

# Probing pretrained models

CS 685, Fall 2020

Introduction to Natural Language Processing

<http://people.cs.umass.edu/~miyyer/cs685/>

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

*most slides from Tu Vu*

# Logistics stuff

- Reminder: don't discuss exam, people are still doing makeups!
- Exam grading: to be done by end of the month
- Remaining todos for you: final project report, a couple quizzes, and (optionally) the extra credit HW

# BERTology



# BERTology

studying the inner working of large-scale Transformer language models like BERT

- what are captured in different model components, e.g., attention / hidden states?





# tools & examples

BERTology - HuggingFace's Transformers

<https://huggingface.co/transformers/bertology.html>



- accessing all the hidden states of BERT
- accessing all the attention weights for each head of BERT
- retrieving heads output values and gradients

## tools & examples (cont.)

Are Sixteen Heads Really Better than One? Michel et al., NeurIPS 2019

large percentage of attention heads can be removed at test time without significantly impacting performance

What Does BERT Look At? An Analysis of BERT's Attention, Clark et al., BlackBoxNLP 2019

substantial syntactic information is captured in BERT's attention

# tools & examples

AllenNLP Interpret  
<https://allennlp.org/interpret>



AllenNLP

## Simple Gradients Visualization

See saliency map interpretations generated by [visualizing the gradient](#).

### Saliency Map:

[CLS] The [MASK] rushed to the **emergency** room to see **her** patient . [SEP]

### Mask 1 Predictions:

- 47.1% **nurse**
- 16.4% **woman**
- 10.0% **doctor**
- 3.4% **mother**
- 3.0% **girl**

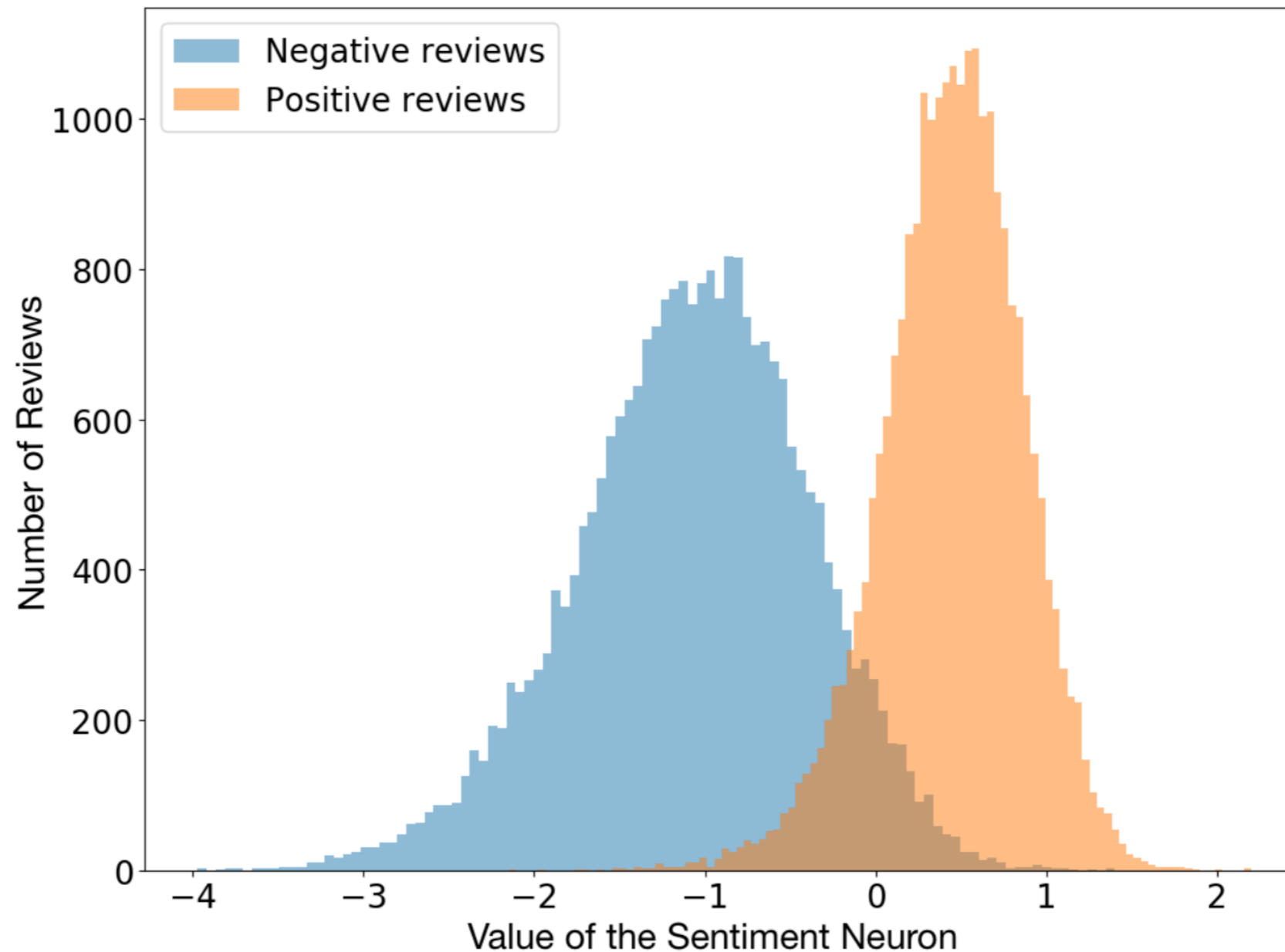
# understanding contextualized representations

two most prominent methods

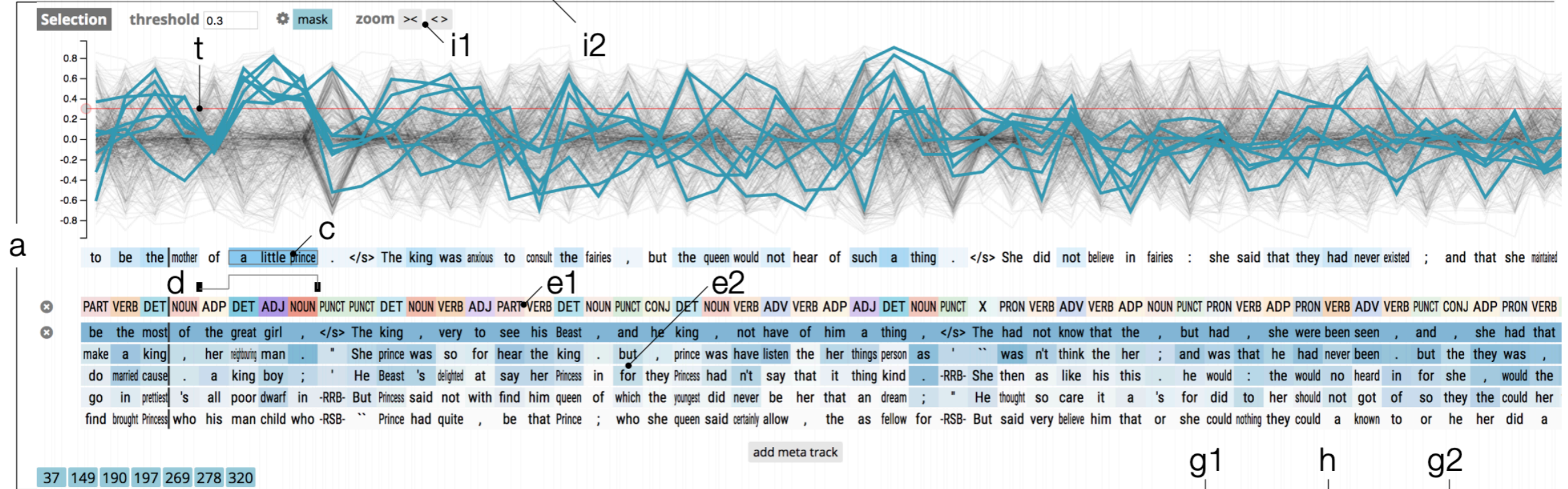
- visualization
- linguistic probe tasks

# Sentiment neuron

While training the linear model with L1 regularization, we noticed it used surprisingly few of the learned units. Digging in, we realized there actually existed a single “sentiment neuron” that’s highly predictive of the sentiment value.

