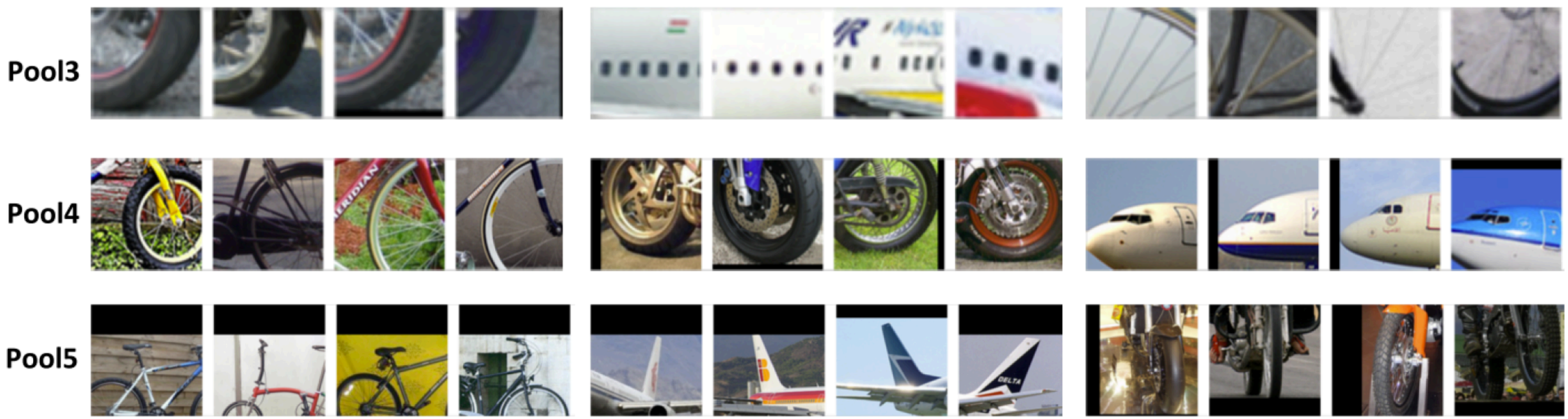
Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

