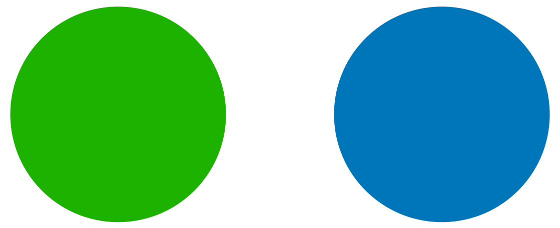
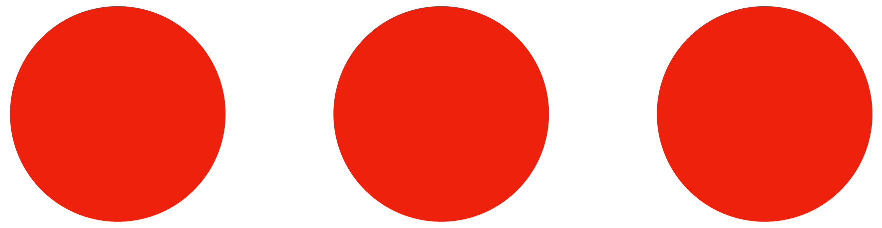
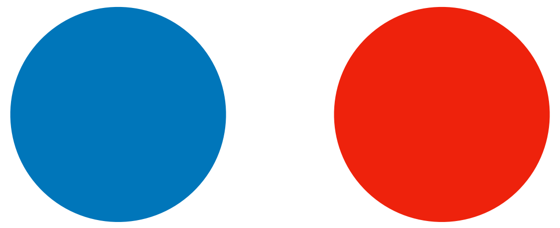
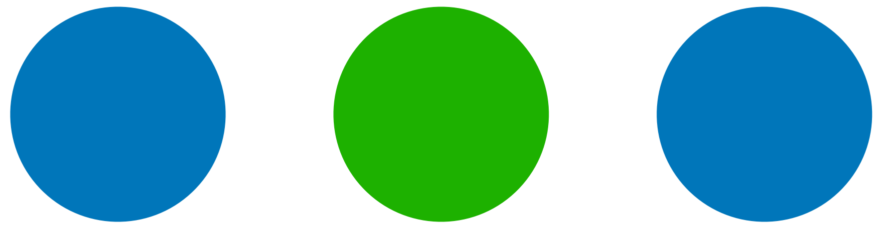
Neuro-symbolic NLP

Jacob Andreas / MIT 6.884 / Fall 2020

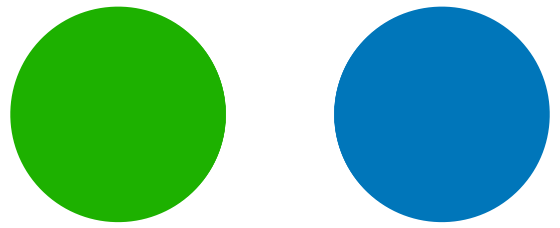
lug kiki wif 

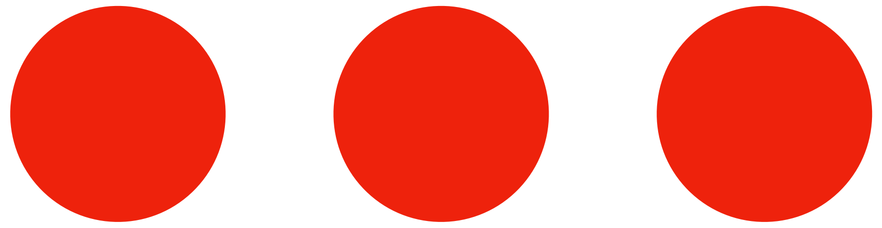
dax fep 

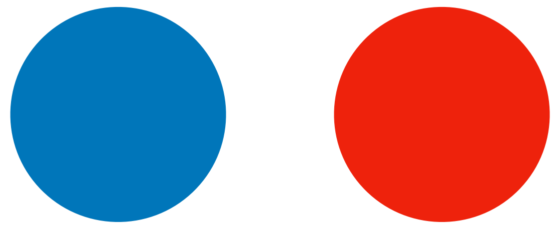
dax kiki lug 

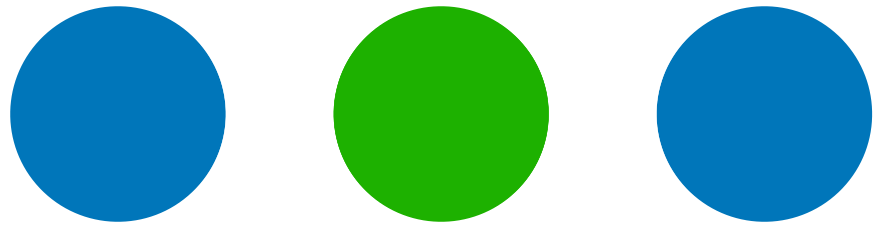
lug blicket wif 

[Lake, Linzen and Baroni 2020]

lug kiki wif 

dax fep 

dax kiki lug 

lug blicket wif 

wif fep

[Lake, Linzen and Baroni 2020]

lug kiki wif

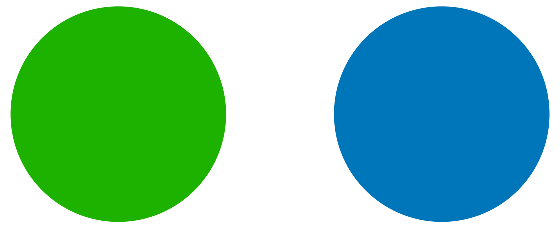
dax fep

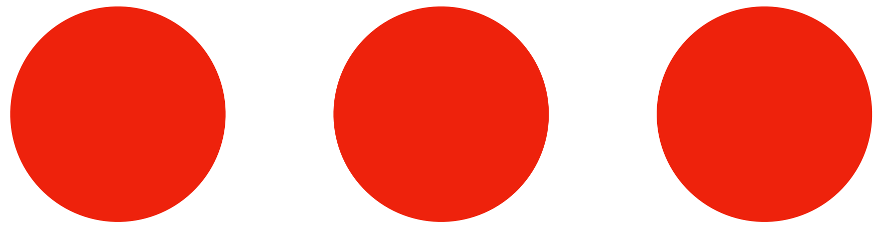
dax kiki lug

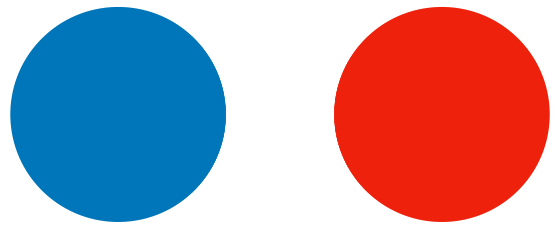
lug blicket wif

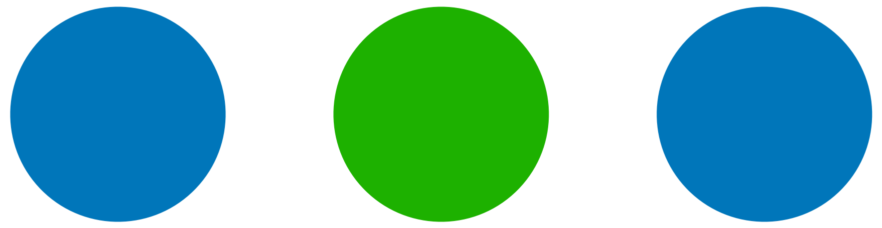
wif fep

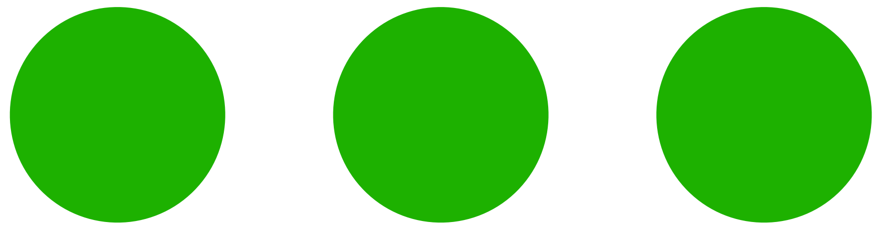
[Lake, Linzen and Baroni 2020]

lug kiki wif 

dax fep 

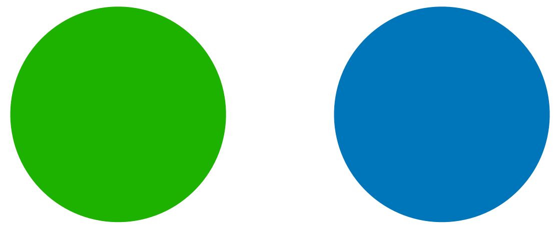
dax kiki lug 

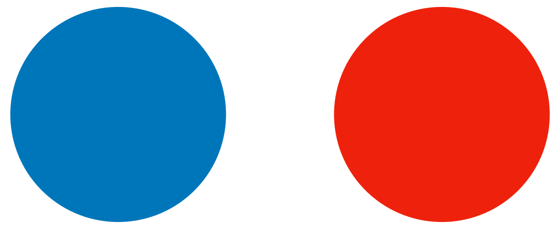
lug blicket wif 

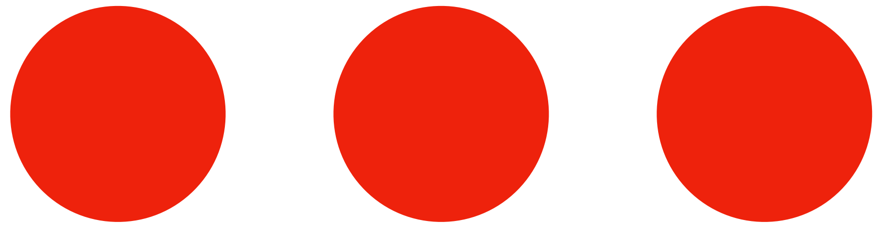
wif fep 

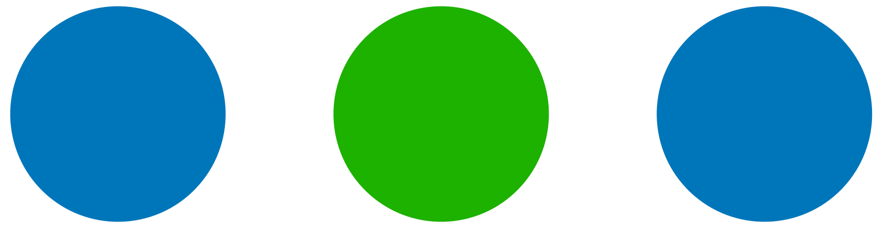
wif blicket lug kiki dax fep

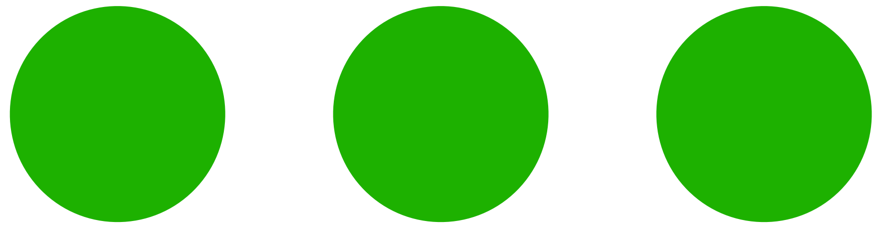
[Lake, Linzen and Baroni 2020]

lug kiki wif 

dax kiki lug 

dax fep 

lug blicket wif 

wif fep 

wif blicket lug kiki dax fep

[Lake, Linzen and Baroni 2020]

lug kiki wif

dax fep

dax kiki lug

lug blicket wif

wif fep

wif blicket lug kiki dax fep

[Lake, Linzen and Baroni 2020]

lug kiki wif

dax fep

dax kiki lug

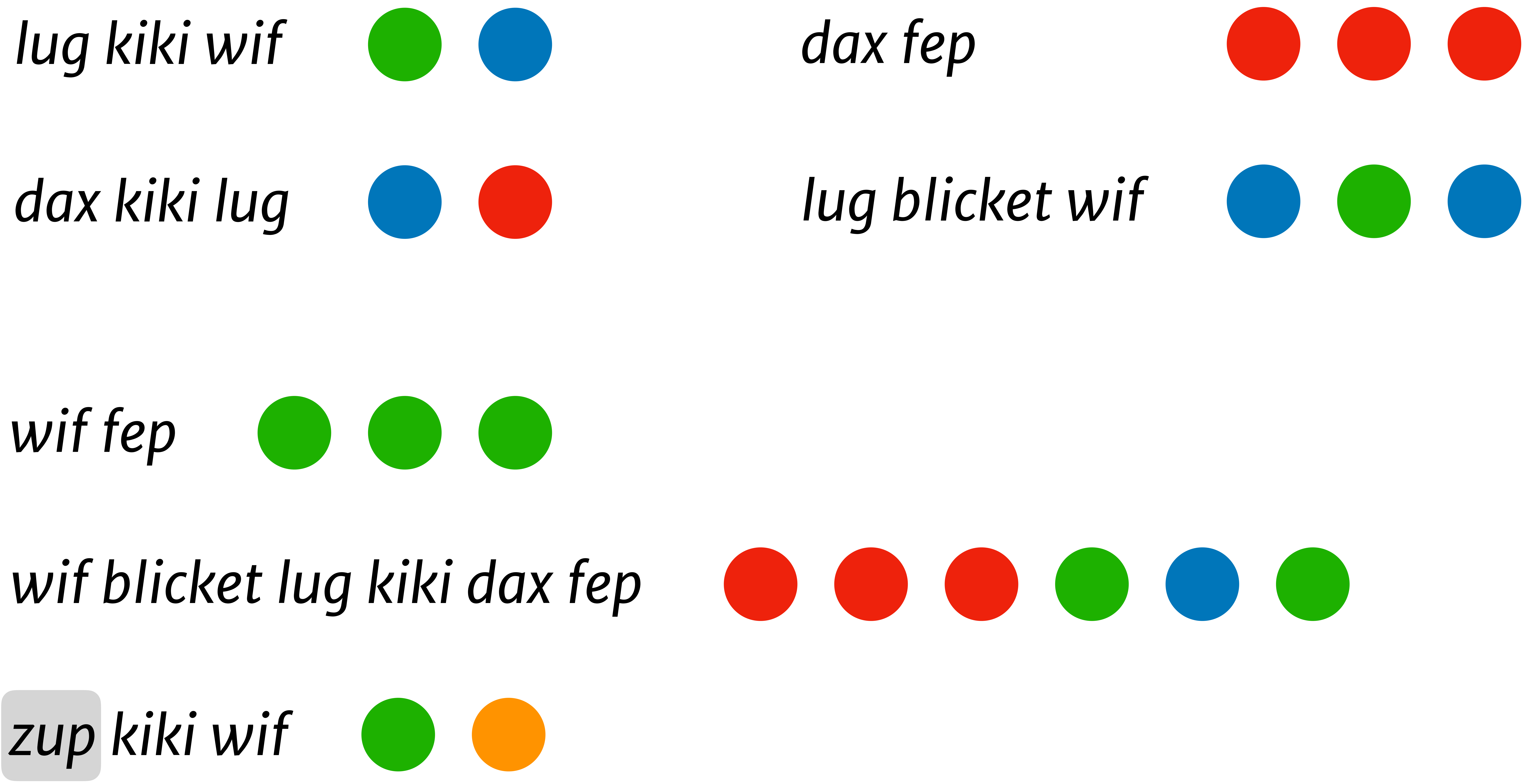
lug blicket wif

wif fep

wif blicket lug kiki dax fep

zup kiki wif

[Lake, Linzen and Baroni 2020]



[Lake, Linzen and Baroni 2020]

lug kiki wif

dax kiki lug

dax fep

lug blicket wif

wif fep

86% of crowd workers

wif blicket lug kiki dax fep

70%

zup kiki wif

~82%

[Lake, Linzen and Baroni 2020]

lug kiki wif

dax fep

dax kiki lug

lug blicket wif

wif fep 86% of turkers

wif blicket lug kiki dax fep 70%

zup kiki wif ~82%

Why?