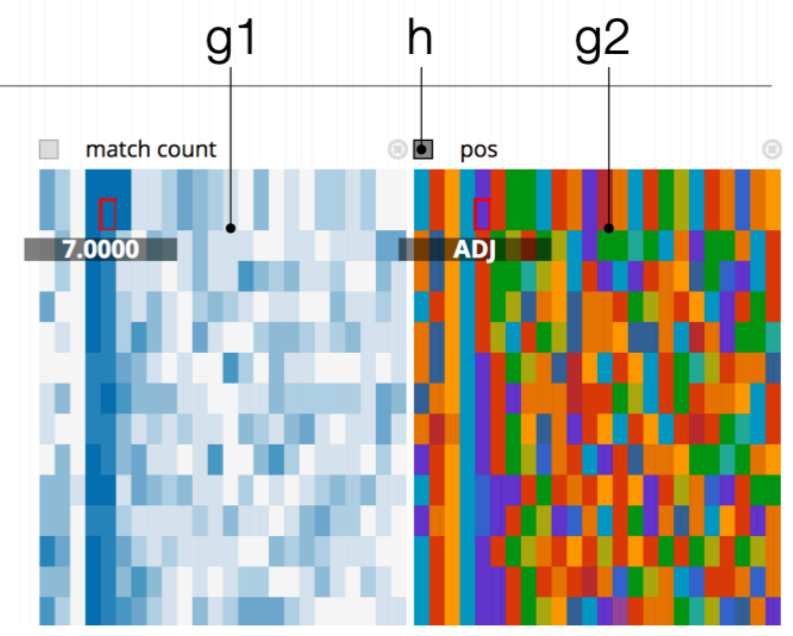


The sentiment neuron within our model can classify reviews as negative or positive, even though the model is trained only to predict the next character in the text.



**Matching** match fast precise mask stats meta match count ner pos

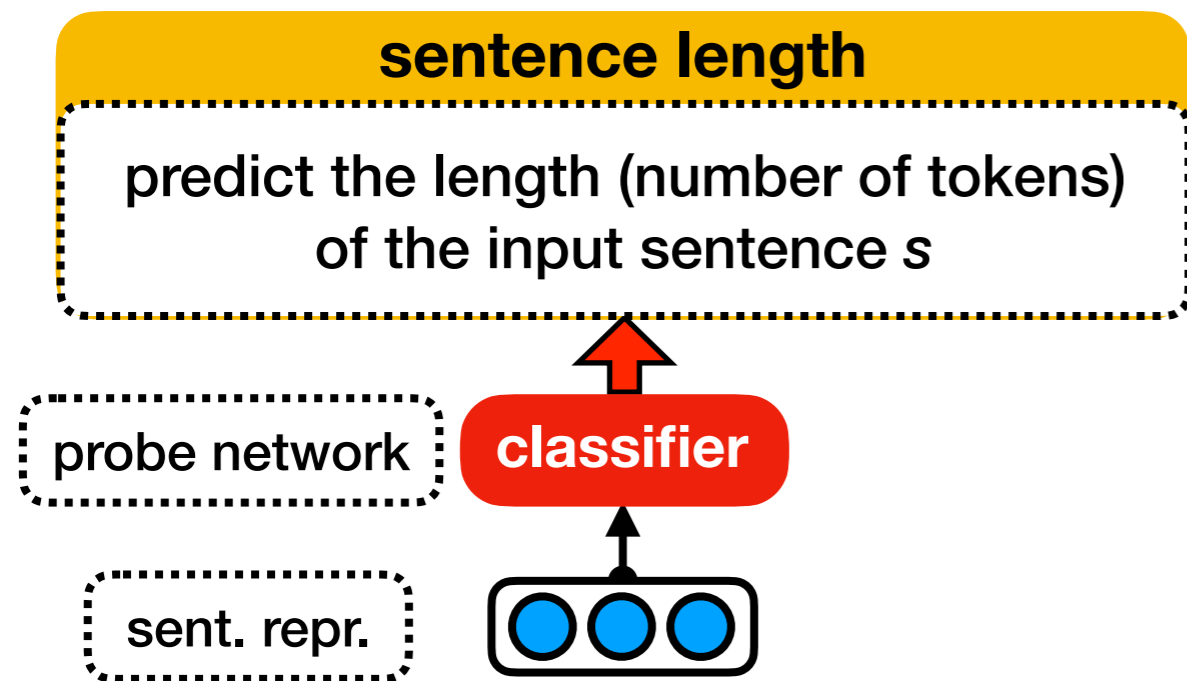
<input checked="" type="checkbox"/>	the mother of	a little prince . </s> The king was anxious to consult the fairies , but the queen would not hear of
<input checked="" type="checkbox"/>	the mother of	a little prince . </s> The king was anxious to consult the fairies , but the queen would not hear of
<input checked="" type="checkbox"/>	transform yourself into	a lion , or elephant , and the like . " </s> " That is true , " answered the ogre
<input checked="" type="checkbox"/>	change yourself into	a lion . ' </s> And in a moment such a fierce creature stood before them , that all the guests
<input checked="" type="checkbox"/>	the court of	a king , and it happened that he was holding games , and giving prizes to the best runners , boxers
<input checked="" type="checkbox"/>	change yourself into	a rat or a mouse ; but I must own to you I take this to be impossible . " </s>
<input checked="" type="checkbox"/>	led him into	a little shed , and chained him up to a ring in the wall . </s> But food was given him
<input checked="" type="checkbox"/>	he fell into	a deep pit which had been made to trap bears , and the hunters , who were hiding in a tree
<input checked="" type="checkbox"/>	happened to want	a little boy , so she threw her ball in the direction of the hunters ' huts . </s> A child
<input checked="" type="checkbox"/>	prime minister of	a great nation , and yet see what a degrading occupation I am reduced to . " </s> " Listen to
<input checked="" type="checkbox"/>	the arm of	a most beautiful young girl , who wore chains of gold on her wrists and was evidently her slave . "
<input checked="" type="checkbox"/>	youngest was of	a very puny constitution , and scarce ever spoke a word , which made them take that for stupidity which was
<input checked="" type="checkbox"/>	a boy of	a bold temper , and took delight in hearing or reading of conjurers , giants , and fairies ; and used
<input checked="" type="checkbox"/>	the humor of	a great many others , who love wives to speak well , but think those very impudent who are continually doing
<input checked="" type="checkbox"/>	the mother of	a great many children , and of them all only one daughter was left . </s> But then she was worth