https://www.aclweb.org/anthology/D19-1002/

A general word of caution



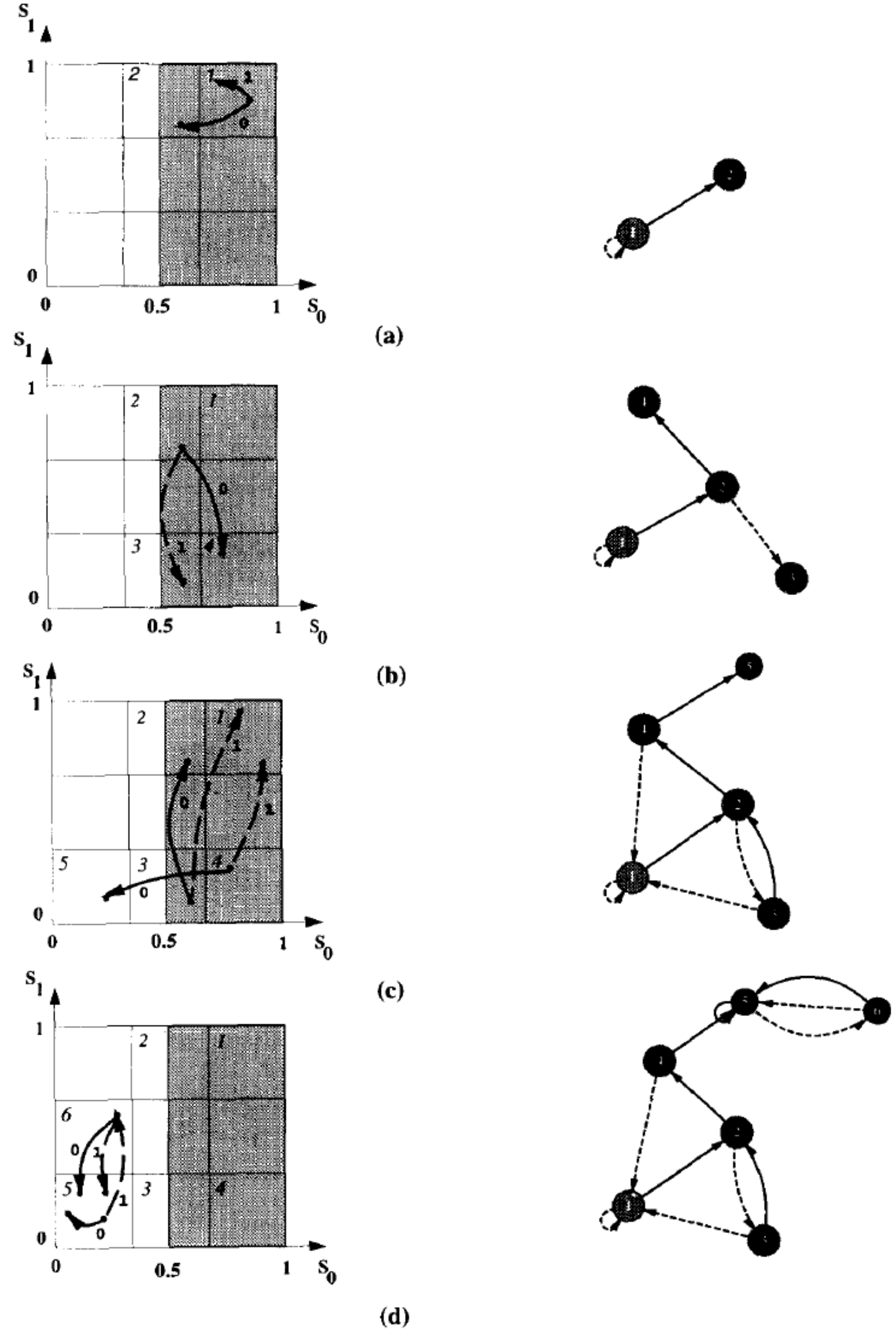
(Wang et al., 2015)

“However, such verbal interpretations may overstate the degree of categoricality and localization, and understate the statistical and distributed nature of these representations” (Kriegeskorte 2015)

Outline

- Characterizing what the network learned using behavioral experiments (“psycholinguistics on neural networks”)
- What information is encoded in intermediate vectors? (“artificial neuroscience”)
- Interpreting attention heads
- **Symbolic approximations of neural networks**