[Lake, Linzen and Baroni 2020]

lug kiki wif

dax fep

dax kiki lug

lug blicket wif

wif fep 0% of randomly initialized RNNs

wif blicket lug kiki dax fep 0%

zup kiki wif 0%

[Lake, Linzen and Baroni 2020]

lug kiki wif

dax fep

dax kiki lug

lug blicket wif

wif fep 0% of randomly initialized RNNs

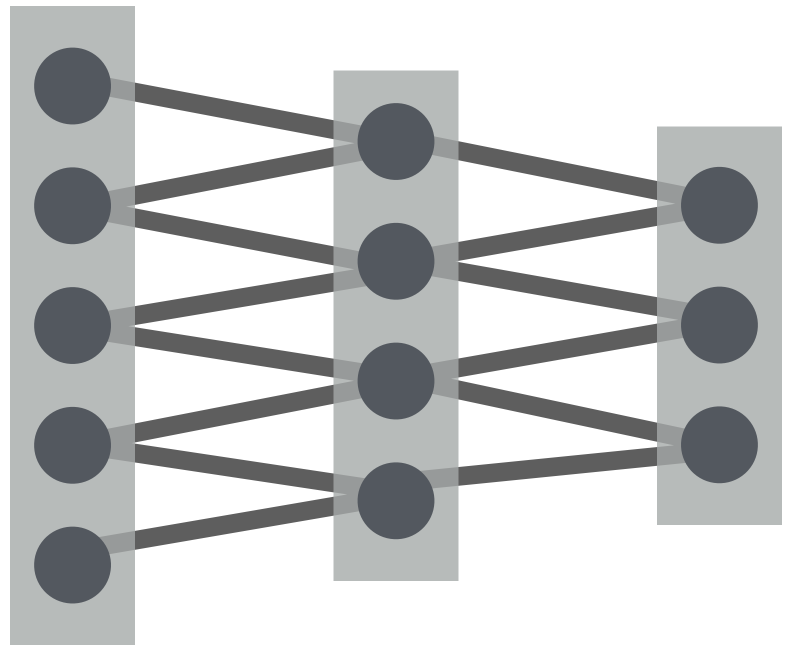
wif blicket lug kiki dax fep 0%

zup kiki wif 0%

Why?

[Lake, Linzen and Baroni 2020]

Successes of deep learning: vision



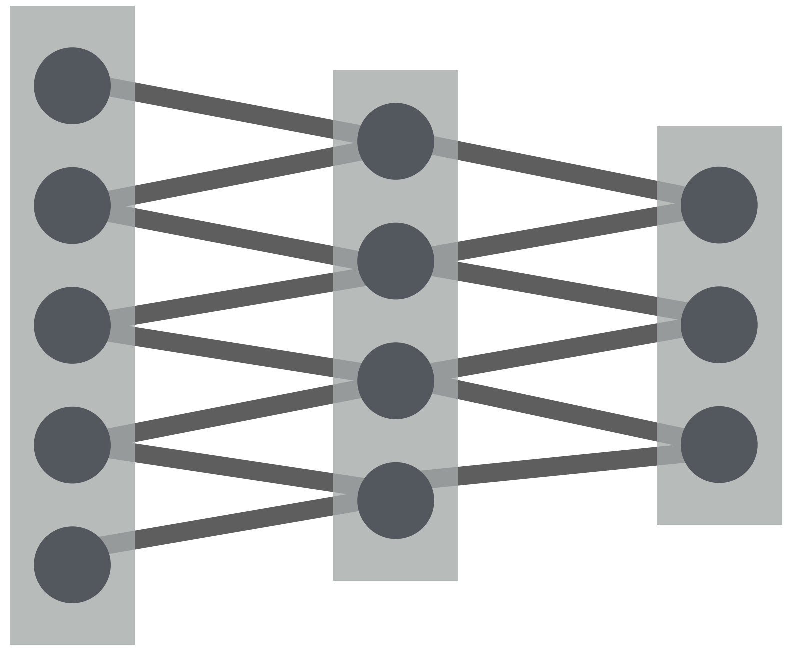
cucumber

telephone_pole

irish_setter

Successes of deep learning: NLP

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha much tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.

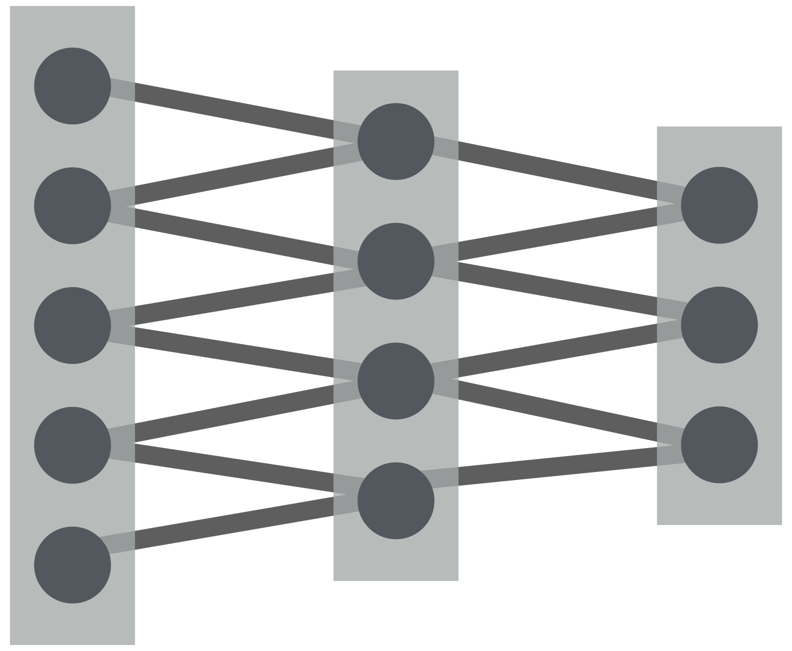


Somewhere in La Mancha, in a place whose name I do not care to remember, a gentleman lived not long ago, one of those who has a lance and ancient shield on a shelf and keeps a skinny nag and a greyhound for racing.

[Grossman 2005]

Successes of deep learning: NLP

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha much tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.



In a place in La Mancha, whose name I do not want to remember, it was not long ago that a nobleman of the shipyard spear, old shield, skinny nag and running greyhound lived.

[http://translate.google.com]

But…

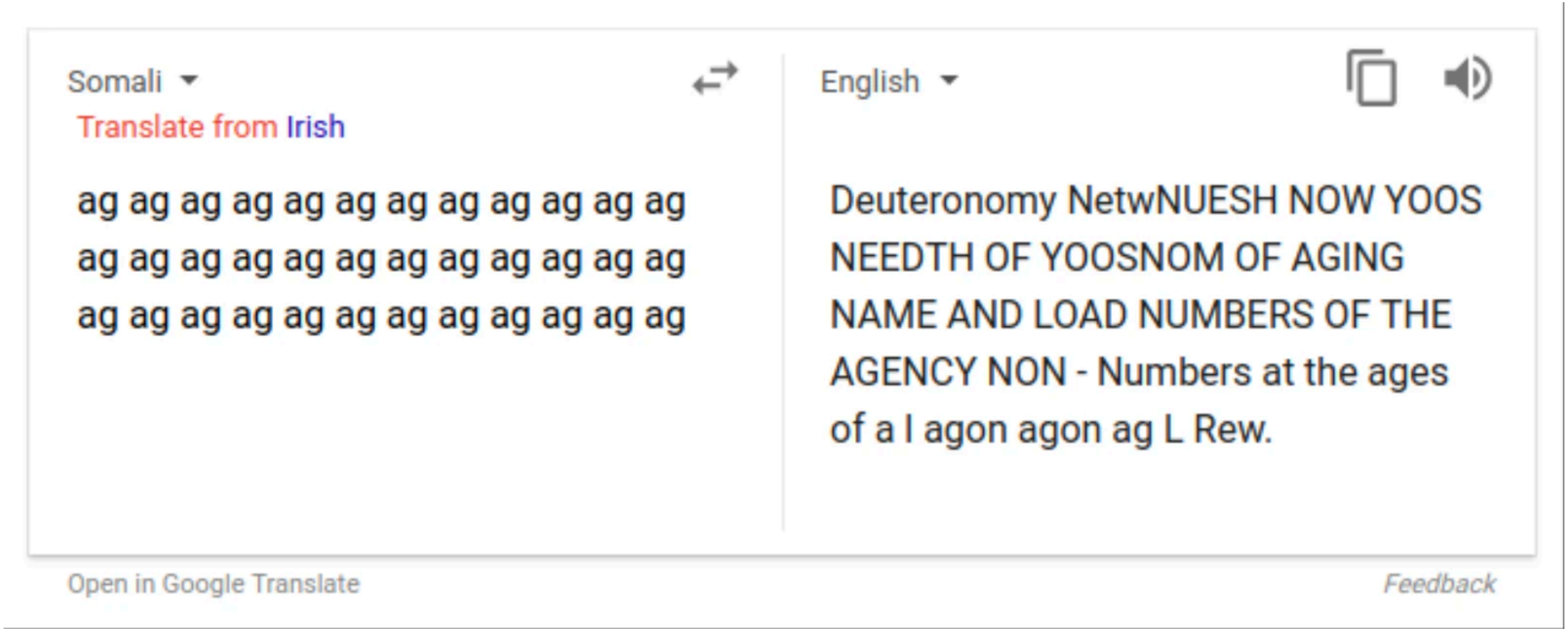
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently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

Q: What was the name of the quarterback who was 38 in Super Bowl XXXIII?

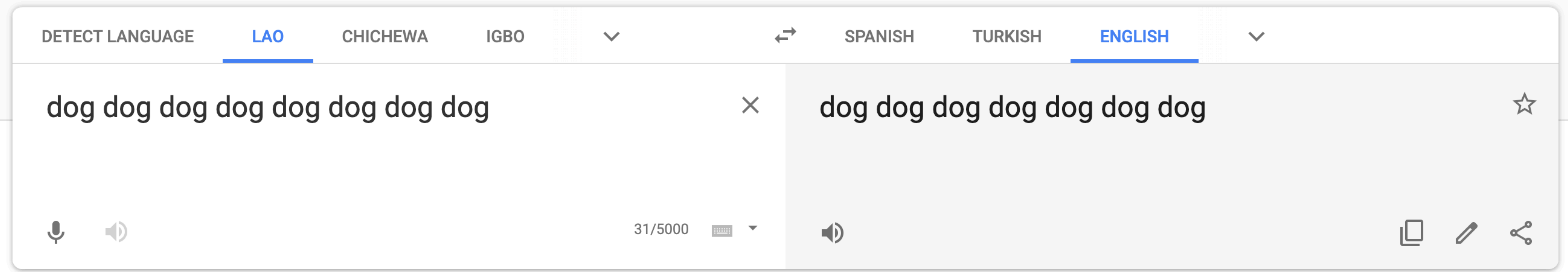
A: Jeff Dean.

[Jia & Liang 2017]

But…

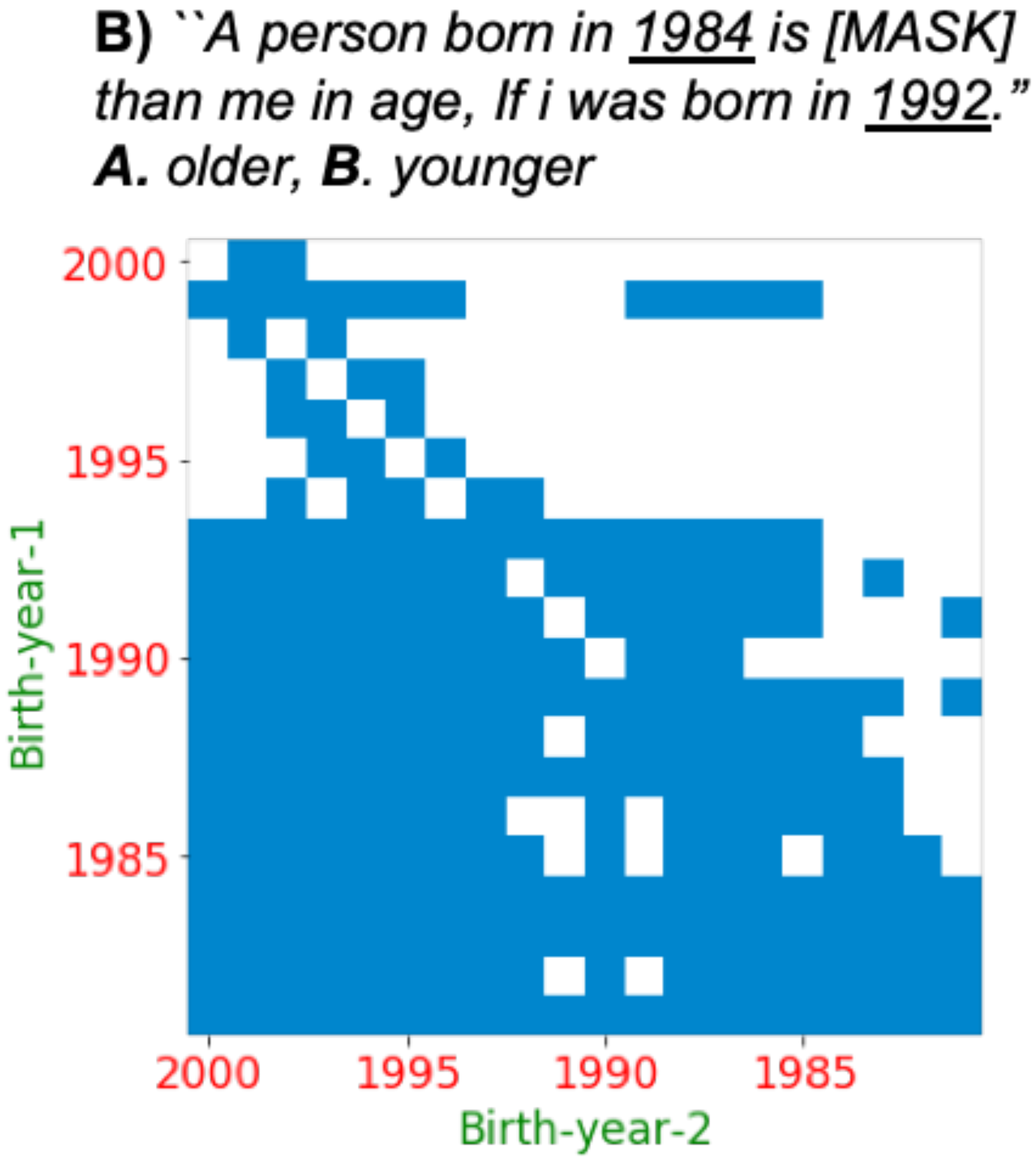

Somali
Translate from Irish
ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag
English
Deuteronomy NetwNUESH NOW YOOS NEEDTH OF YOOSNOM OF AGING NAME AND LOAD NUMBERS OF THE AGENCY NON - Numbers at the ages of a l agon agon ag L Rew.
Open in Google Translate
Feedback

But…



DETECT LANGUAGE
LAO
CHICHEWA
IGBO
SPANISH
TURKISH
ENGLISH
dog dog dog dog dog dog dog dog
dog dog dog dog dog dog dog
31/5000

But…



[Talmor et al. 2019]

But???

Q: If a b c changes to a b d , what does p q r change to?
A: p q s

Q: If a b c changes to a b d , what does i j k change to?

[Mitchell 2020]

But???

Q: If a b c changes to a b d , what does p q r change to?
A: p q s

Q: If a b c changes to a b d , what does i j k change to?
A: i j l

[Mitchell 2020]

But??????

Q: If a b c changes to a b d , what does p q r change to?
A: p q s

Q: If a b c changes to a b d , what does i j k change to?
A: i j l

Q: If a b c changes to a b d , what does p q r change to?
A: p q s

Q: If a b c changes to a b d , what does i j k l m change to?
A: i j k l n

Q: If a b c changes to a b d , what does r s t u v w change to?
A: r s t u x

[Mitchell 2020]

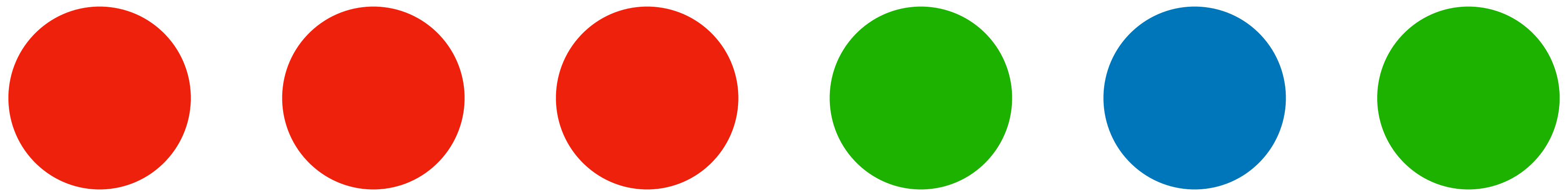
Symbolic NLP

What do these problems all have in common?