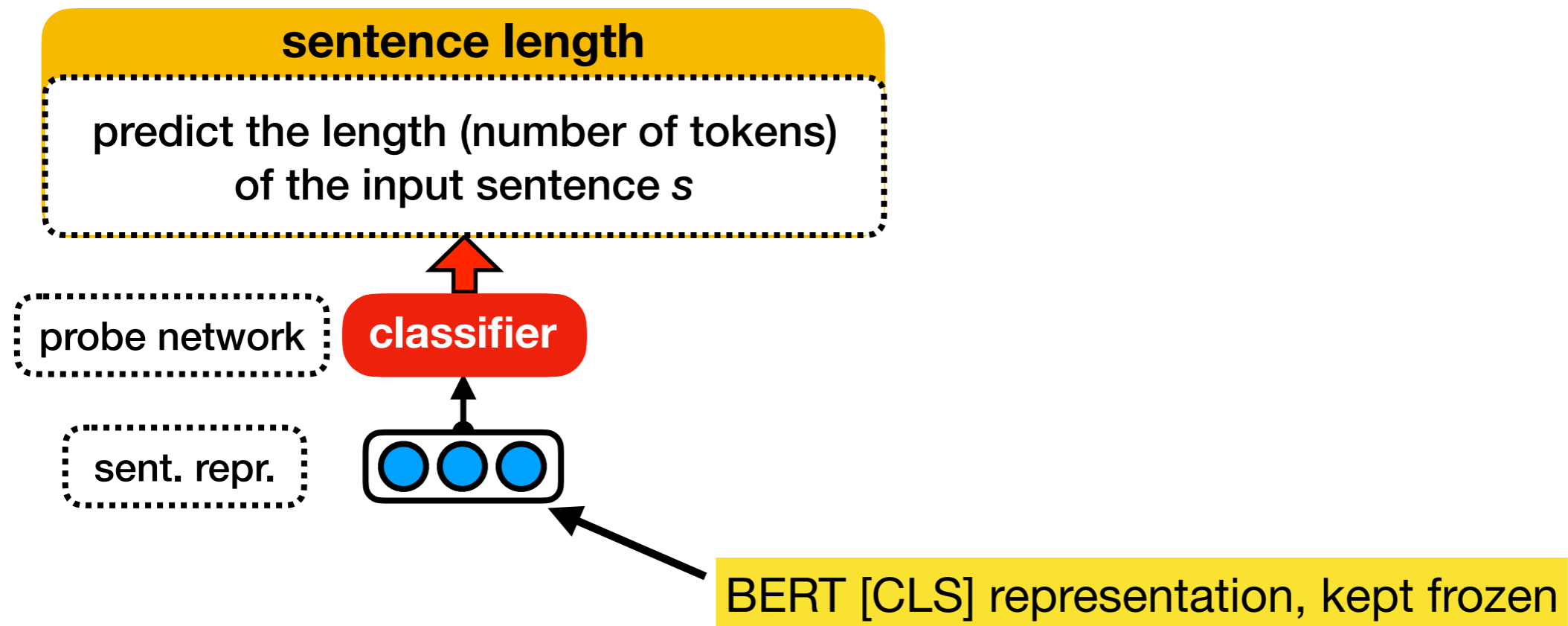


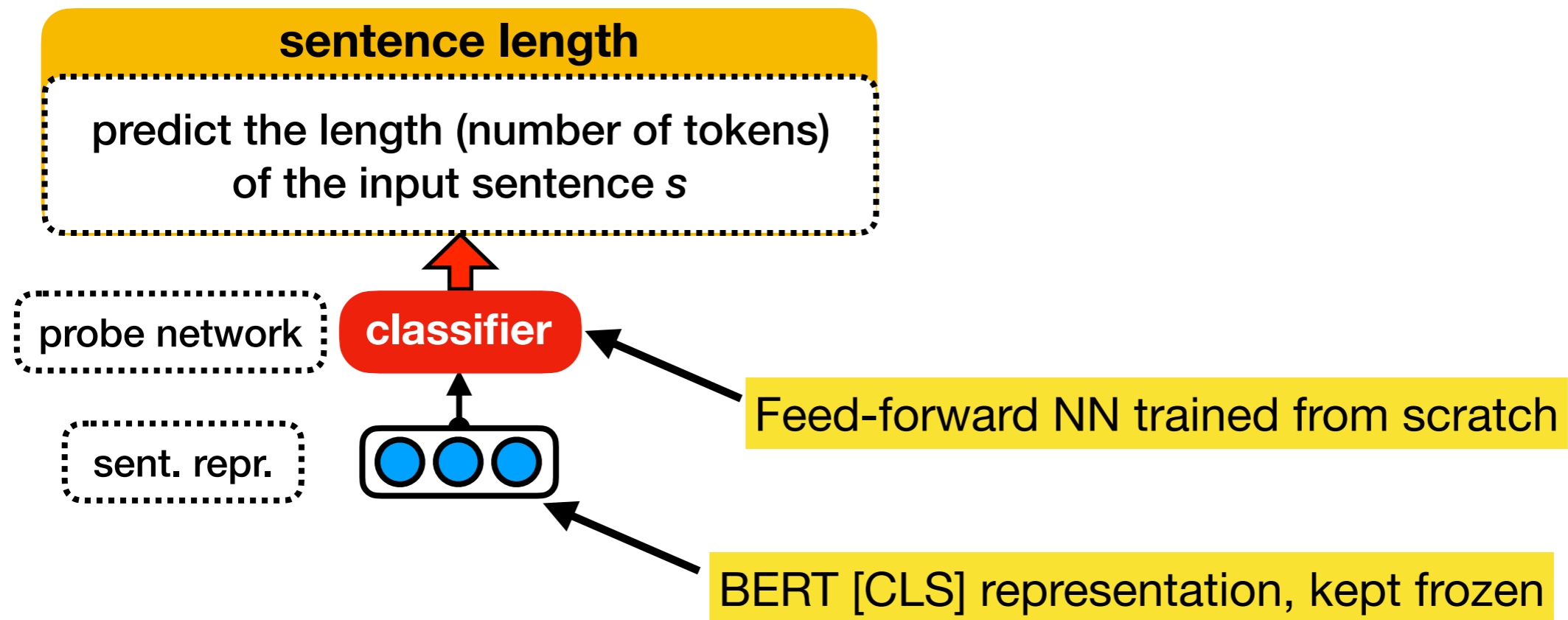


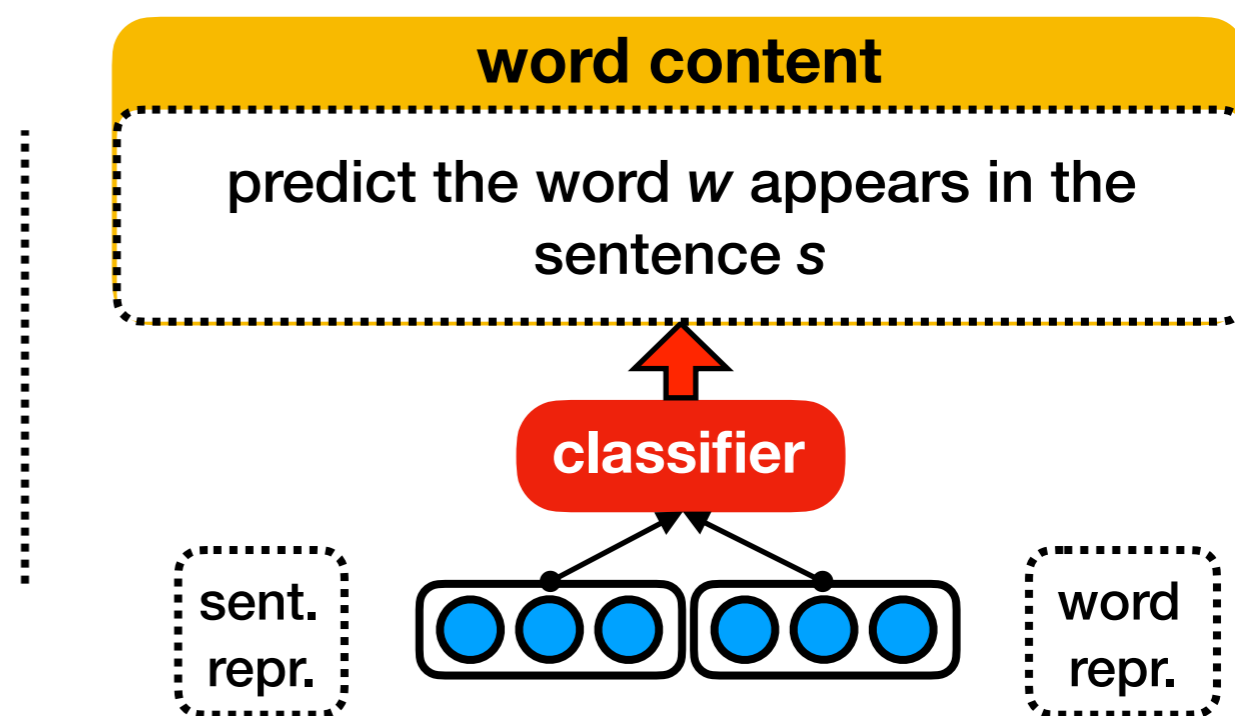
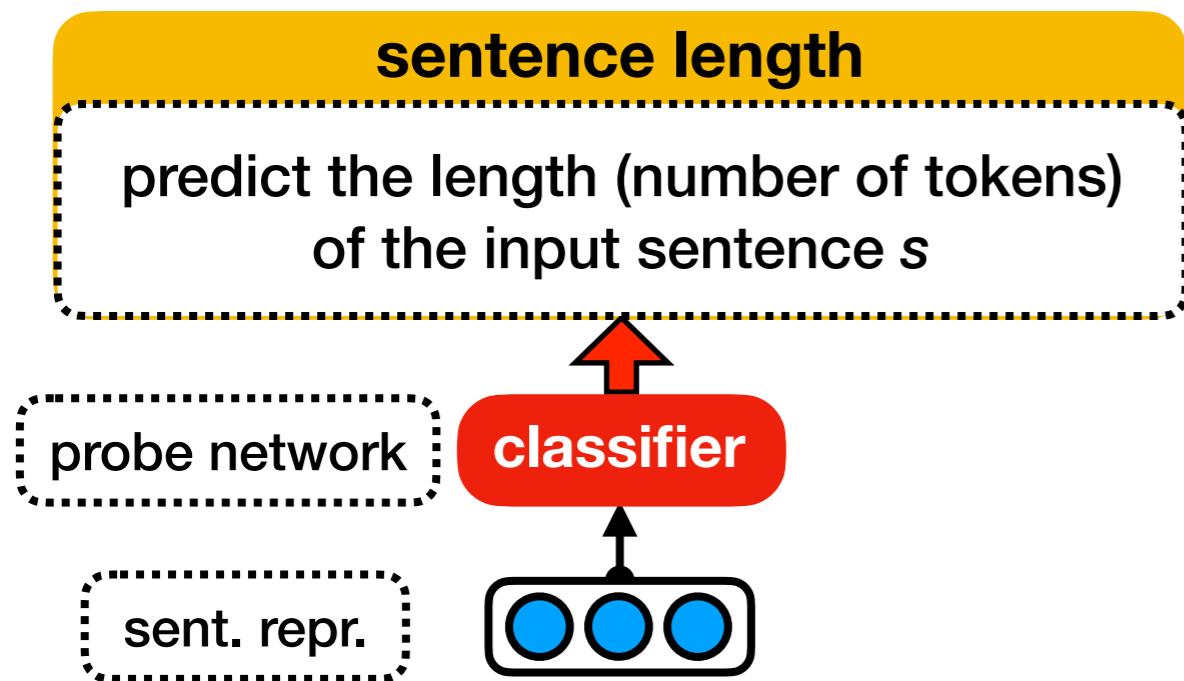
# what is a linguistic probe task?

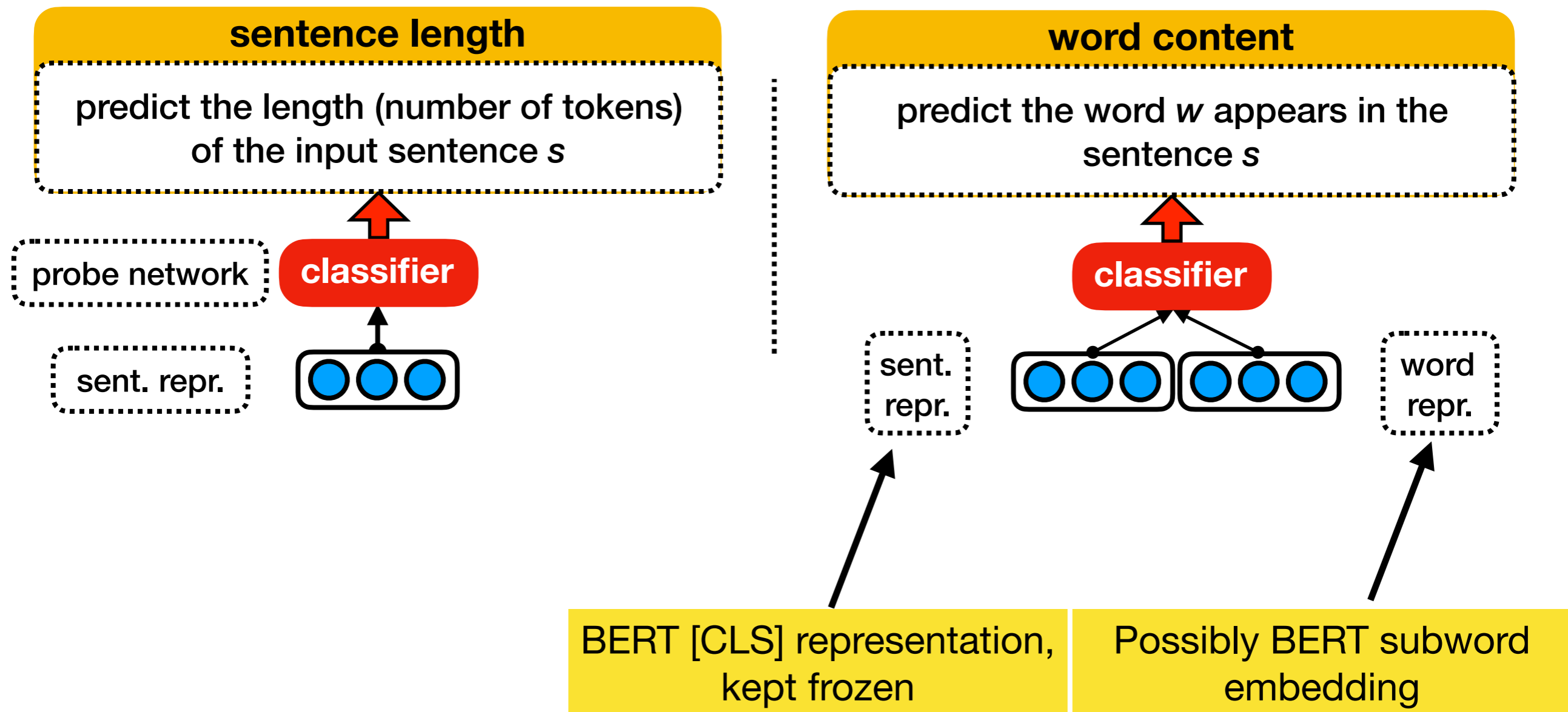
given an encoder model (e.g., BERT) pre-trained on a certain task, we use the representations it produces to train a classifier (without further fine-tuning the model) to predict a linguistic property of the input text