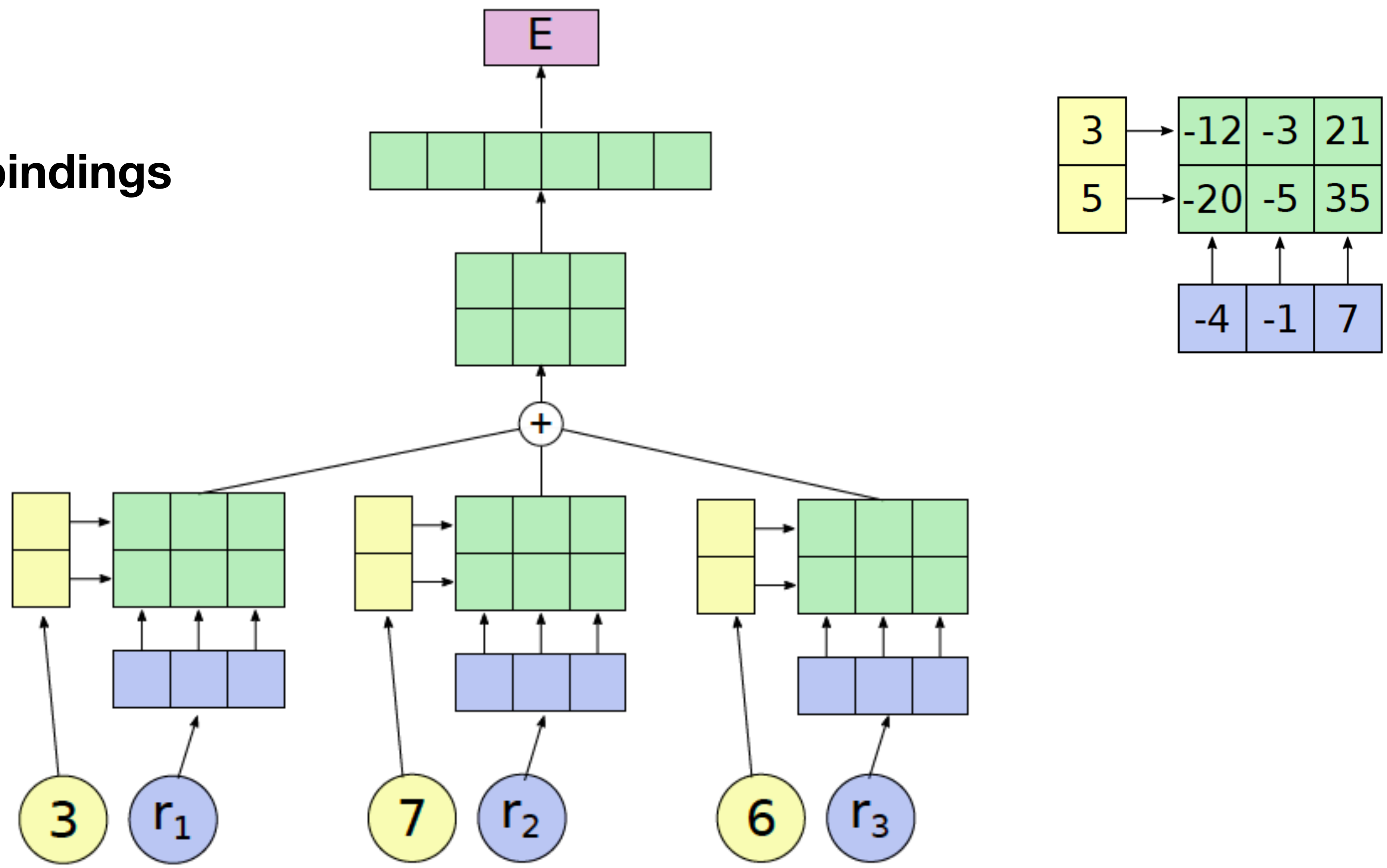
DFA extraction



(Omlin & Giles, 1996,
Weiss et al., 2018, ICML)

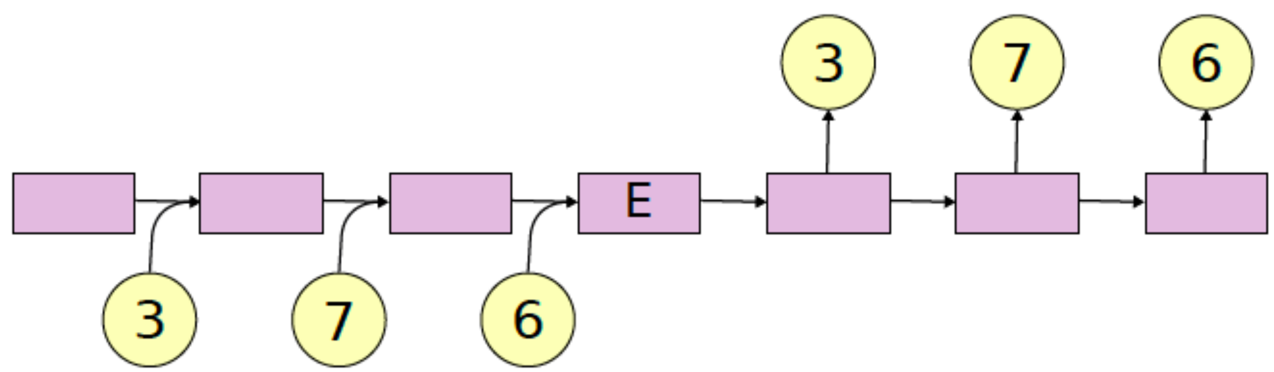
Method: Tensor Product Decomposition Networks