Symbolic NLP

What do these problems all have in common?

wif blicket lug kiki dax fep

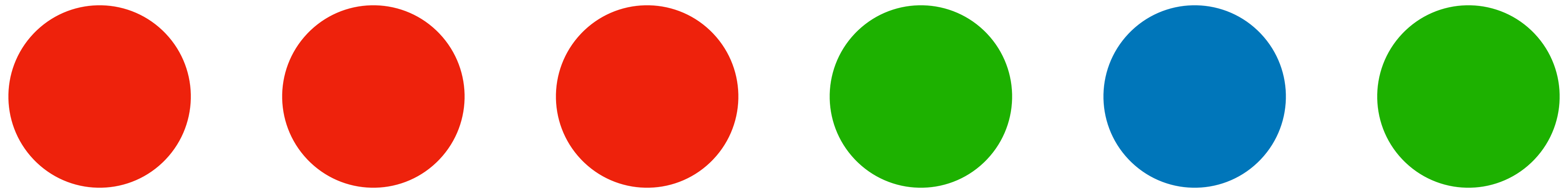


green around blue after red thrice

Symbolic NLP

What do these problems all have in common?

wif blicket lug kiki dax fep

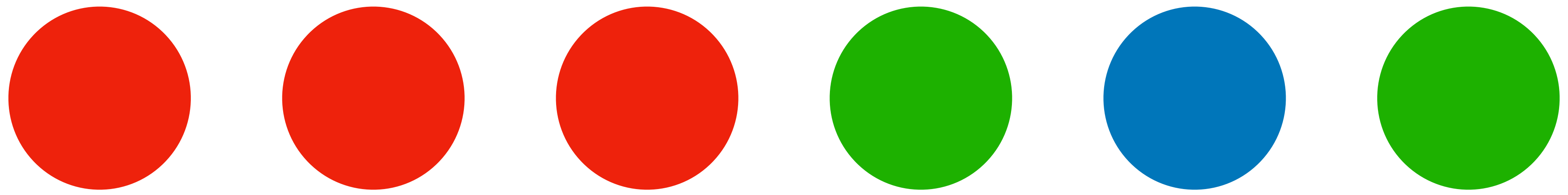


green around blue after red thrice

`after(around(green, blue), thrice(red))`

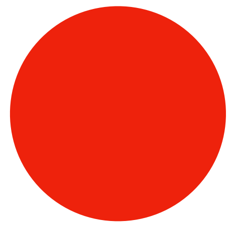
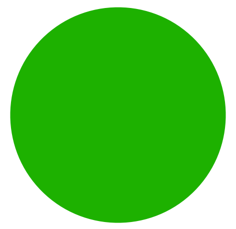
Symbolic NLP

What do these problems all have in common?

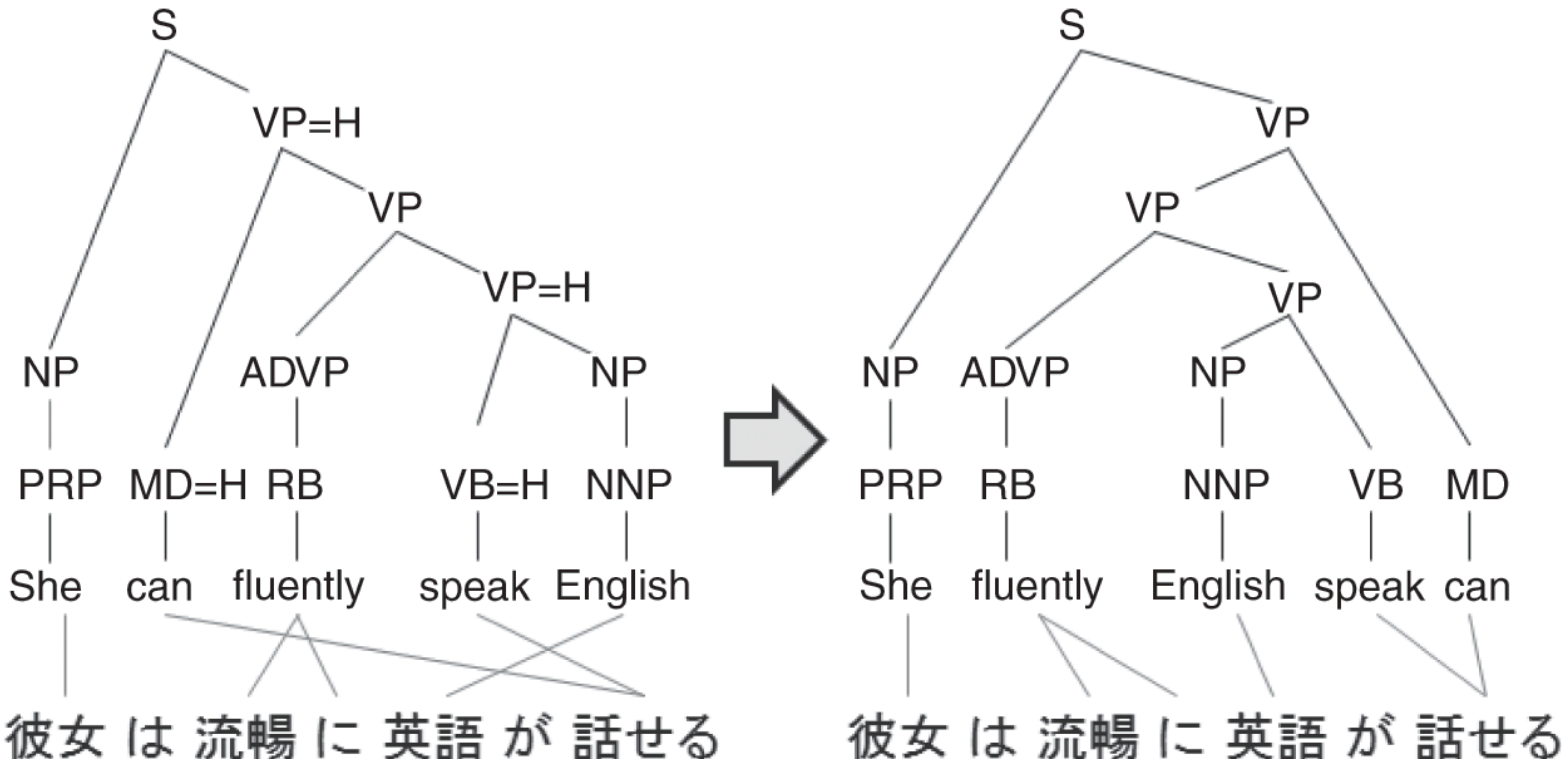
wif blicket lug kiki dax fep 

green around blue after red thrice

`after(around(green, blue), thrice(red))`

fep:  *wif:*  *blicket:* `(x, y)` $\mapsto$ `x y x`

Symbolic NLP



Symbolic NLP

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.*"

Q: What was the name of the quarterback who was 38 in Super Bowl XXXIII?

```
name(e1, John Elway)
type(e1, Person)
name(e2, Super Bowl XXXIII)
type(e2, Event)
role(e1, e2, Quarterback)
name(e3, Peyton Manning)...

---

name(x1, a)
role(x1, x2, Quarterback)
name(x2, Super Bowl XXXIII)
a?
```

Compositionality

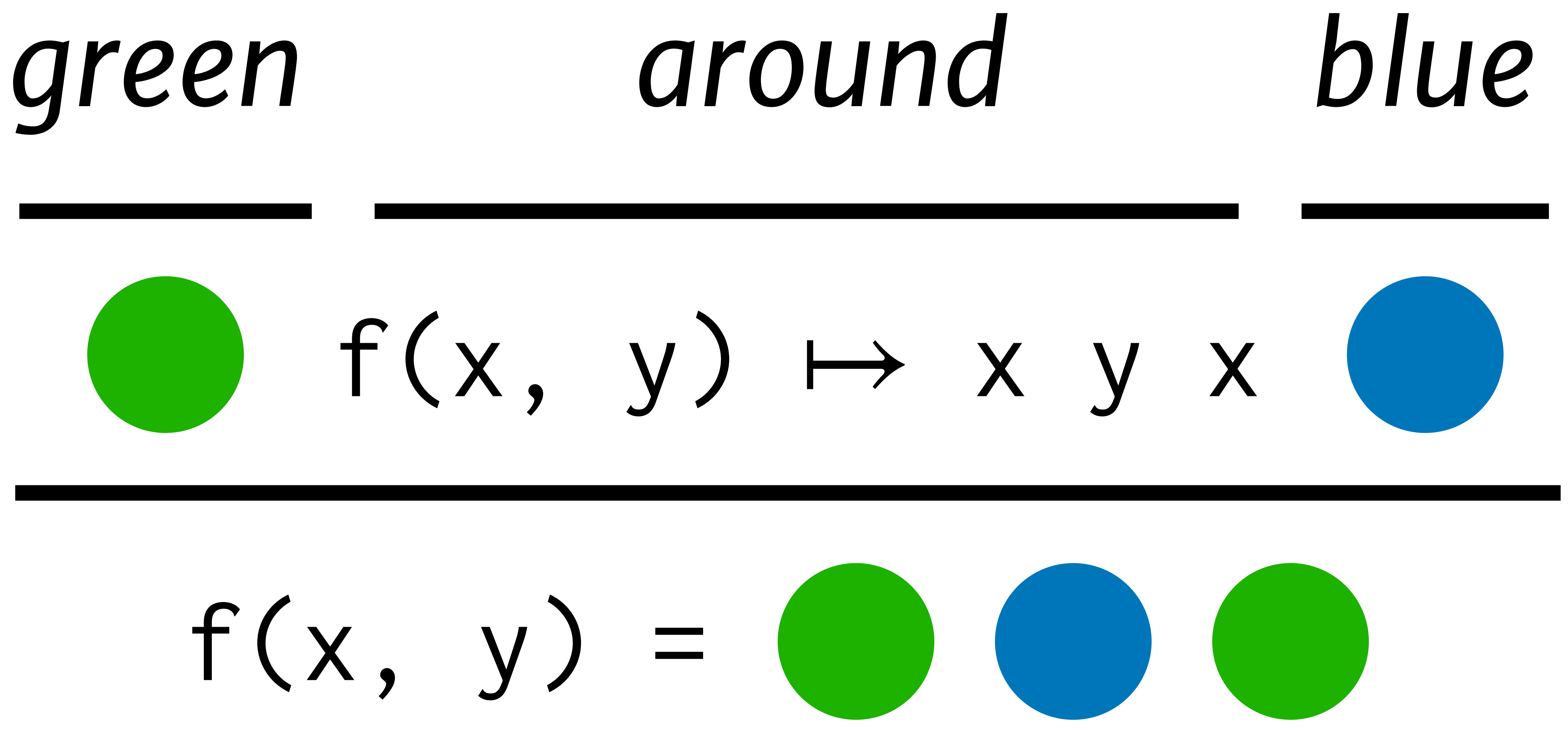
Sentence interpretation is a **homomorphism** from inputs to outputs.

green *around* *blue*

$f(x, y) \mapsto x\ y\ x$

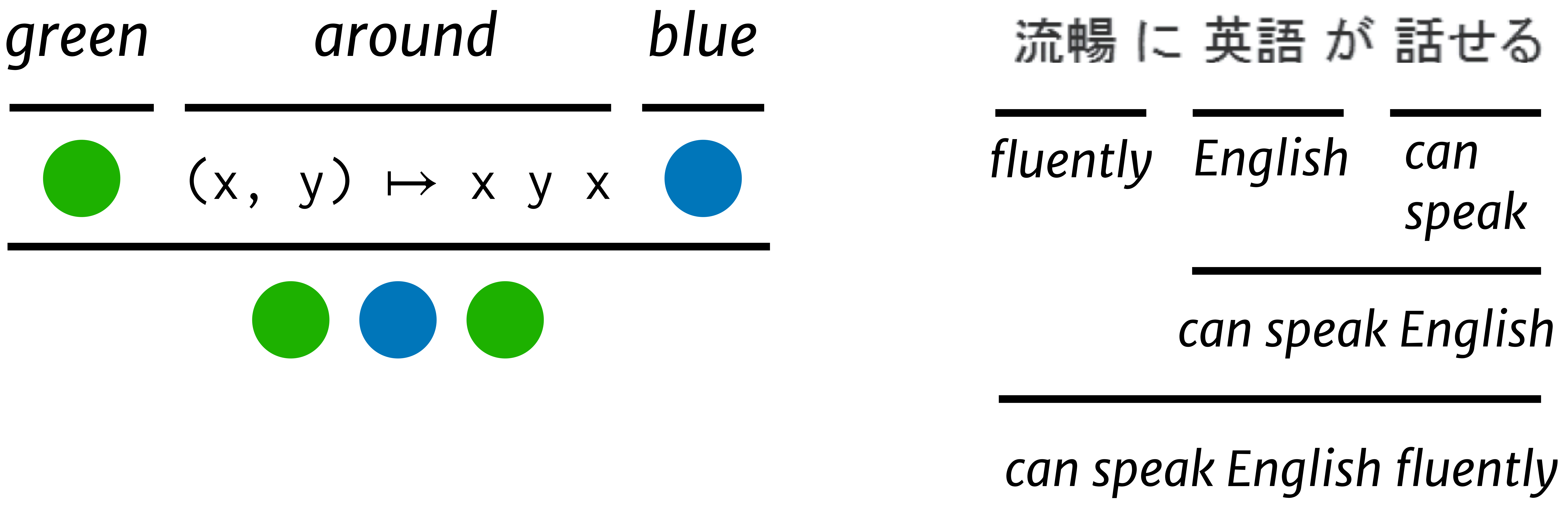
Compositionality

Sentence interpretation is a **homomorphism** from inputs to outputs.



Compositionality

Sentence interpretation is a **homomorphism** from inputs to outputs.



Symbol processing is about more than compositionality!

X → *X after X*, *X thrice*, *X around X*, *red*, *green*, …

green around blue after red thrice ✔

around around red after ✘

Three theories

<u>Symbol processing is correct as a **mechanistic** model of language.</u>

We'll never get human-level NLP without explicit symbols in our models.

Three theories

<u>Symbol processing is correct as a **descriptive** model of language.</u>

We may not need them at the implementation level, but symbolic models are useful for characterizing the kinds of data distributions and generalizations that matter.

Three theories

Symbol processing is the **wrong model.**

Real languages (and other human representational systems) are too messy and have too many exceptions to admit a useful symbolic description at any level of representation.

Cultures in artificial intelligence research

```
name(e1, John Elway)
type(e1, Person)
name(e2, Super Bowl XXXIII)
type(e2, Event)
role(e1, e2, Quarterback)
name(e3, Peyton Manning)...

---

name(x1, a)
role(x1, x2, Quarterback)
name(x2, Super Bowl XXXIII)
a?
```

Symbolists

Model behavior is produced by discrete composition of primitive reasoning operations.

Implementation-level understanding is most important.

Tools: grammars, logics, formal systems.

Inductive program synthesis

Input v_1	*Input* v_2	*Output*
Alex	*Asst.*	*Alex(Asst.)*
Jim	*Manager*	*Jim(Manager)*
Ryan	ϵ	ϵ
ϵ	*Asst.*	ϵ

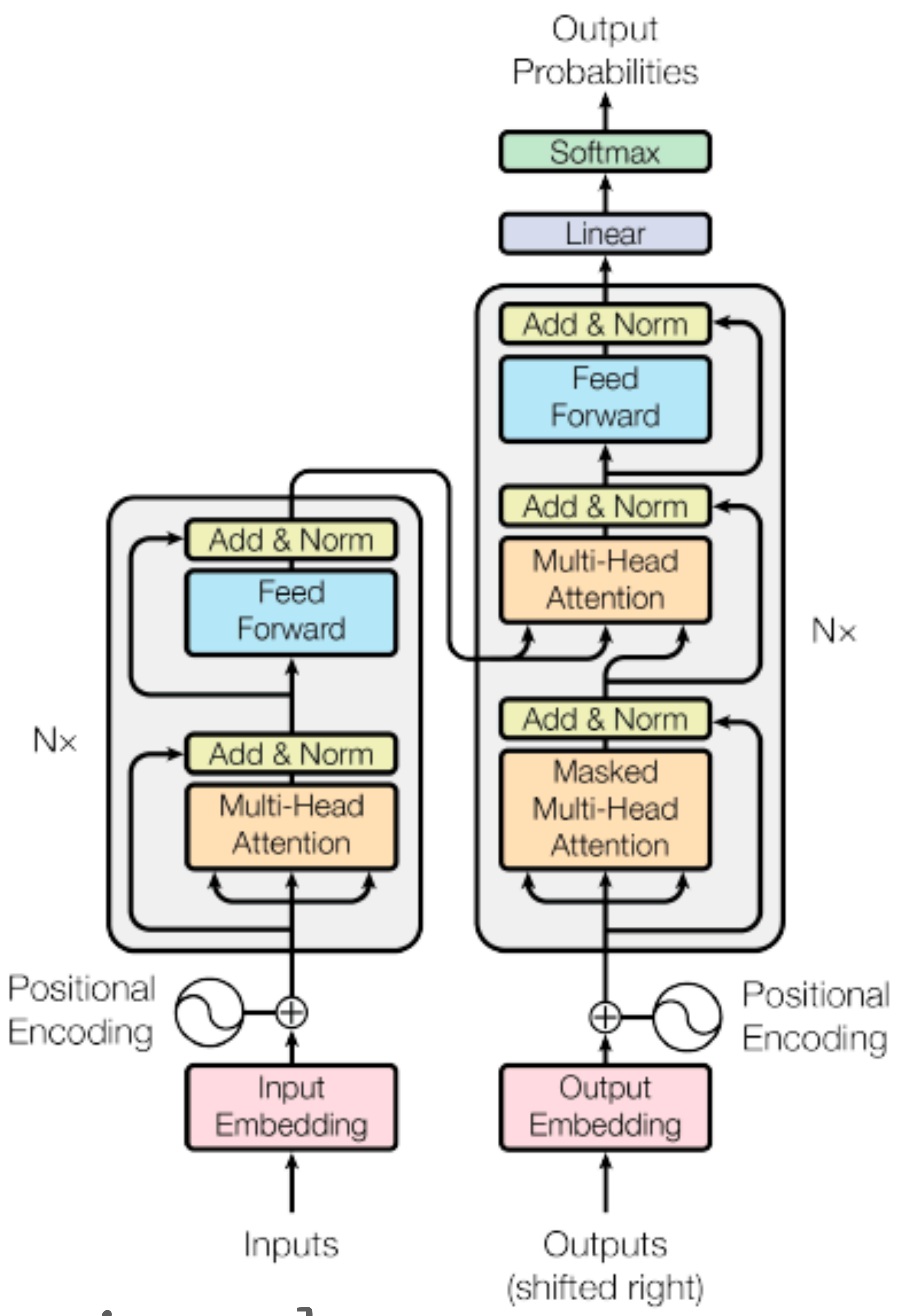
String Program:
$\texttt{Switch}((b_1, e_1), (b_2, \epsilon))$, *where*
$b_1 \equiv \texttt{Match}(v_1, \mathit{CharTok}) \wedge \texttt{Match}(v_2, \mathit{CharTok})$,
$e_1 \equiv \texttt{Concatenate}(v_1, \texttt{ConstStr}(\text{“(”}), v_2, \texttt{ConstStr}(\text{“)”}))$,
$b_2 \equiv \neg\texttt{Match}(v_1, \mathit{CharTok}) \vee \neg\texttt{Match}(v_2, \mathit{CharTok})$.