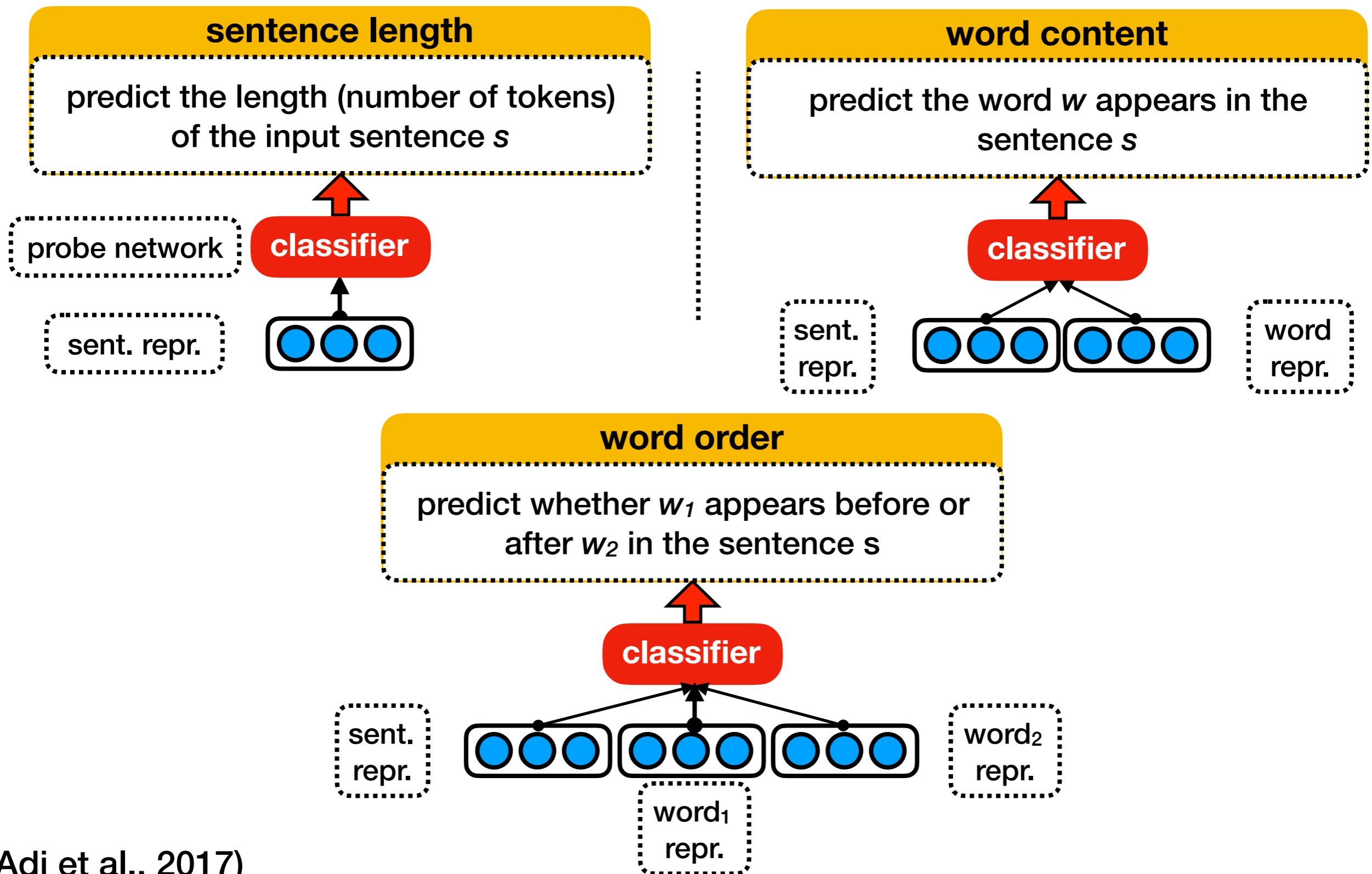




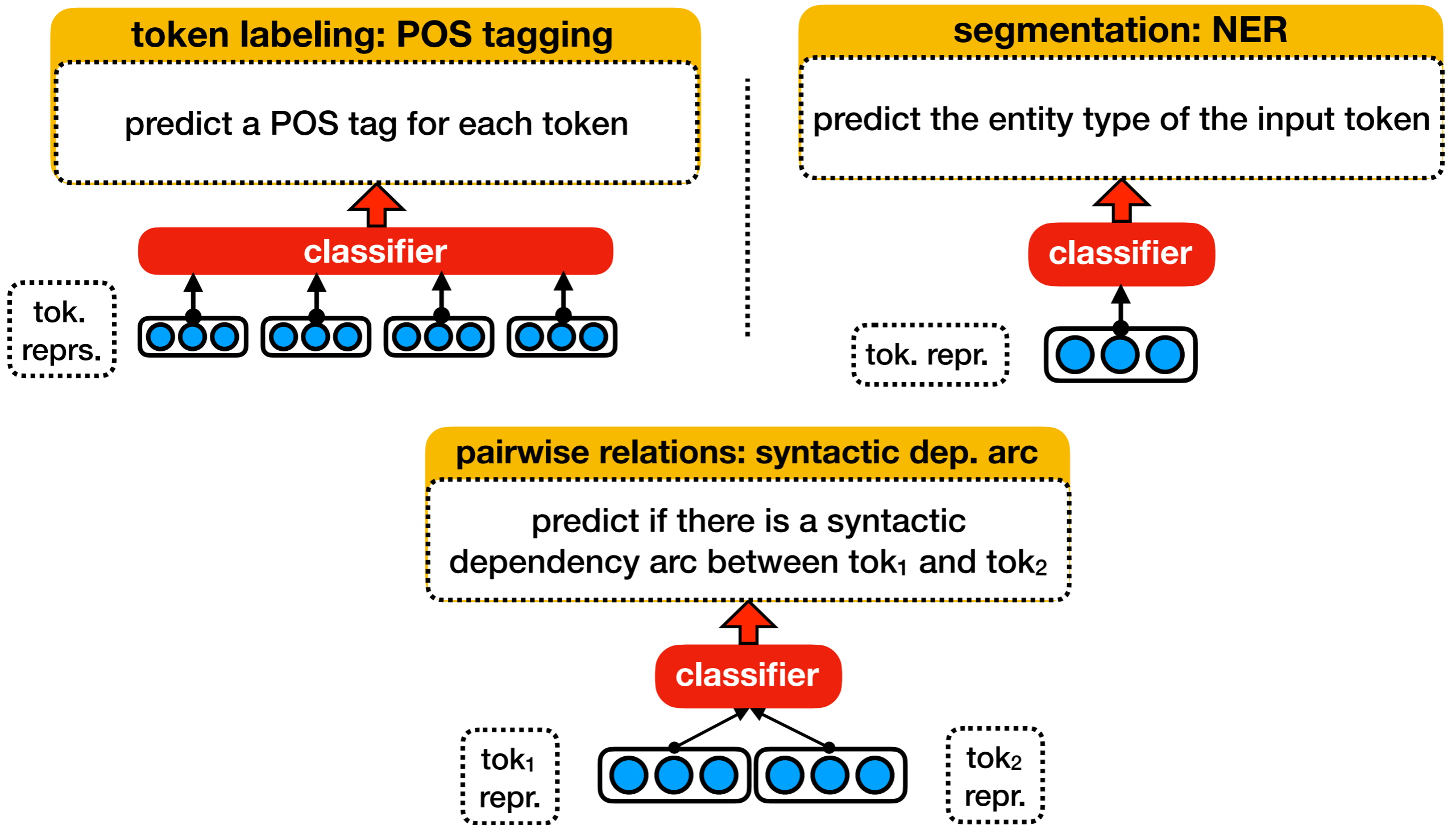








(Adi et al., 2017)