Sum of filler-role bindings



(McCoy, Linzen, Dunbar & Smolensky, 2019, ICLR)

Test case: sequence autoencoding



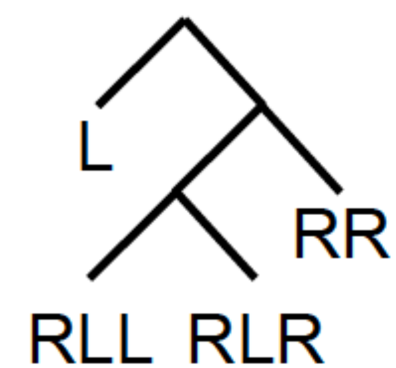
Hypothesis:

= 4:first + 2:second + 7:third + 9:fourth

Experimental setup: role schemes

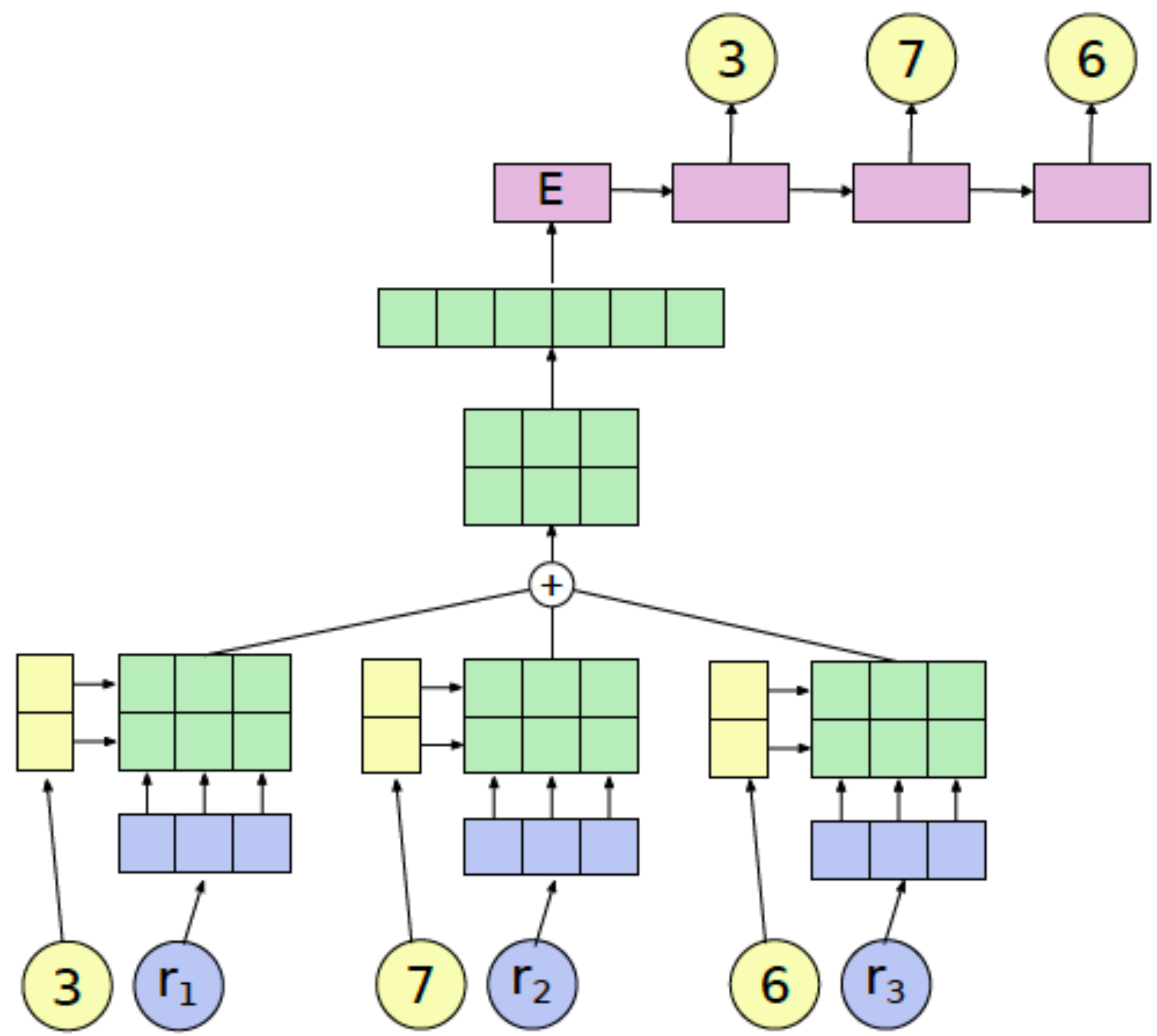
 = 4:first + 2:second + 7:third + 9:fourth