Input v_1	*Output*
01/21/2001	*01*
22.02.2002	*02*
2003-23-03	*03*

String Program:
$\texttt{Switch}((b_1, e_1), (b_2, e_2), (b_3, e_3))$, *where*
$b_1 \equiv \texttt{Match}(v_1, \mathit{SlashTok})$, $b_2 \equiv \texttt{Match}(v_1, \mathit{DotTok})$,
$b_3 \equiv \texttt{Match}(v_1, \mathit{HyphenTok})$,
$e_1 \equiv \texttt{SubStr}(v_1, \texttt{Pos}(\mathit{StartTok}, \epsilon, 1), \texttt{Pos}(\epsilon, \mathit{SlashTok}, 1))$
$e_2 \equiv \texttt{SubStr}(v_1, \texttt{Pos}(\mathit{DotTok}, \epsilon, 1), \texttt{Pos}(\epsilon, \mathit{DotTok}, 2))$
$e_3 \equiv \texttt{SubStr}(v_1, \texttt{Pos}(\mathit{HyphenTok}, \epsilon, 2), \texttt{Pos}(\mathit{EndTok}, \epsilon, 1))$

[Gulwani 2010]

Cultures in artificial intelligence research



[Vaswani 2017]

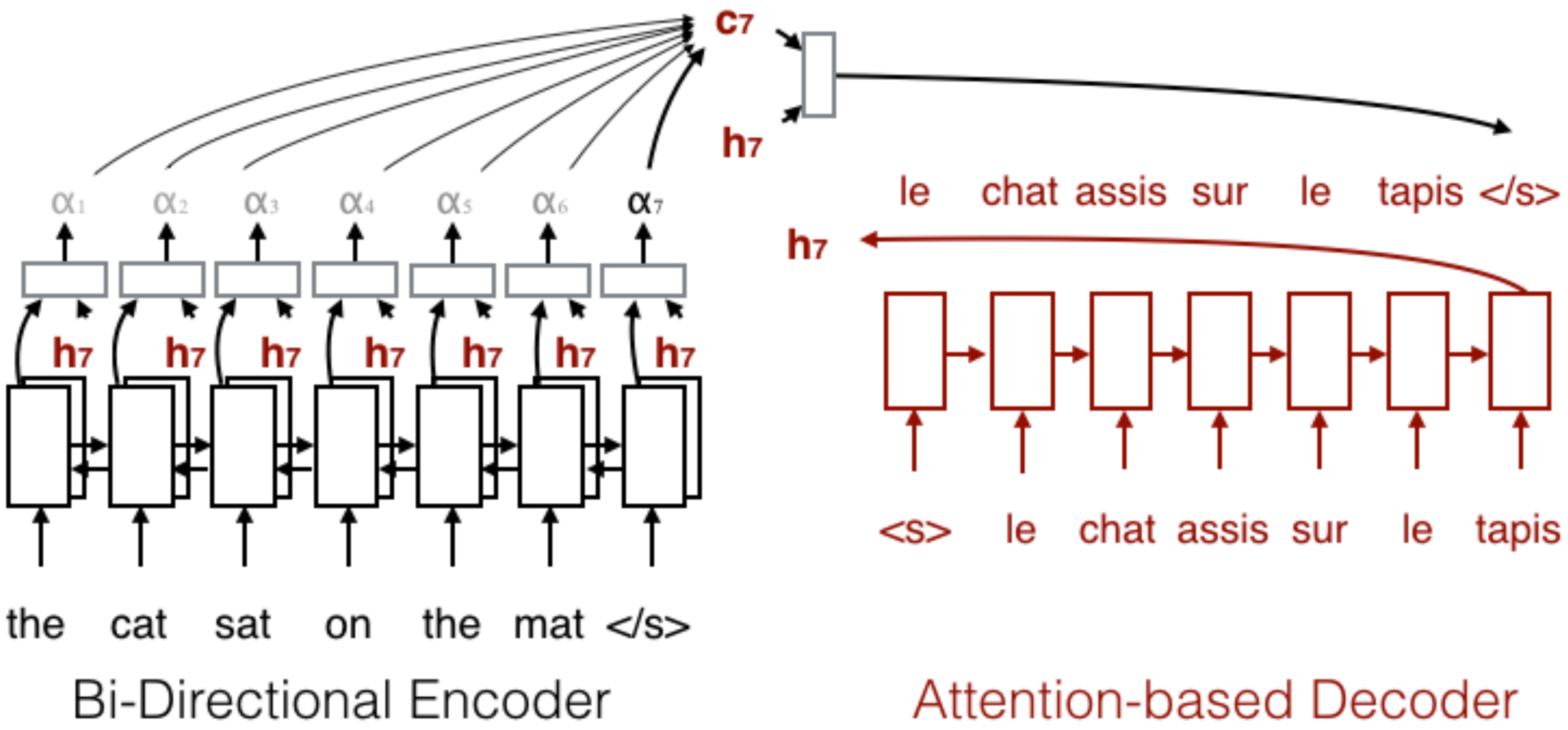
Connectionists

Model behavior is produced by generic operations on continuous representations.

Application-level understanding is most important.

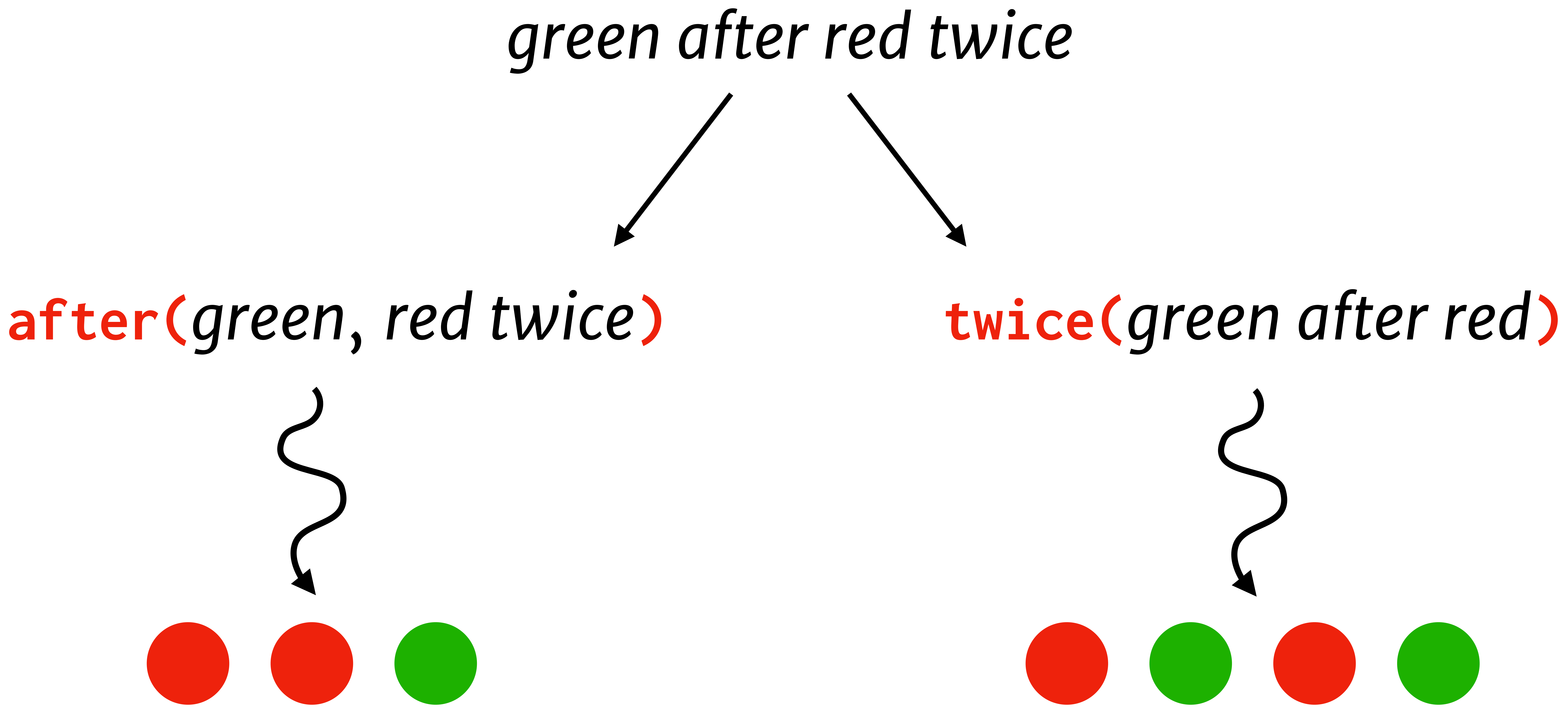
Tools: humongous datasets, data augmentation, pretraining.

Neural machine translation

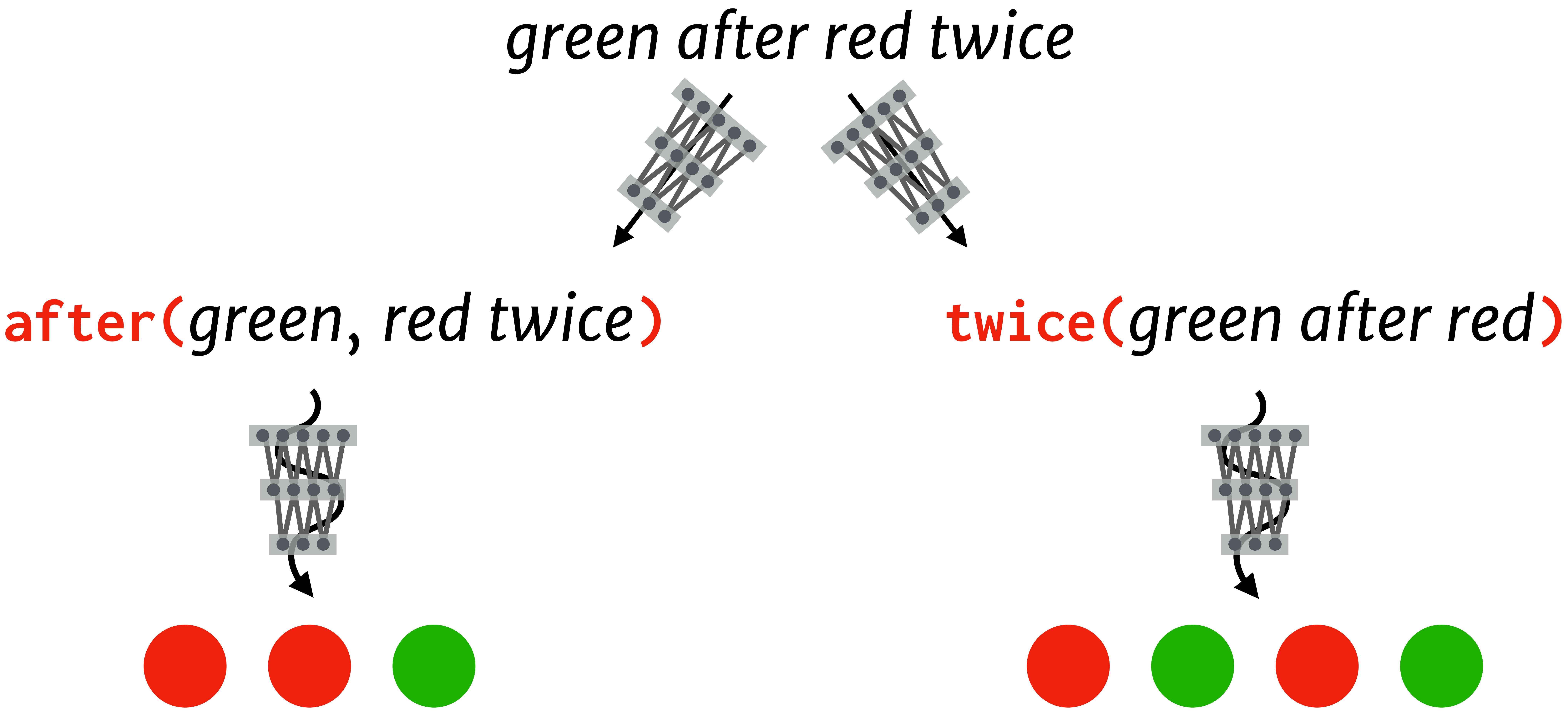


[Bahdanau 2015]

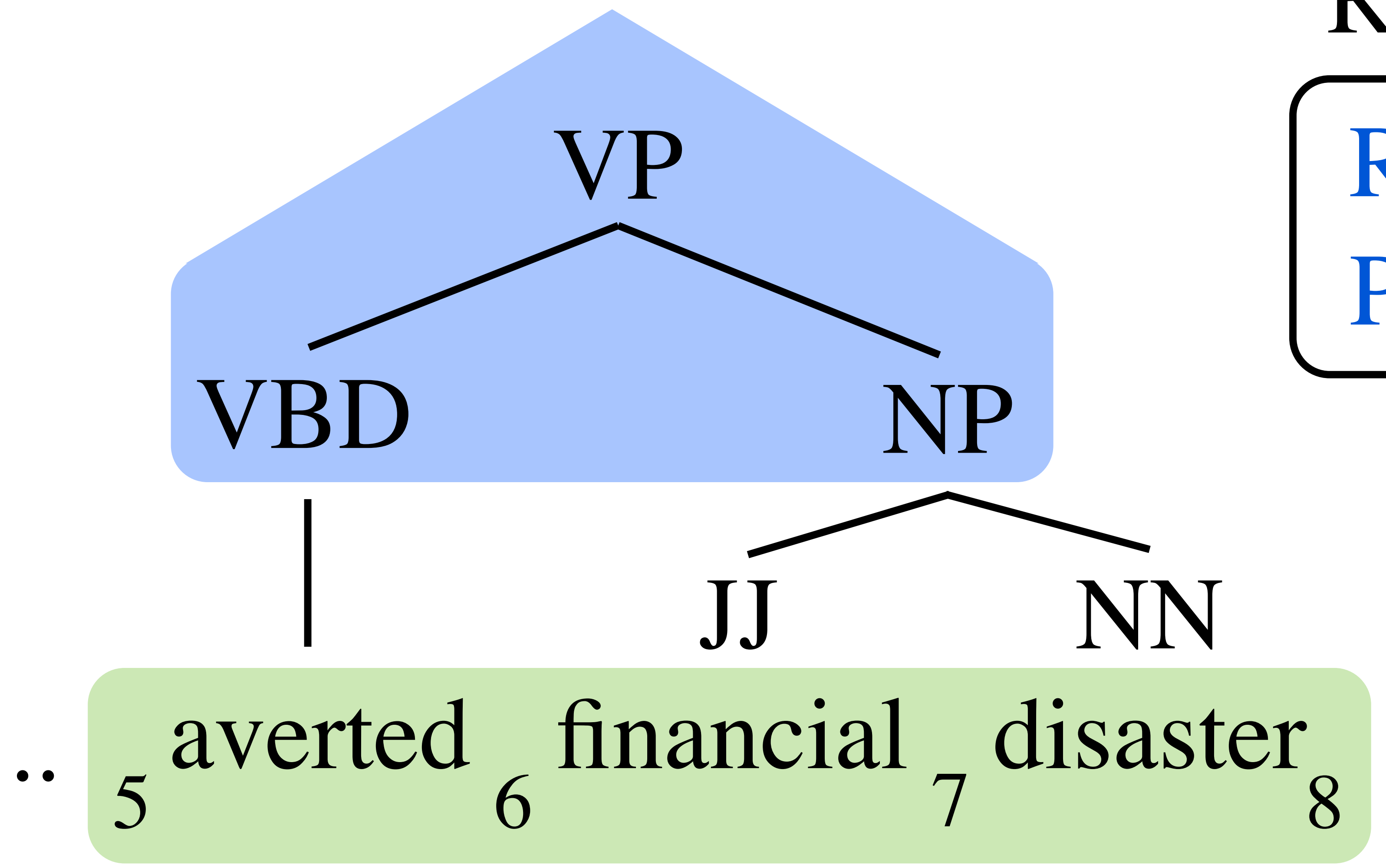
Connectionist scores in symbolist models