## edge probing: coreference

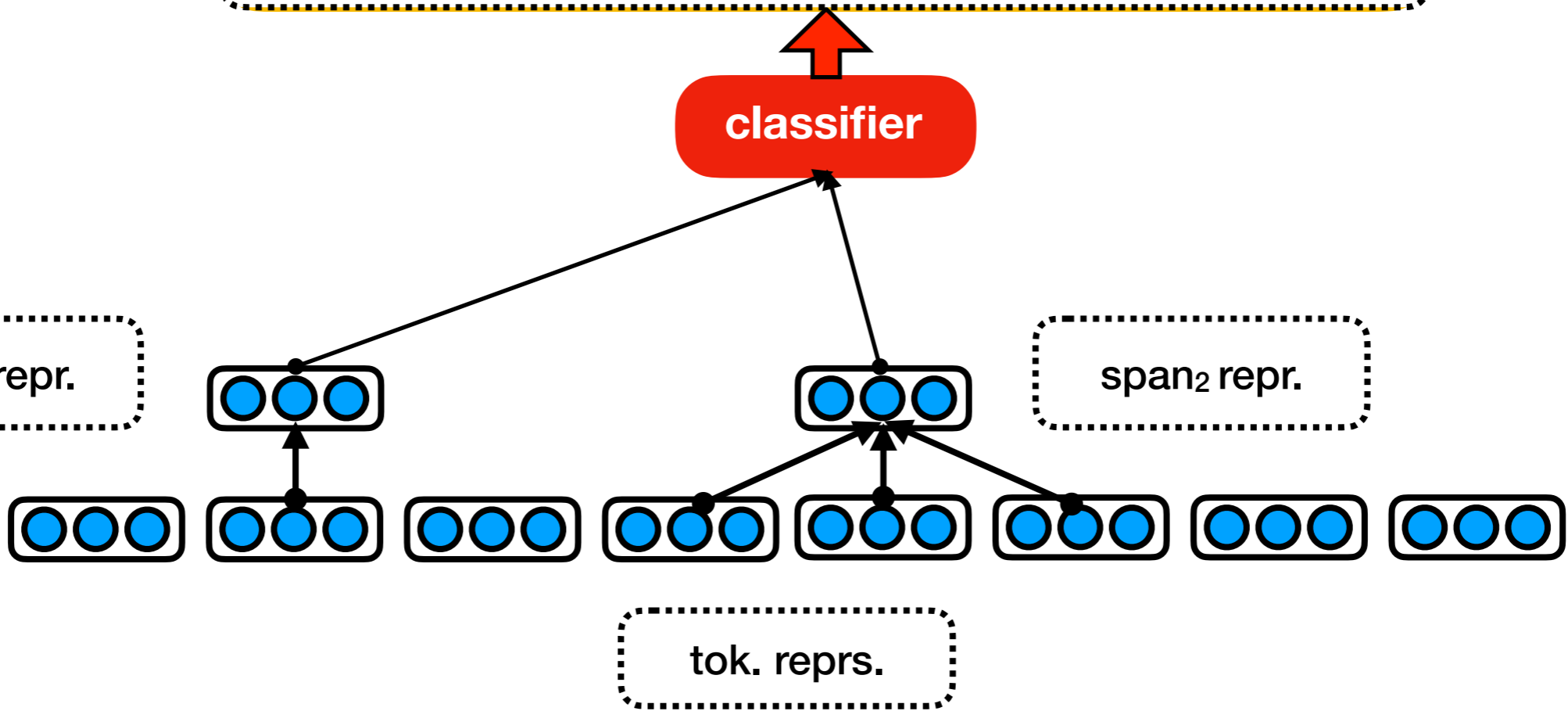
predict whether two spans of tokens (“mentions”) refer to the same entity (or event)

classifier

span<sub>1</sub> repr.

span<sub>2</sub> repr.

tok. reprs.