	3	1	1	6
Left-to-right	0	1	2	3
Right-to-left	3	2	1	0
Bidirectional	(0, 3)	(1, 2)	(2, 1)	(3, 0)
Wickelroles	#_1	3_1	1_6	1_#
Tree	L	RLL	RLR	RR
Bag of words	r_0	r_0	r_0	r_0

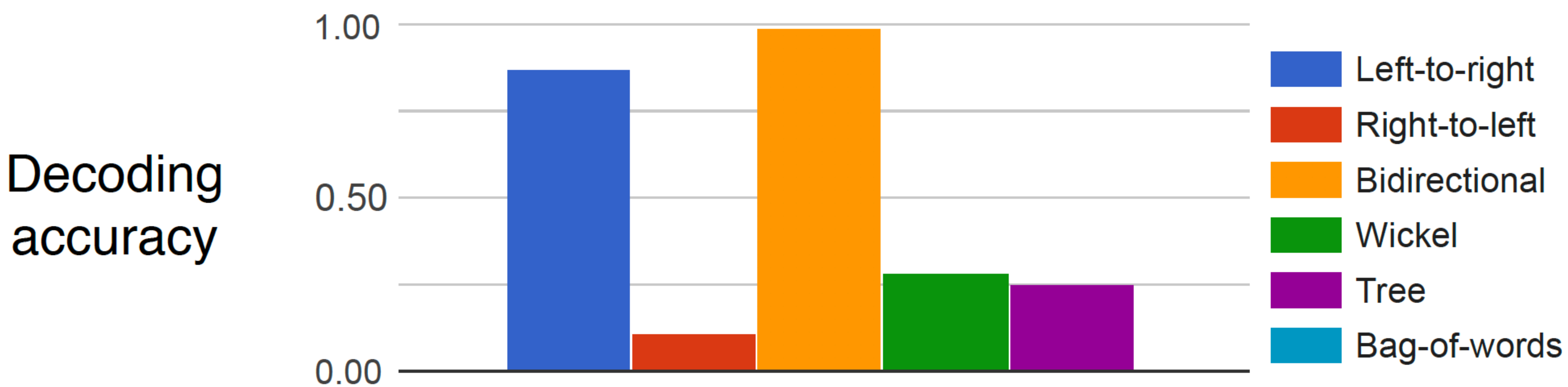


Tree roles

Evaluation: substitution accuracy

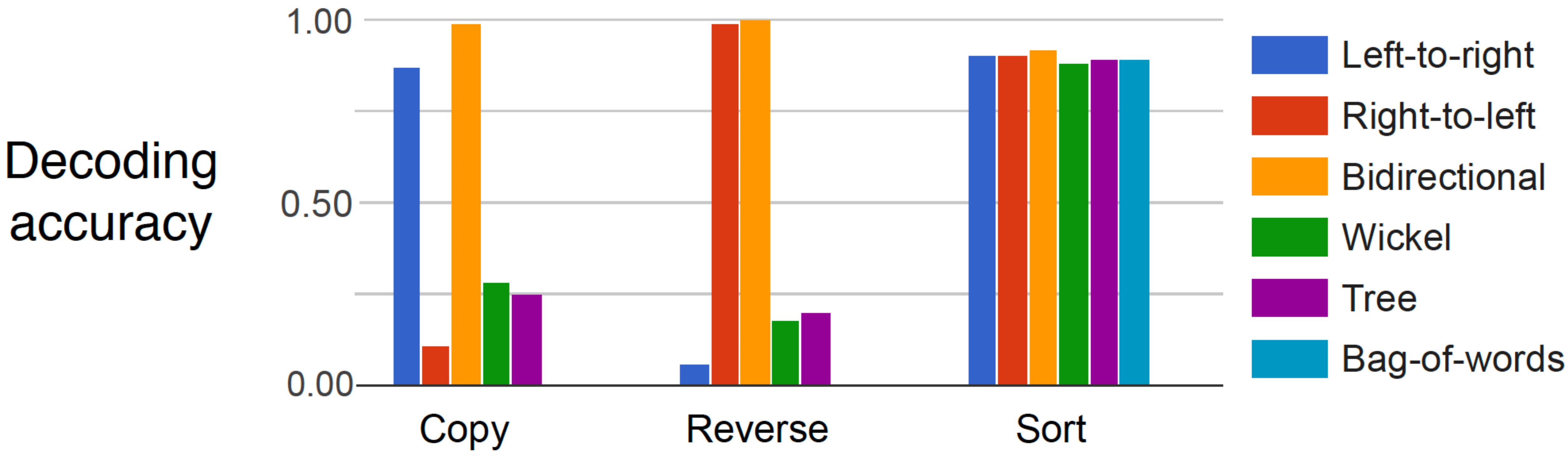


RNN autoencoders can be approximated almost perfectly



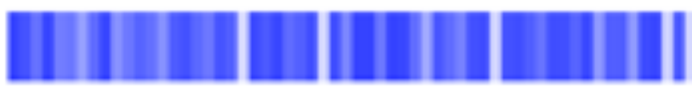
(McCoy, Linzen, Dunbar & Smolensky, 2019, ICLR)

Different tasks favor different role schemes



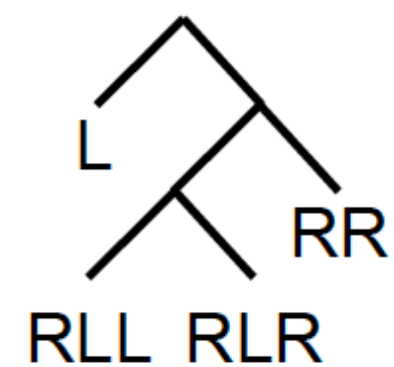
(McCoy, Linzen, Dunbar & Smolensky, 2019, ICLR)