Connectionist scores in symbolist models



Connectionist scores in symbolist models



Rule backoffs

RULE = VP → VBD NP
PARENT = VP

Span properties

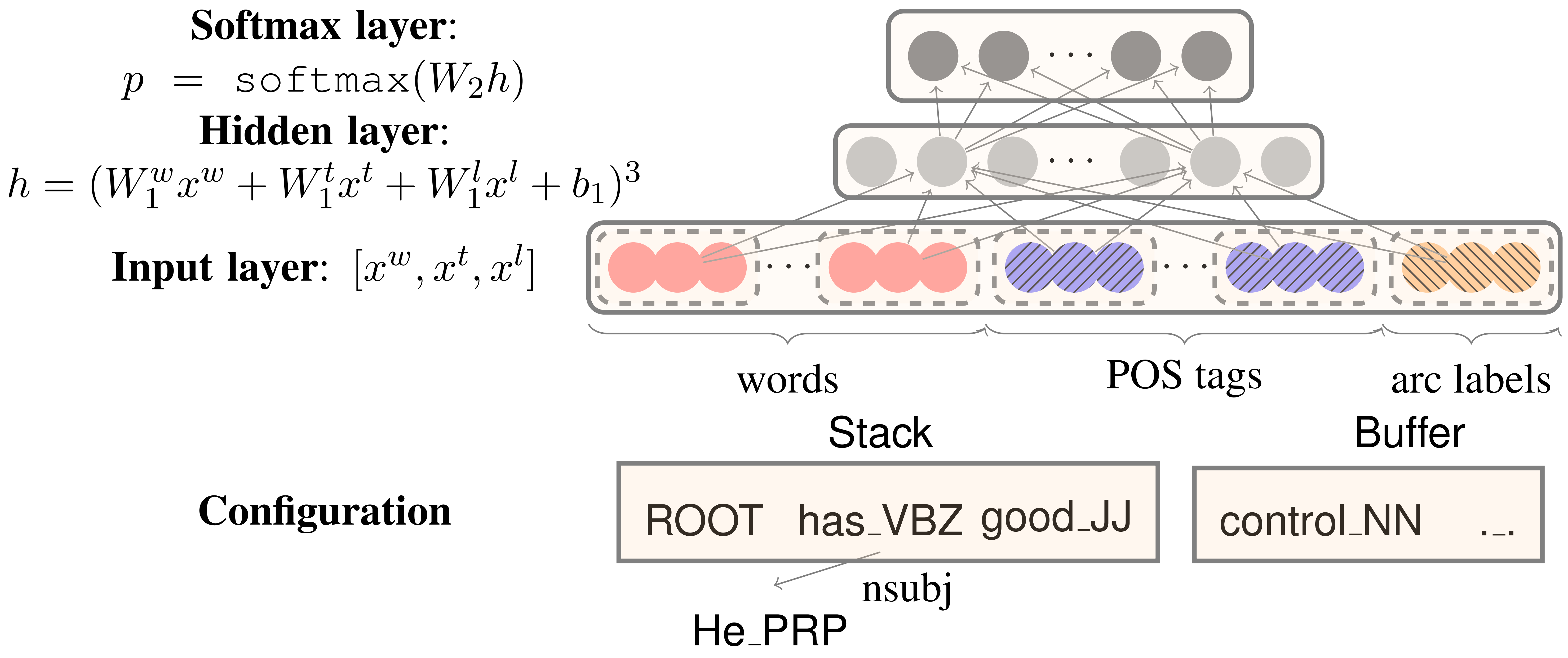
FIRSTWORD = averted
LASTWORD = disaster
LENGTH = 3

Features

FIRSTWORD = averted ⊗ PARENT = VP
FIRSTWORD = averted ⊗ RULE = VP → VBD NP
LASTWORD = disaster ⊗ PARENT = VP
...

[Hall et al. 2014]

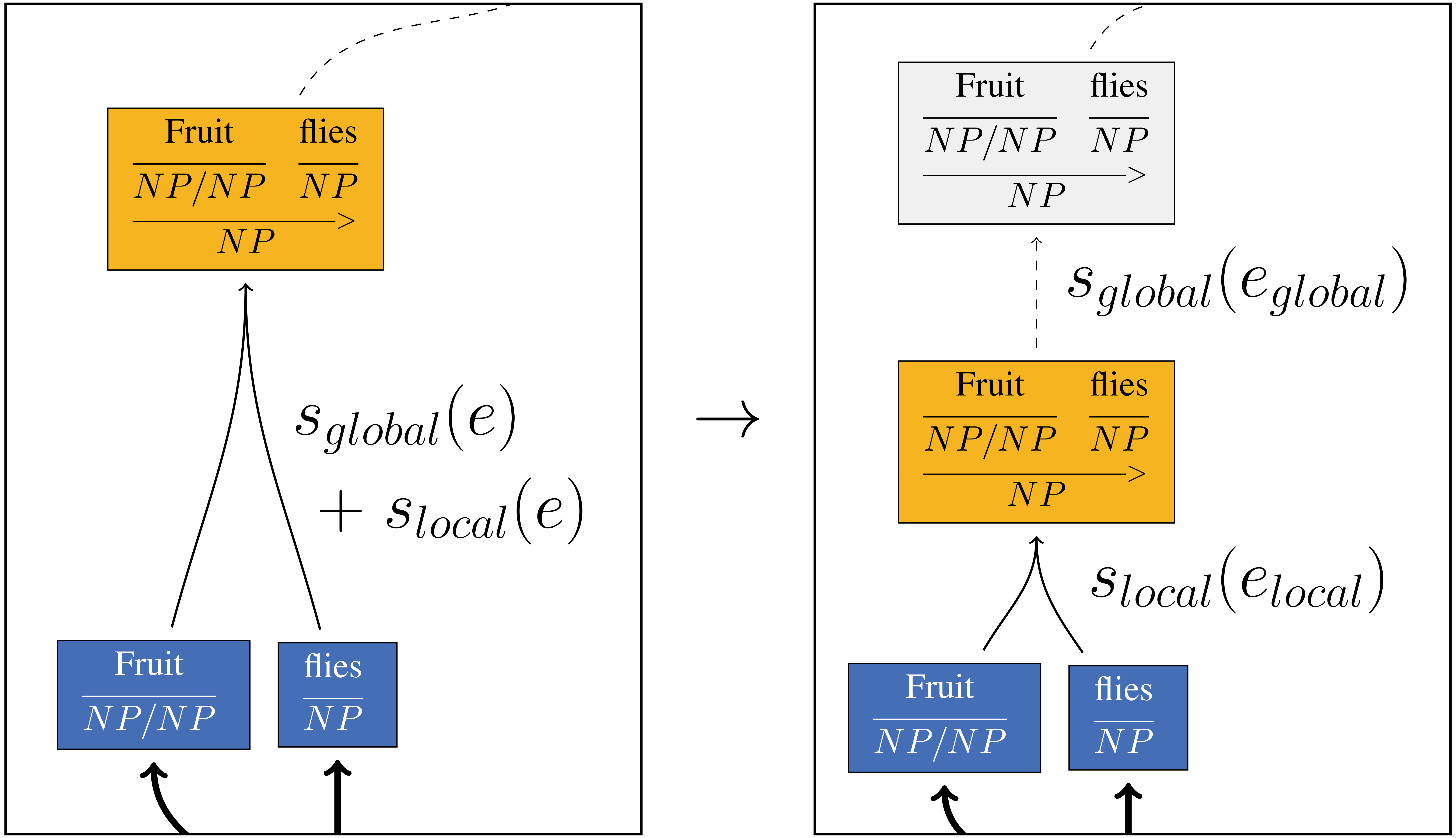
Connectionist scores in symbolist models



Transition	Stack	Buffer	A
	[ROOT]	[He has good control .]	$\emptyset$
`SHIFT`	[ROOT He]	[has good control .]	
`SHIFT`	[ROOT He has]	[good control .]	
`LEFT-ARC(nsubj)`	[ROOT has]	[good control .]	$A \cup$ nsubj(has,He)
`SHIFT`	[ROOT has good]	[control .]	
`SHIFT`	[ROOT has good control]	[.]	
`LEFT-ARC(amod)`	[ROOT has control]	[.]	$A \cup$amod(control,good)
`RIGHT-ARC(dobj)`	[ROOT has]	[.]	$A \cup$ dobj(has,control)
...	...	...	...
`RIGHT-ARC(root)`	[ROOT]	[]	$A \cup$ root(ROOT,has)

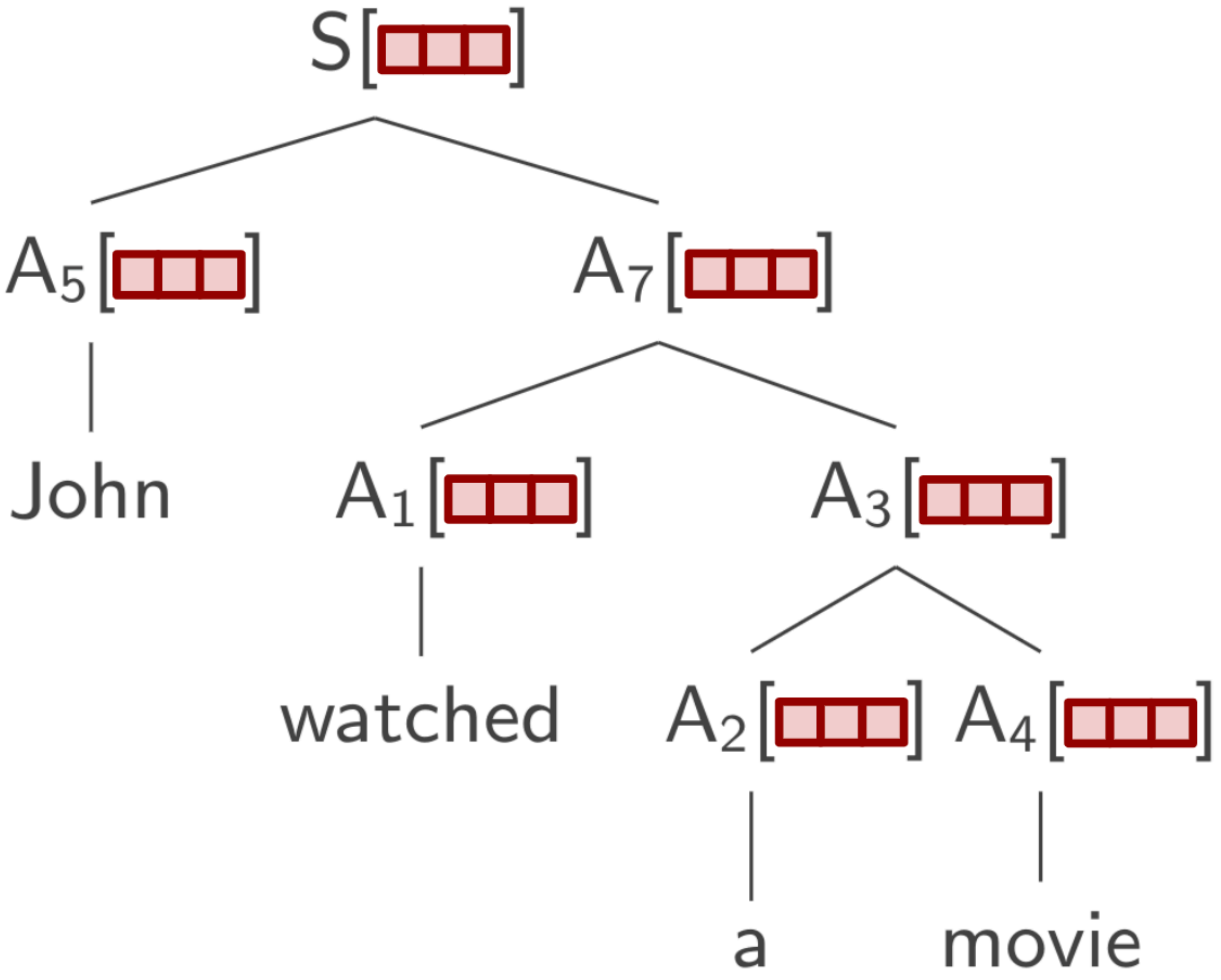
[Chen et al. 2014]

Connectionist scores in symbolist models



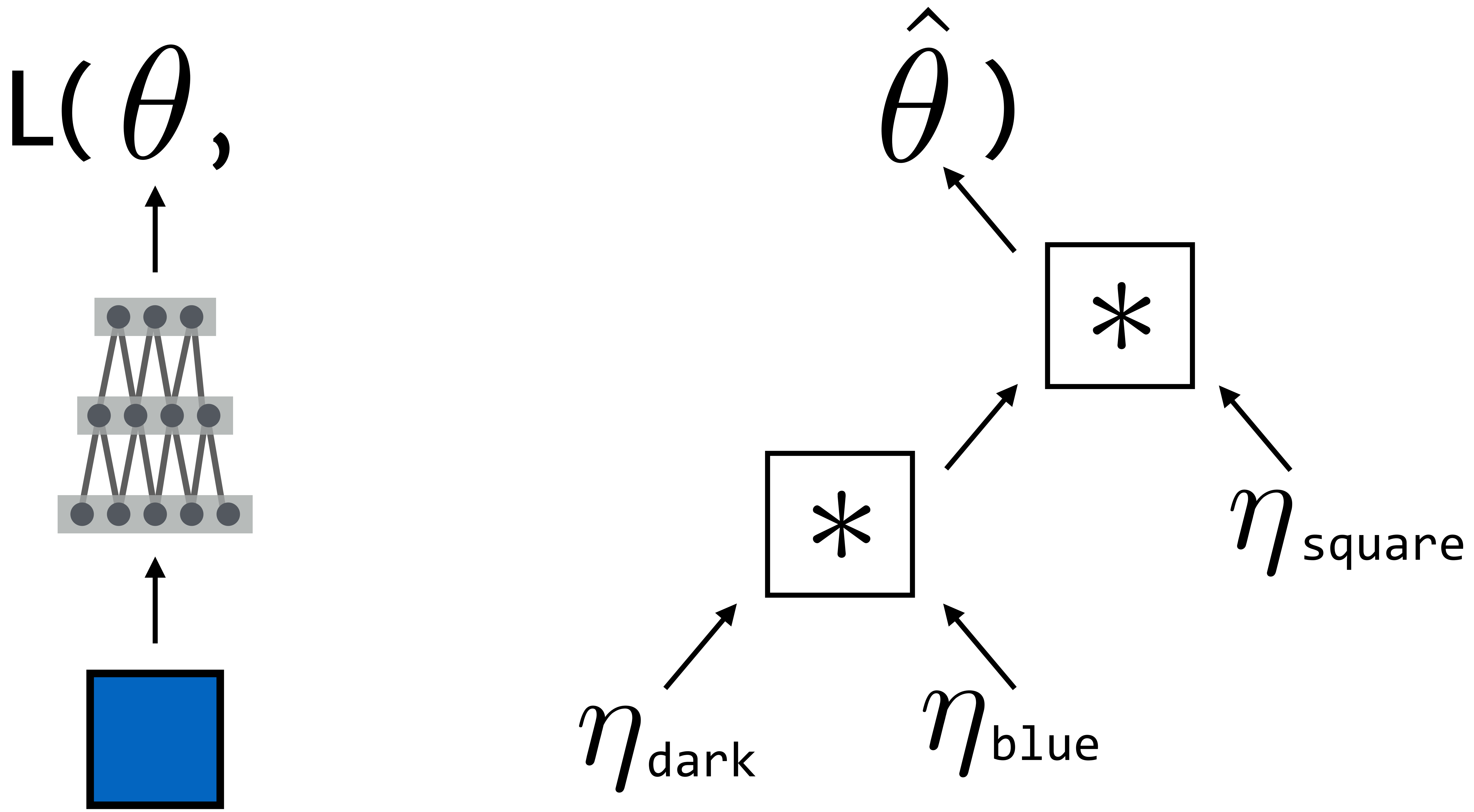
[Lee et al. 2016]

Connectionist scores in symbolist models



[Kim et al. 2019]