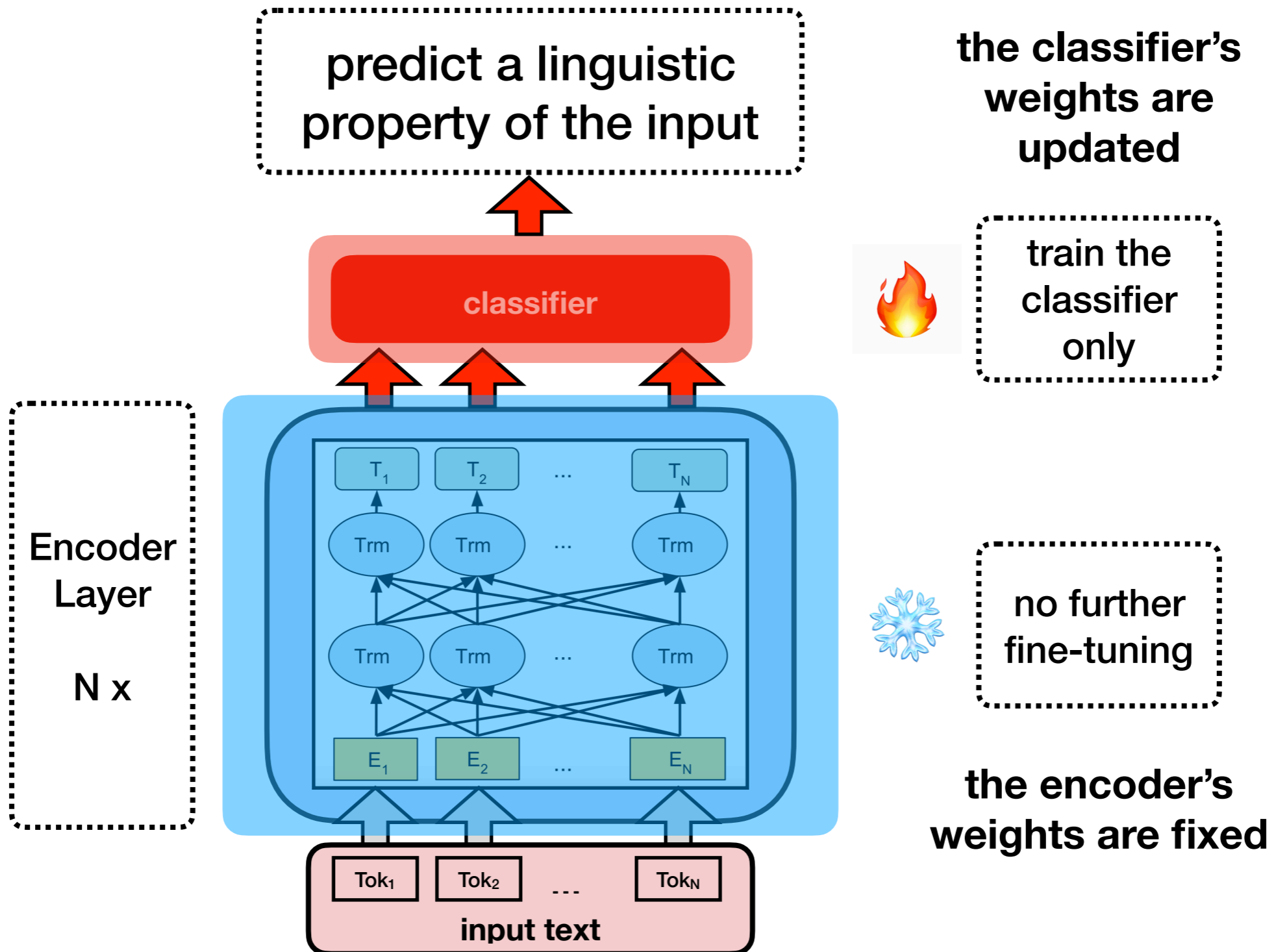
# motivation of probe tasks

- if we can train a classifier to predict a property of the input text based on its representation, it means the property is encoded somewhere in the representation
- if we cannot train a classifier to predict a property of the input text based on its representation, it means the property is not encoded in the representation or not encoded in a useful way, considering how the representation is likely to be used

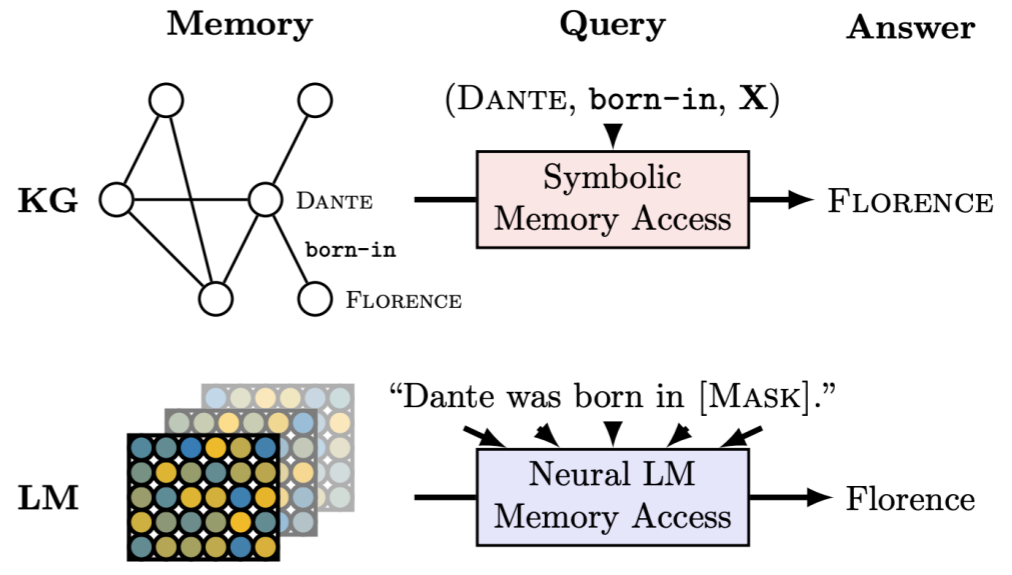
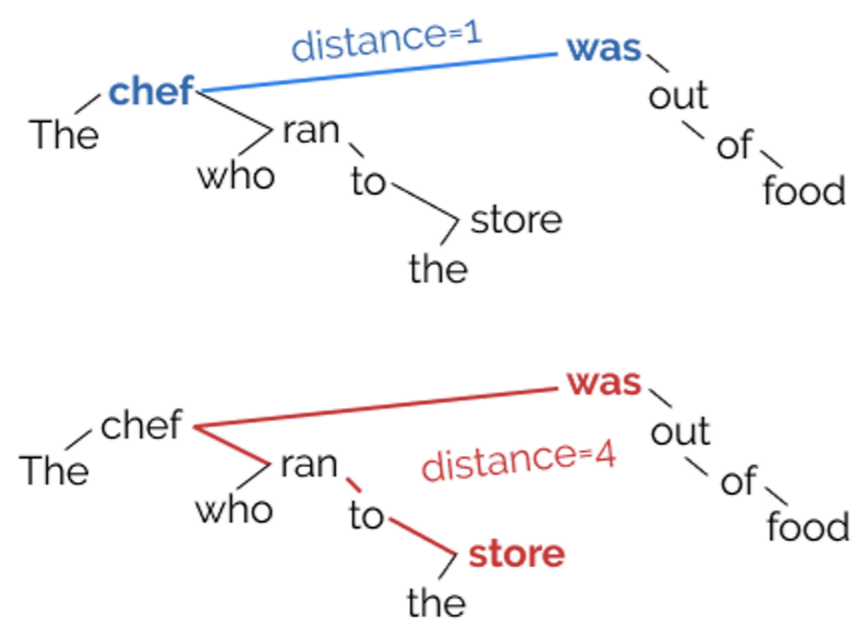
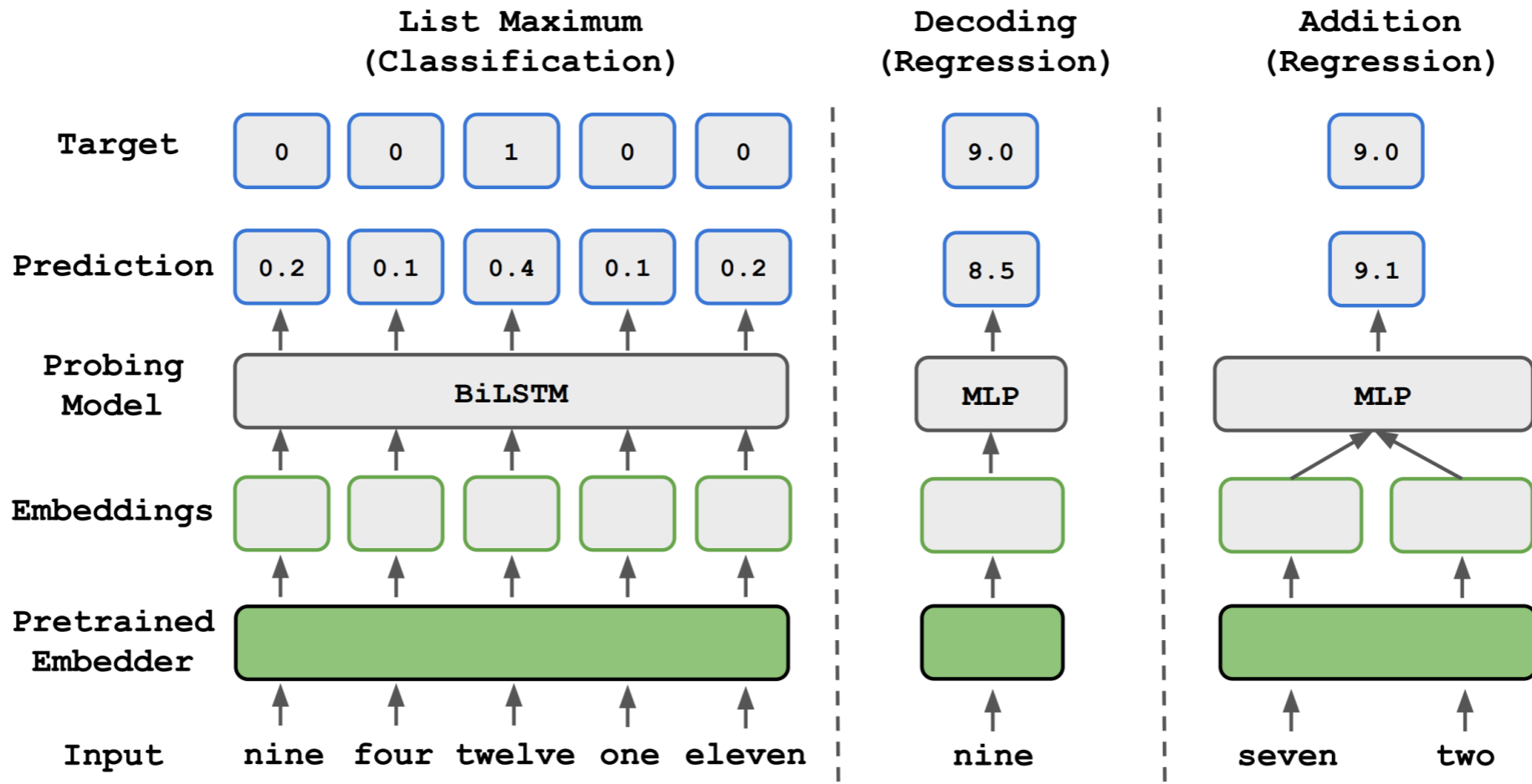
# characteristics of probe tasks