This experiment required assuming a particular role scheme



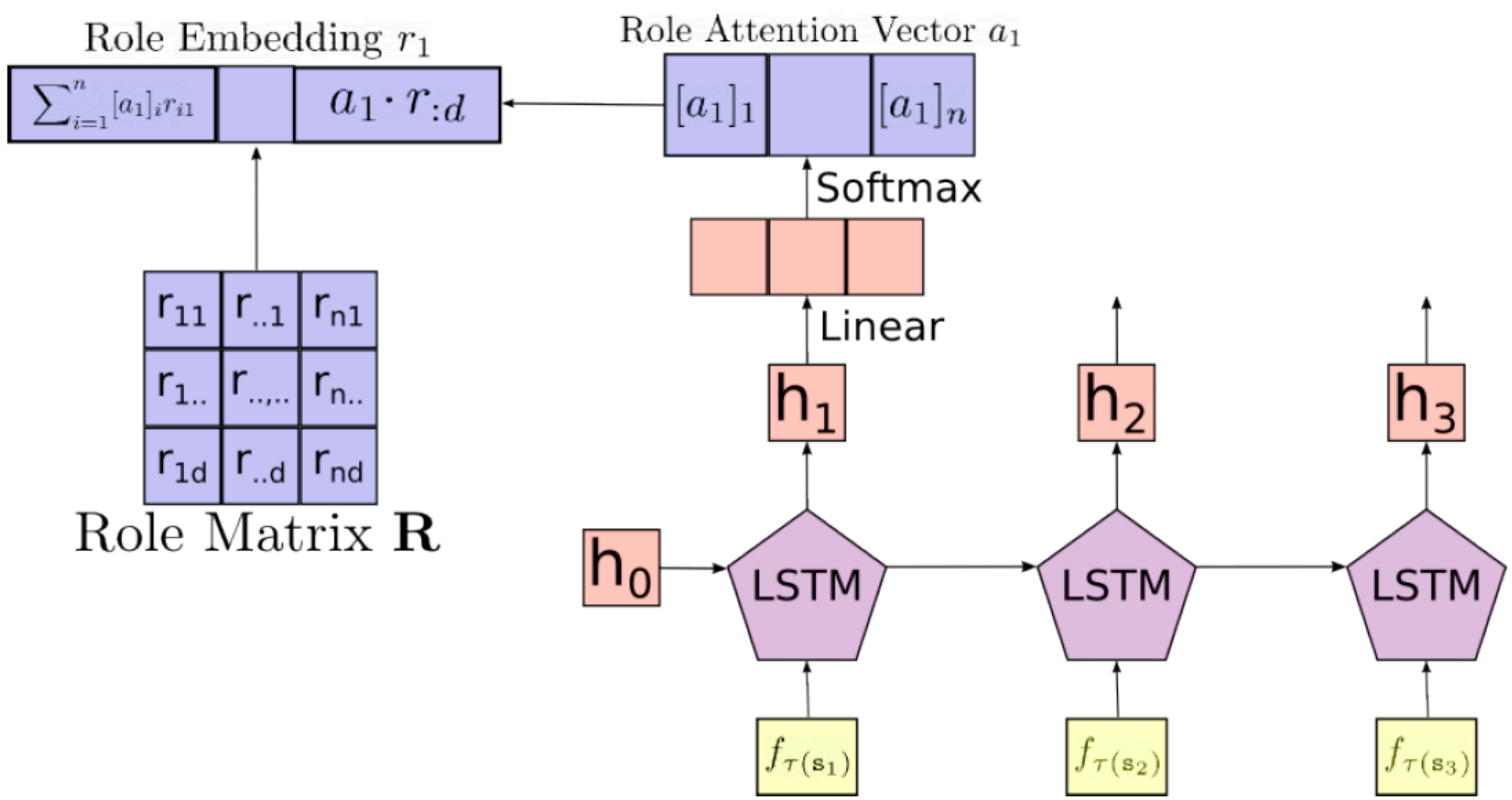
= 4:first + 2:second + 7:third + 9:fourth

	3	1	1	6
Left-to-right	0	1	2	3
Right-to-left	3	2	1	0
Bidirectional	(0, 3)	(1, 2)	(2, 1)	(3, 0)
Wickelroles	#_1	3_1	1_6	1_#
Tree	L	RLL	RLR	RR
Bag of words	r_0	r_0	r_0	r_0



Tree roles

Learning the role scheme



(Soulos, McCoy, Linzen & Smolensky, 2019)

Summary

- Symbolic approximations are currently successful only for synthetic data
- It is difficult to understand how massive end-to-end neural networks do what they're able to do, though the field has some ideas
- If interpretability and explainability are important:
  - Use networks that operate over human-interpretable symbolic structure
  - Use a pipeline approach with interpretable intermediate products