Symbolist losses for connectionist models

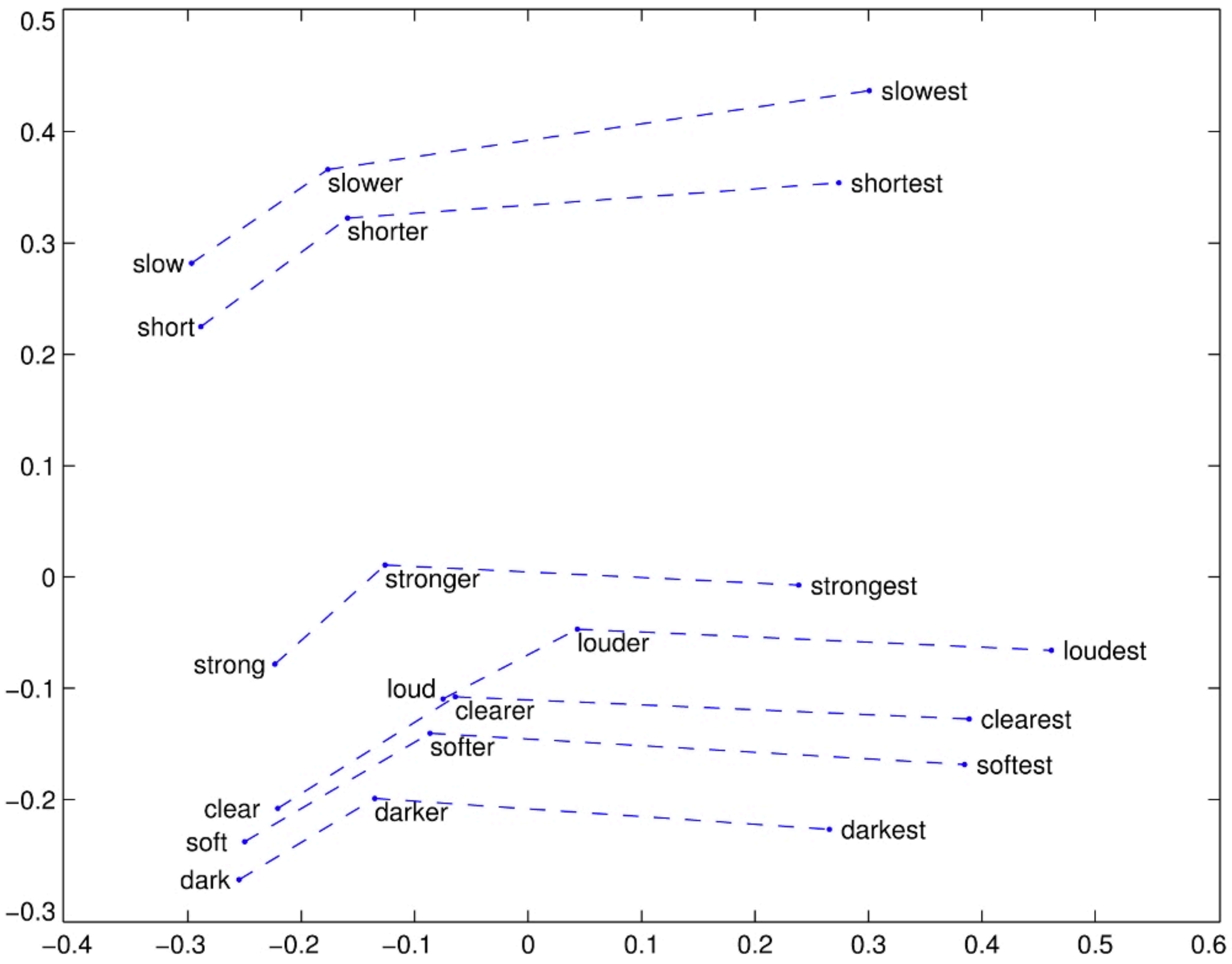


model

compositional approx.

[Andreas et al. 2016]

Symbolist losses for connectionist models



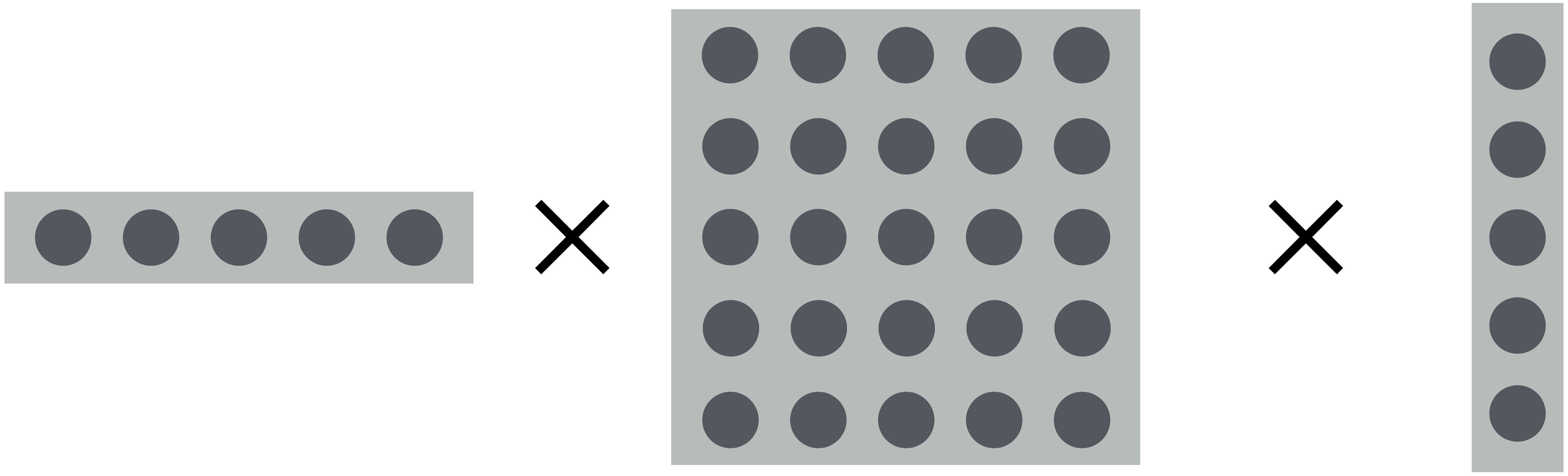
$$\mathcal{L}_{sim} = \mathbb{E}_{\mathbf{g}_{A\ldots D} \sim \mathcal{G}_{sim}} \left[\| \Delta \left(\mathbf{g}_A, \mathbf{g}_B \right) - \Delta \left(\mathbf{g}_C, \mathbf{g}_D \right) \|^2 \right]$$

$$\mathcal{L}_{dis} = \mathbb{E}_{\mathbf{g}_{A\ldots D} \sim \mathcal{G}_{dis}} \left[\left(\tau_{dis} - \| \Delta \left(\mathbf{g}_A, \mathbf{g}_B \right) - \Delta \left(\mathbf{g}_C, \mathbf{g}_D \right) \| \right)_+^2 \right]$$

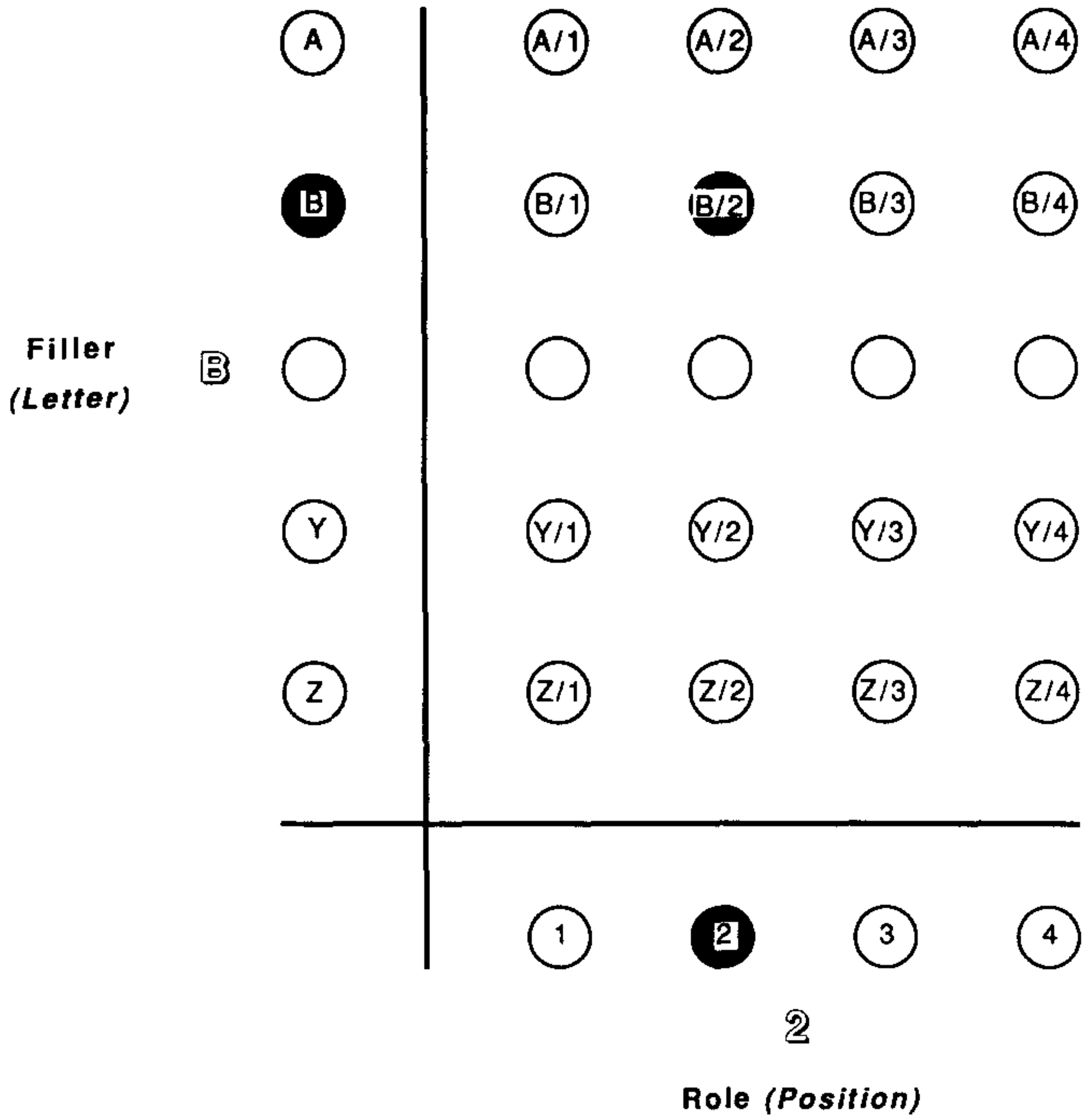
[Oh et al. 2017]

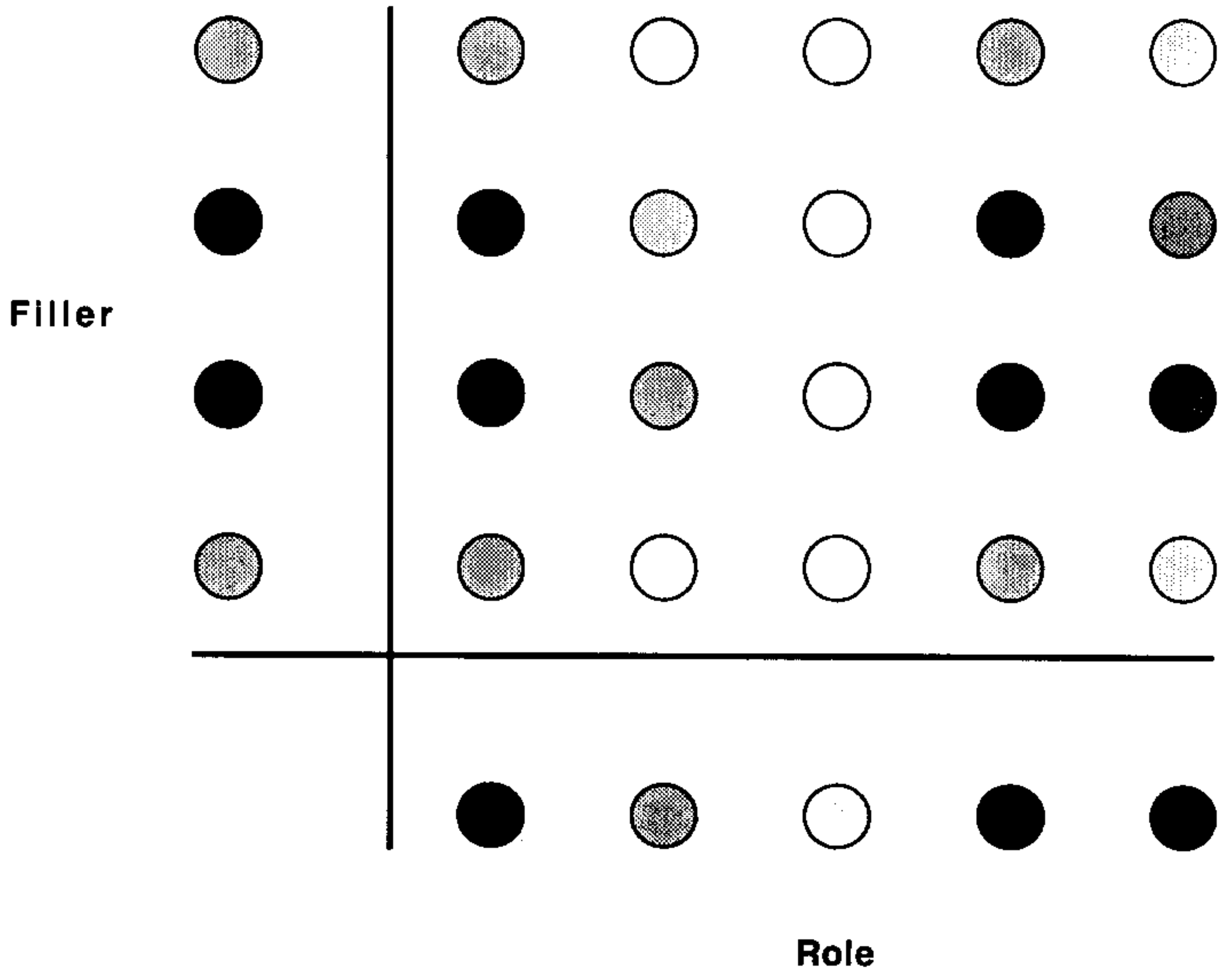
Symbol processing in connectionist models

green *after* *red twice*



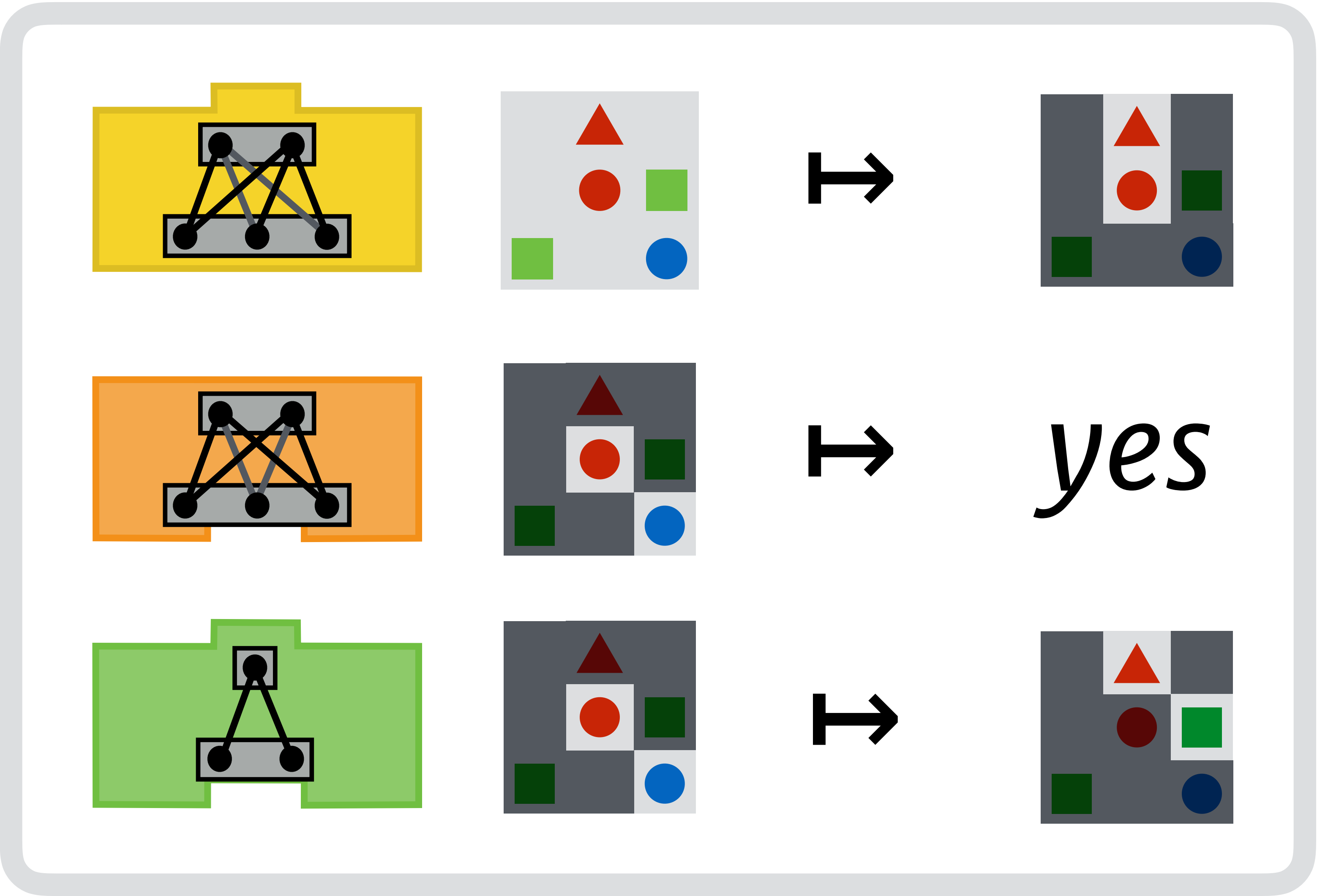
Symbol processing in connectionist models

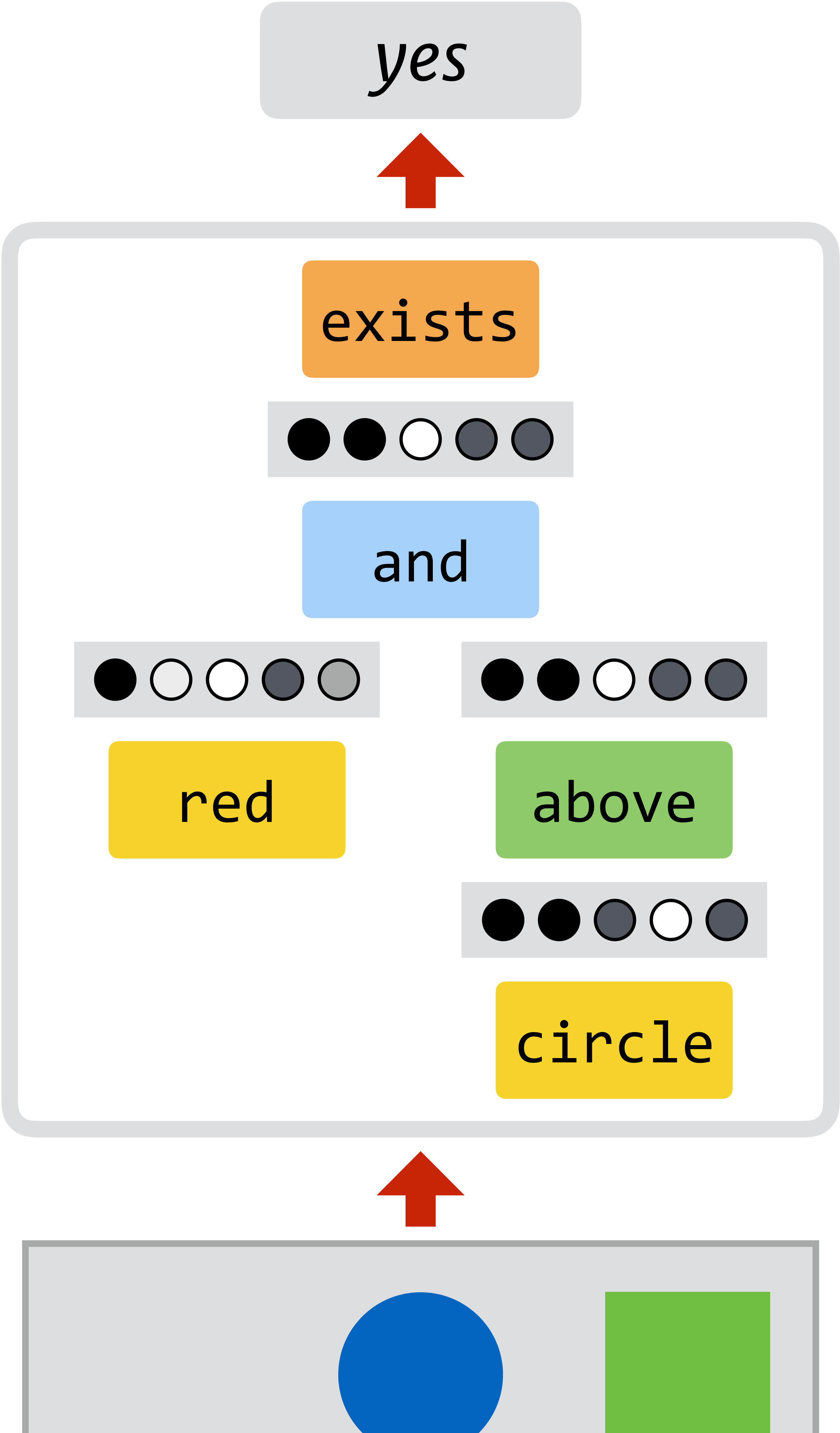




[Smolensky 1990]

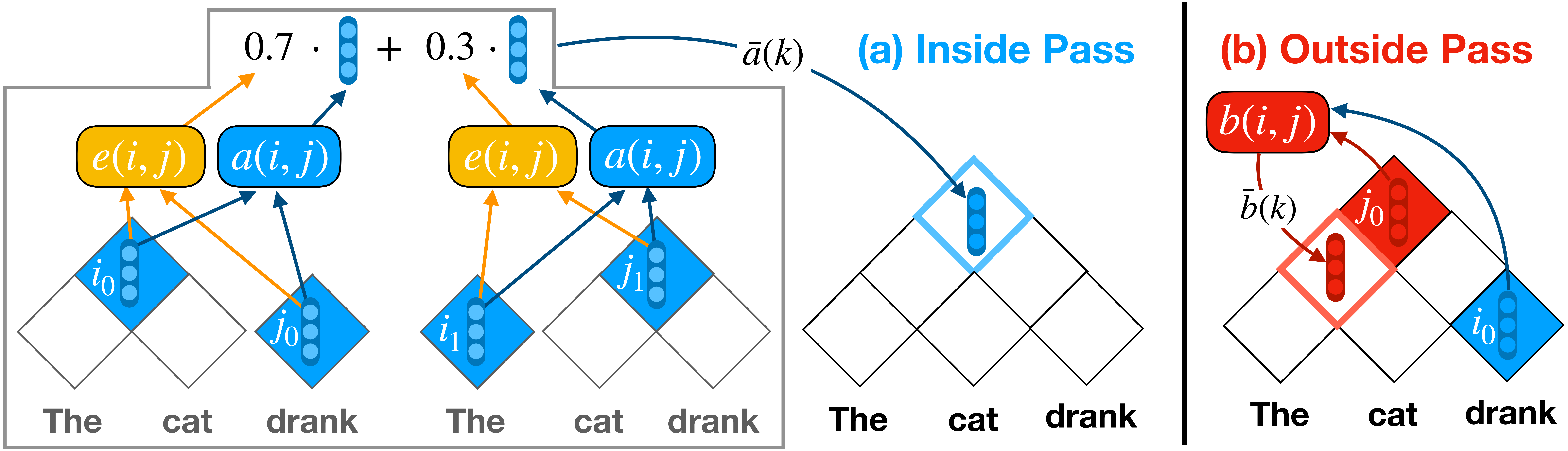
Symbolist processing in connectionist models





[Andreas et al. 2016]

(Is this even the right taxonomy?)



[Drozdov et al. 2018]

(How much of this do we need?)

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
3	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
4	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6

[https://super.gluebenchmark.com/leaderboard/]

(How much of this do we need?)

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
+	3	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	4	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

Break

There's a lot we still don't know

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way?

(2) Under what conditions do generic neural models already succeed at symbolic generalization?

(3) To what extent are successes supported by implicit symbol manipulation operations in vector space?

(4) What modeling tools are available to us beyond the standard seq2seq toolkit for dealing with failures?

There's a lot we still don't know

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way?

(2) Under what conditions do generic neural models already succeed at symbolic generalization?

(3) To what extent are successes supported by implicit symbol manipulation operations in vector space?

(4) What modeling tools are available to us beyond the standard seq2seq toolkit for dealing with failures?

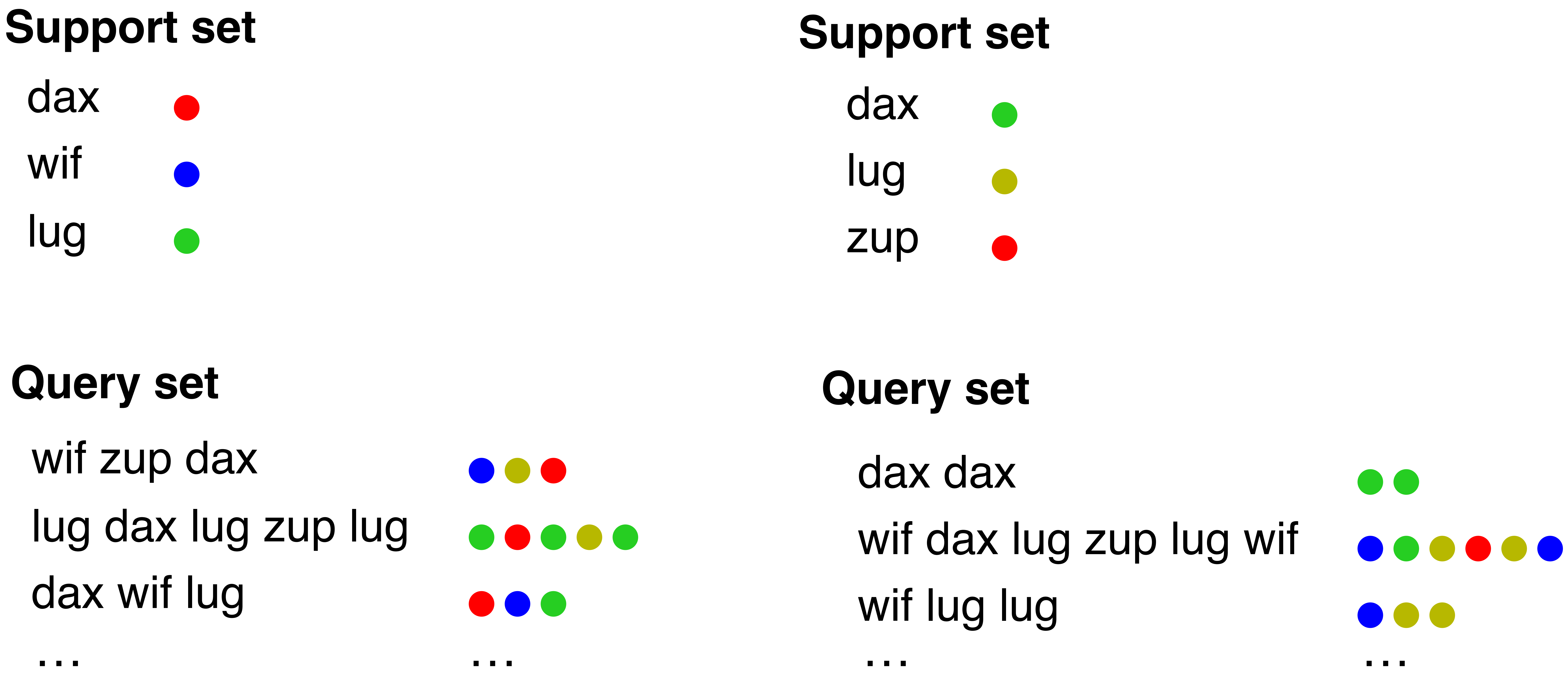
Goal for this 6.884: answer these questions!

Symbolic generalization

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way? To what extent do current models already do it?

Symbolic generalization

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way? To what extent do current models already do it?



Characterizing generalization

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way? To what extent do current models already do it?

```
d a → d a
b c → b c
a b → a b
---
c a → c a
```

4 chars
→ 52% accuracy

Characterizing generalization

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way? To what extent do current models already do it?

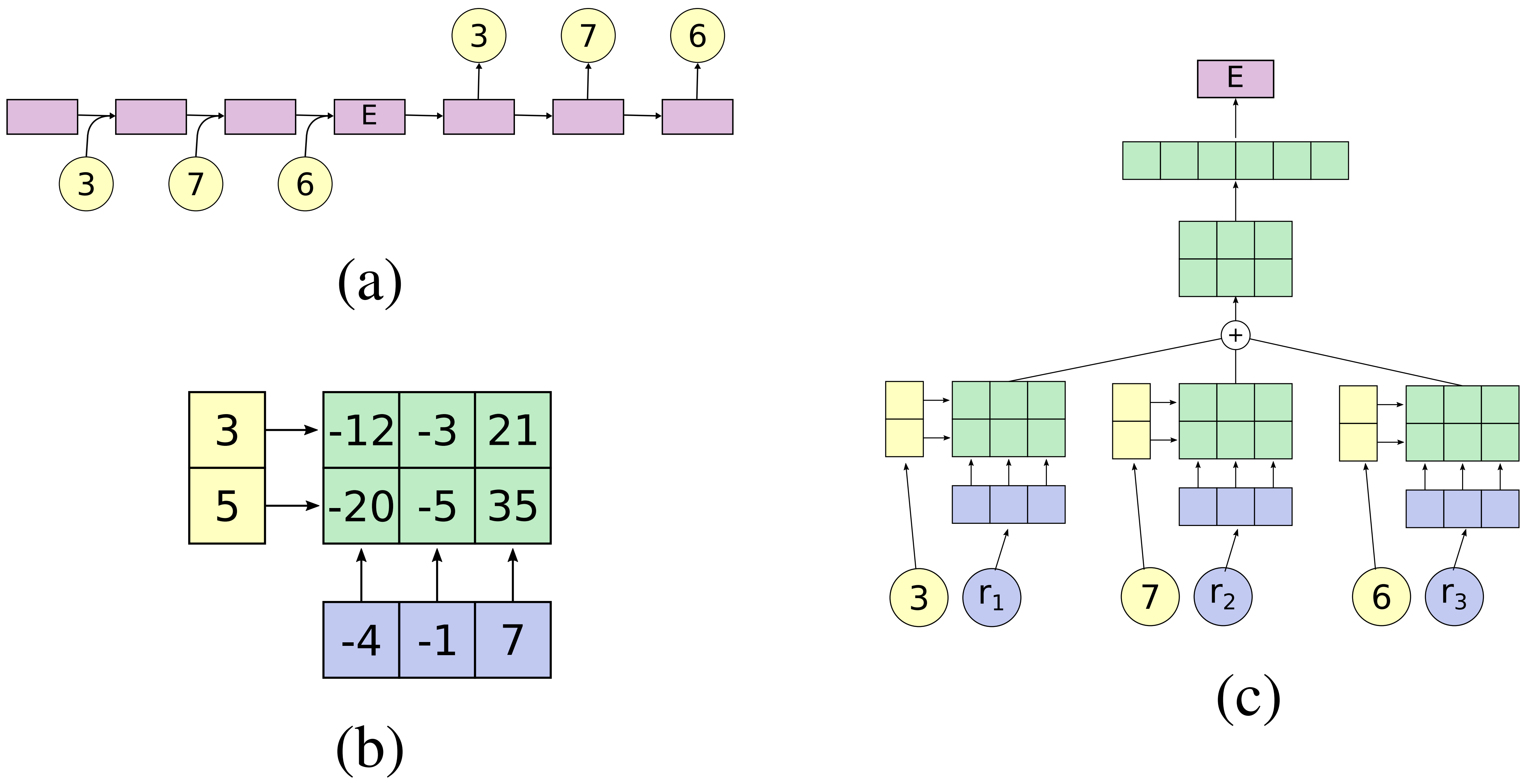
d a → d a	d a → d a	d e → d e
b c → b c	b c → b c	b c → b c
a b → a b	a b → a b	h b → h b
---	---	---
c a → c a	c a → c a	c a → c a
4 chars → 52% accuracy	6 chars → 82% accuracy	7 chars → 100% accuracy

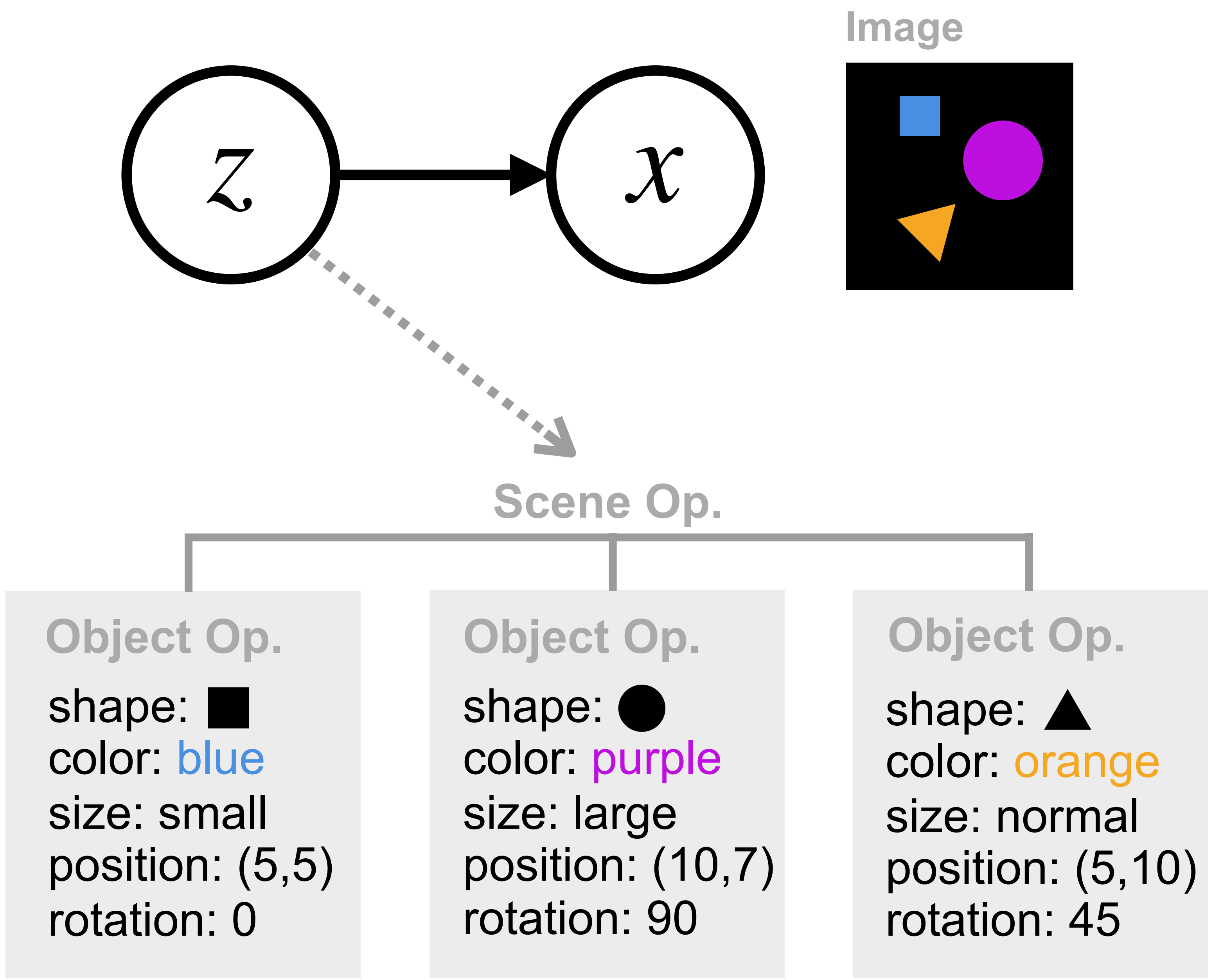
Recognizing symbolic processing

(2) To what extent are successes supported by implicit symbol-manipulation operations? How are these operations implemented?

Recognizing symbolic processing

(2) To what extent are successes supported by implicit symbol-manipulation operations? How are these operations implemented?



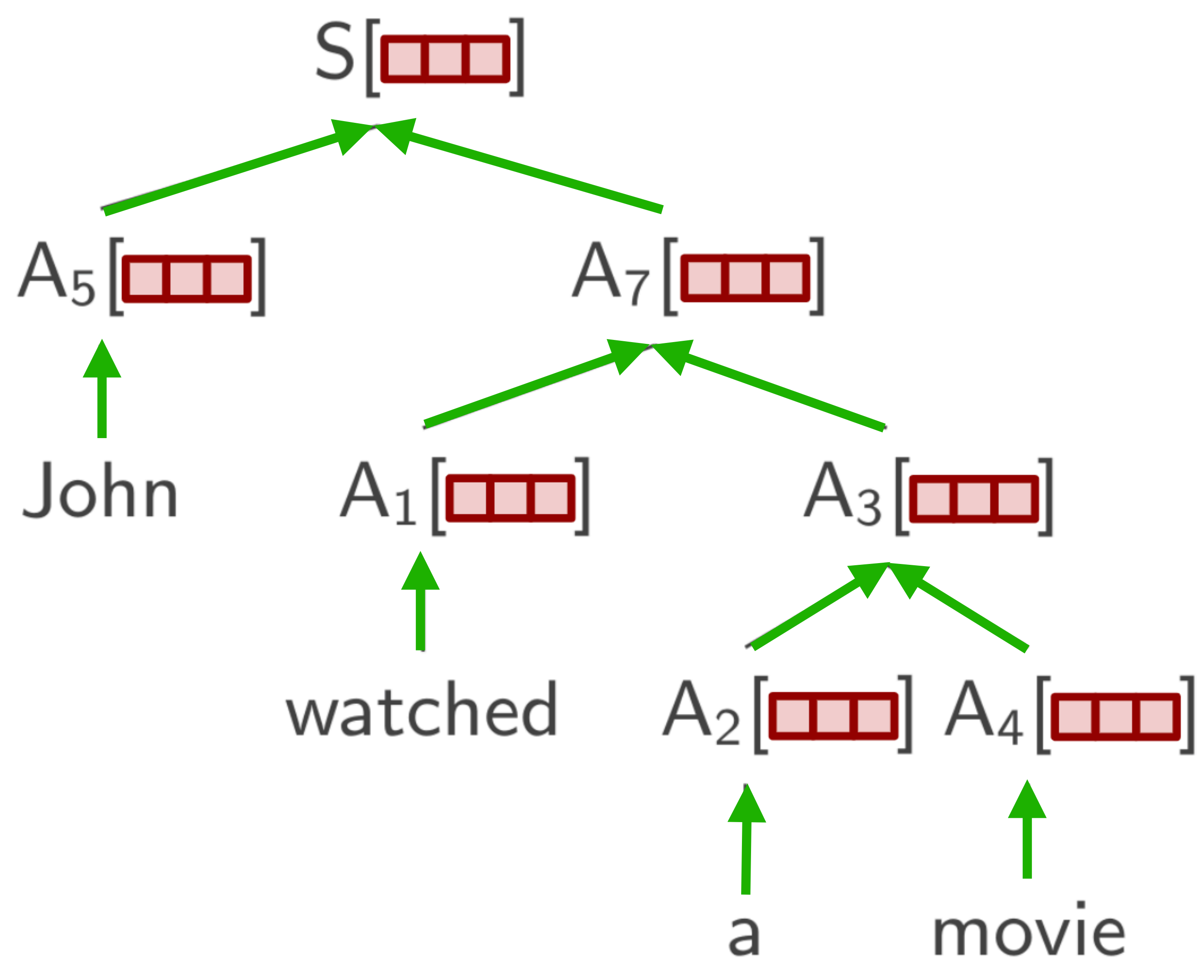


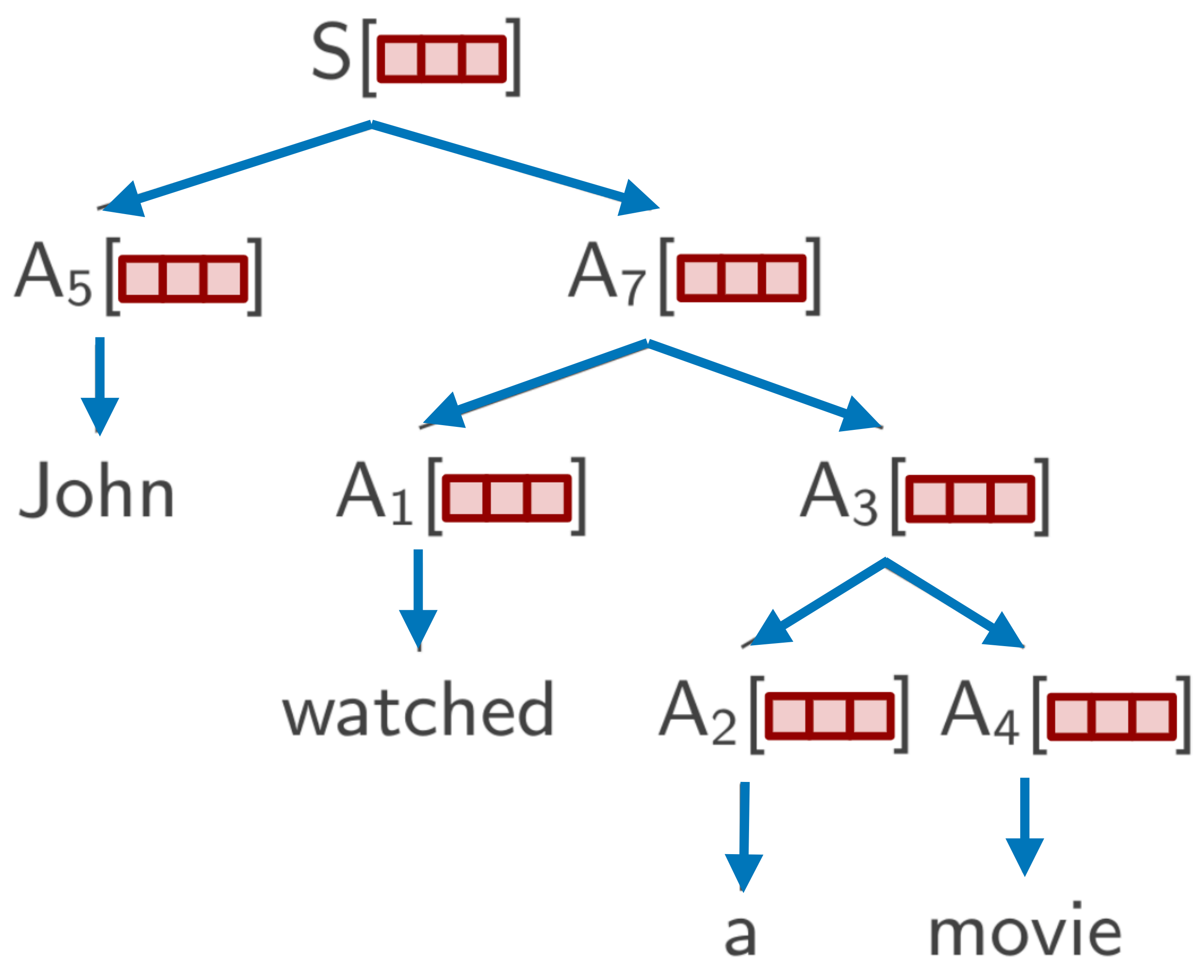
Improving model performance

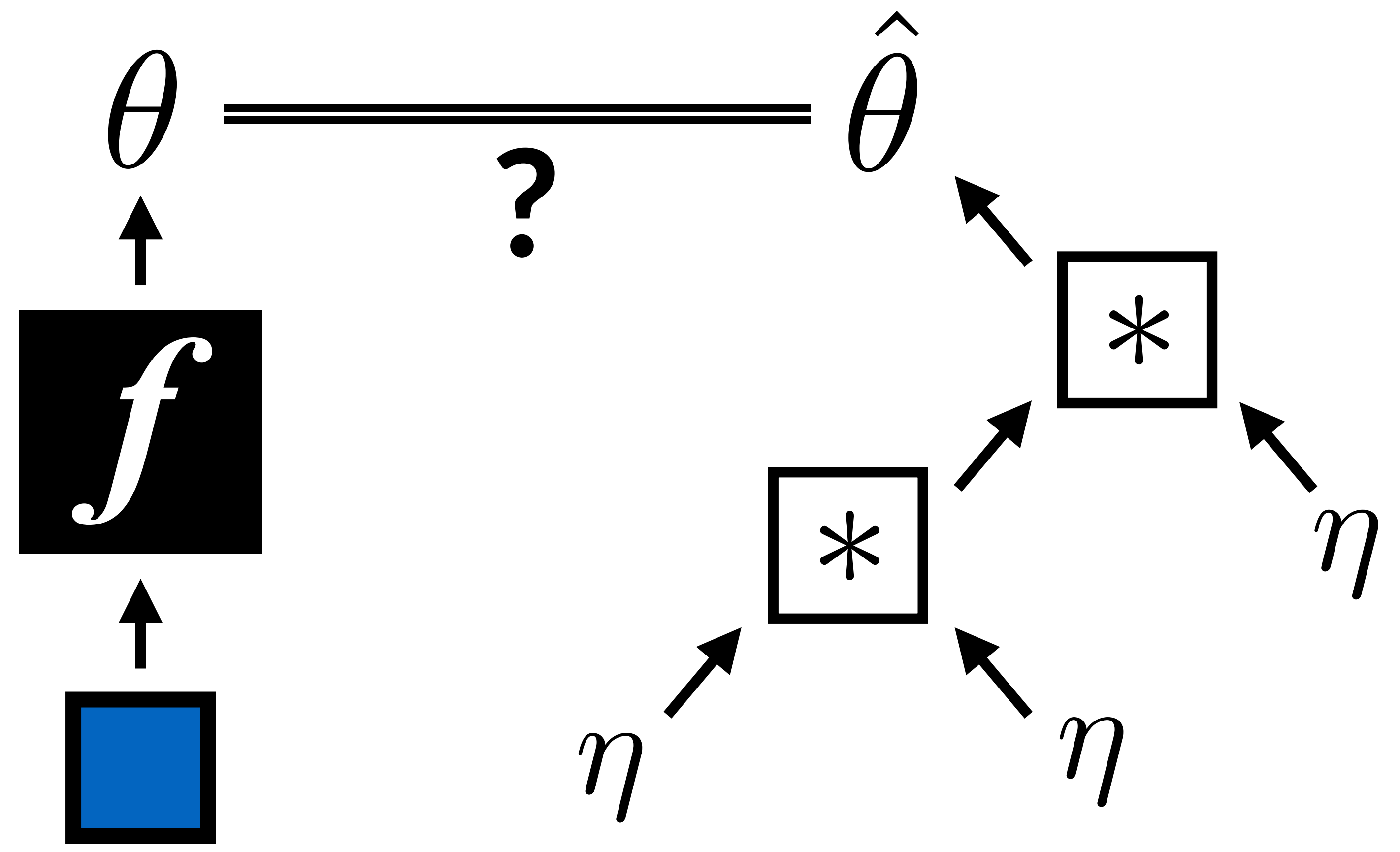
(3) What modeling tools are available to us beyond the standard seq2seq toolkit for dealing with failures?

Improving model performance

(3) What modeling tools are available to us beyond the standard seq2seq toolkit for dealing with failures?







Admin

A note on enrollment

A note on enrollment

A note on enrollment

Final decisions will go out this weekend.

Priority for:
- PhD students
- students with prior NLP / ling experience

Syllabus online!

Background

Neural networks

Deep learning, Goodfellow and Courville

NLP

http://web.mit.edu/jda/www/teaching/6.864/

Language

Linguistic fundamentals for natural language processing, Bender

"Classical AI"

AI: a Modern Approach, Russell and Norvig

Course components

Reading responses & participation

In-class presentations

Final project

Reading responses & participation

By Thursday night before class:

add a comment to the Piazza thread for the day's readings
or
write a response to someone else's comment!

In class:

ask lots of questions!

Reading logistics

Piazza:

https://piazza.com/class/kecwn7kgtec743

Feel free to use for other discussion as well.

In-class presentations

Sign up for a presentation slot.

Meet with me on Wednesday before your presentation.

Present in class!

Presentation logistics

Sign-up spreadsheet:

https://docs.google.com/spreadsheets/d/1VEmxvc-tKo7AeDypkPQgHRe0dVFSUJex6E-jY5JPKas/edit?usp=sharing

Email me [jda@mit.edu] to set up a check-in time.

(Tues morning & Weds are flexible.)

Final project

Say something new about neuro-symbolic NLP!

(combining with your own research / other projects
is **strongly encouraged**)

Preference for groups of 2-4.

Project logistics

Use Piazza to find groups.

2 **written** assignments:

Project proposal (due 2 Oct)

Project writeup (due 4 Dec)

2 **in-person assignments:**

Preliminary discussion

Final presentation

Feel free to reach out with other questions!

Class / Zoom logistics

I prefer that people have their cameras on (but it's fine if you don't want to).

Please mute yourself except when speaking and use the "raise hand" feature.

Class runs from 11:35–1:25 with a 10-minute break.

Office hours 2–4 on Thurs (email me to schedule).

"Grading"

33% participation

33% presentation

33% project

Other disclaimers

THIS IS NOT A NORMAL SEMESTER

I hope this class is a source of joy.

If it becomes a source of stress,
let me know and we'll find a way
to fix it!

Important websites

Course homepage

http://web.mit.edu/jda/www/teaching/6.884/

Piazza

https://piazza.com/class/kecwn7kgtec743

Project signup

https://docs.google.com/spreadsheets/d/1VEmxvc-tKo7AeDypkPQgHRe0dVFSUJex6E-jY5JPKas/edit?usp=sharing

Topic outline

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way? To what extent do current models already do it?

11 September: composition

18 September: syntax and reasoning

25 September: pretraining and scale

Recognizing symbolic processing

(2) To what extent are successes supported by implicit symbol-manipulation operations? How are these operations implemented?

2 October: connectionist symbol processing

9 October: discrete representations

16 October: modular representations

23 October: modular computations

Encouraging symbolic generalization

(3) What modeling tools are available to us beyond the standard seq2seq toolkit for dealing with failures?

11 September: structured neural models

18 September: structured losses

25 September: structured data and meta-learning

Sample project 1

(1) Can we usefully formalize "symbol-like" generalization in a task-independent way? To what extent do current models already do it?

Large-scale empirical study of generalization on synthetic sequence data.

How do training set size, vocabulary size, syntactic complexity and frequency distribution affect empirical properties of symbolic generalization?

Sample project 2

(2) To what extent are successes supported by implicit symbol-manipulation operations? How are these operations implemented?

Investigation of implicit types and type constraints in neural sequence models.

Do pretrained sequence model representations encode abstract notions of syntactic type and constituency?

Sample project 3

(3) What modeling tools are available to us beyond the standard seq2seq toolkit for dealing with failures?

New sequence modeling architectures (e.g. transformers with tree-shaped attention).

Does imposing linguistically motivated structure on generic sequence models improve their generalization on ordinary tasks and hard ones?

Your jobs for next class

1. Have a nice weekend 😀
2. Sign up for a project presentation
3. Do the reading for next class

See you next week!