- usually classification problems that focus on simple linguistic properties
- ask simple questions, minimizing interpretability problems
- because of their simplicity, it is easier to control for biases in probing tasks than in downstream tasks
- the probing task methodology is agnostic with respect to the encoder architecture, as long as it produces a vector representation of input text
- does not necessarily correlate with downstream performance

# probe approach

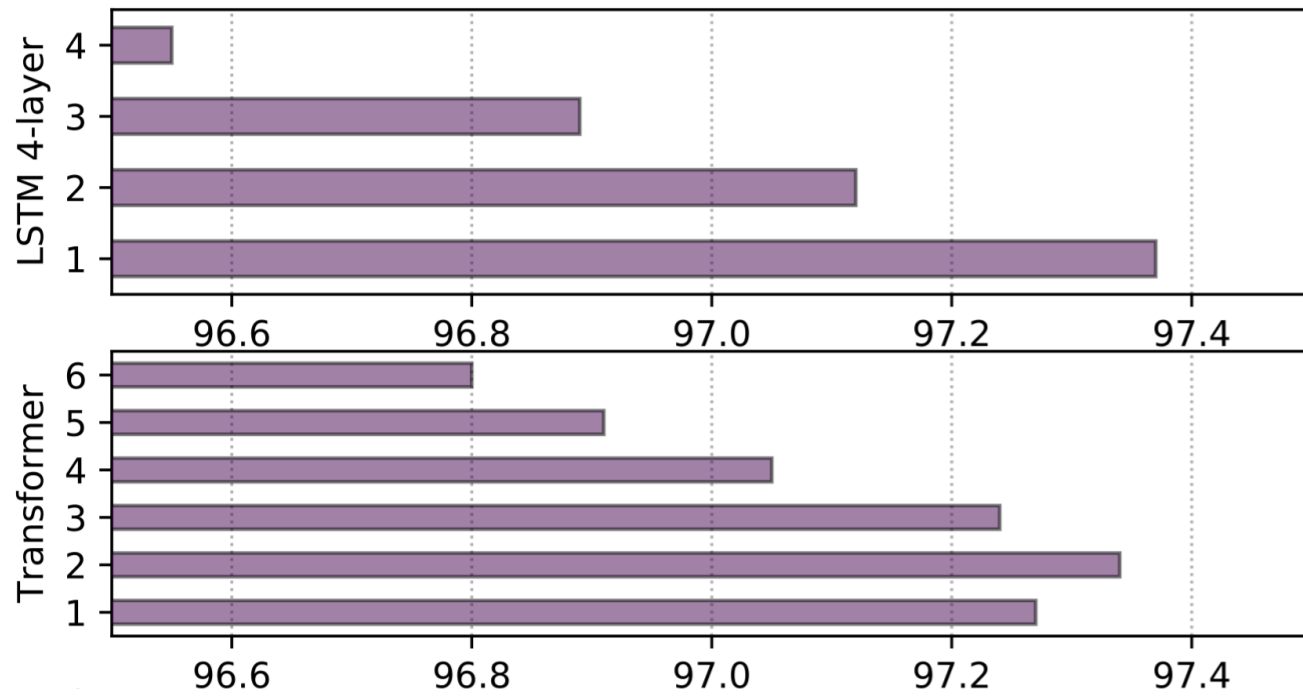




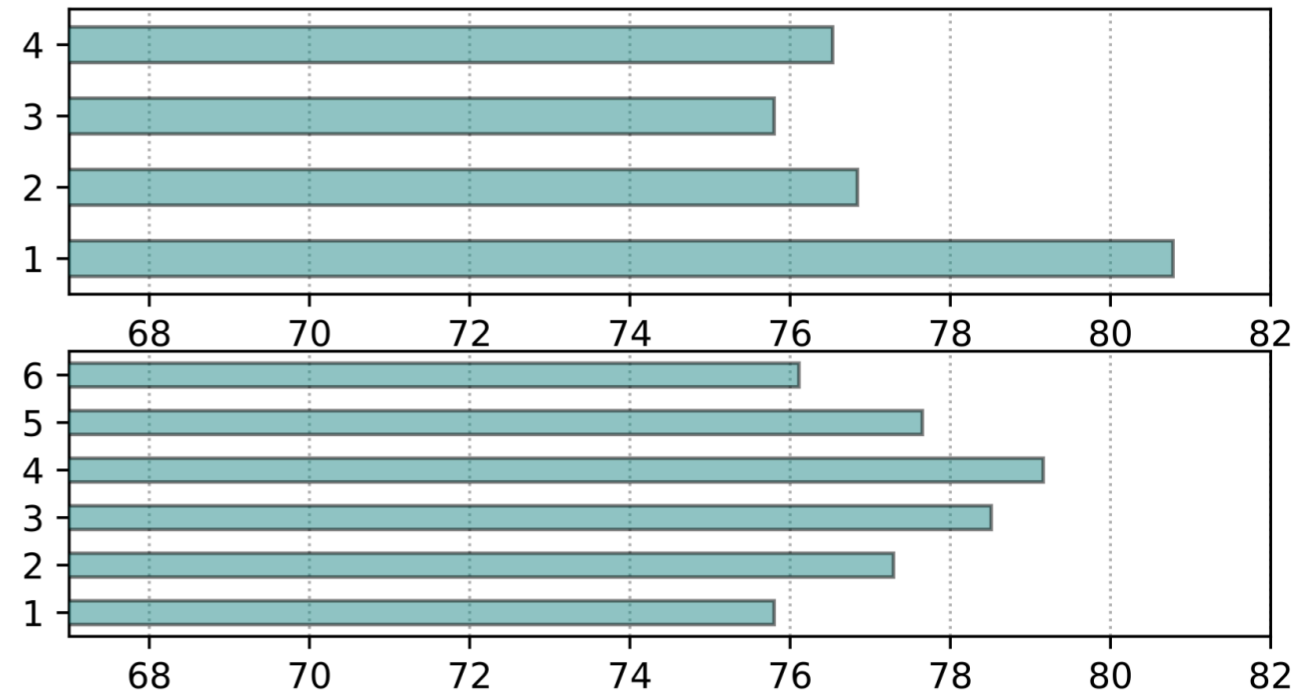


# lowest layers focus on local syntax, while upper layers focus more semantic content

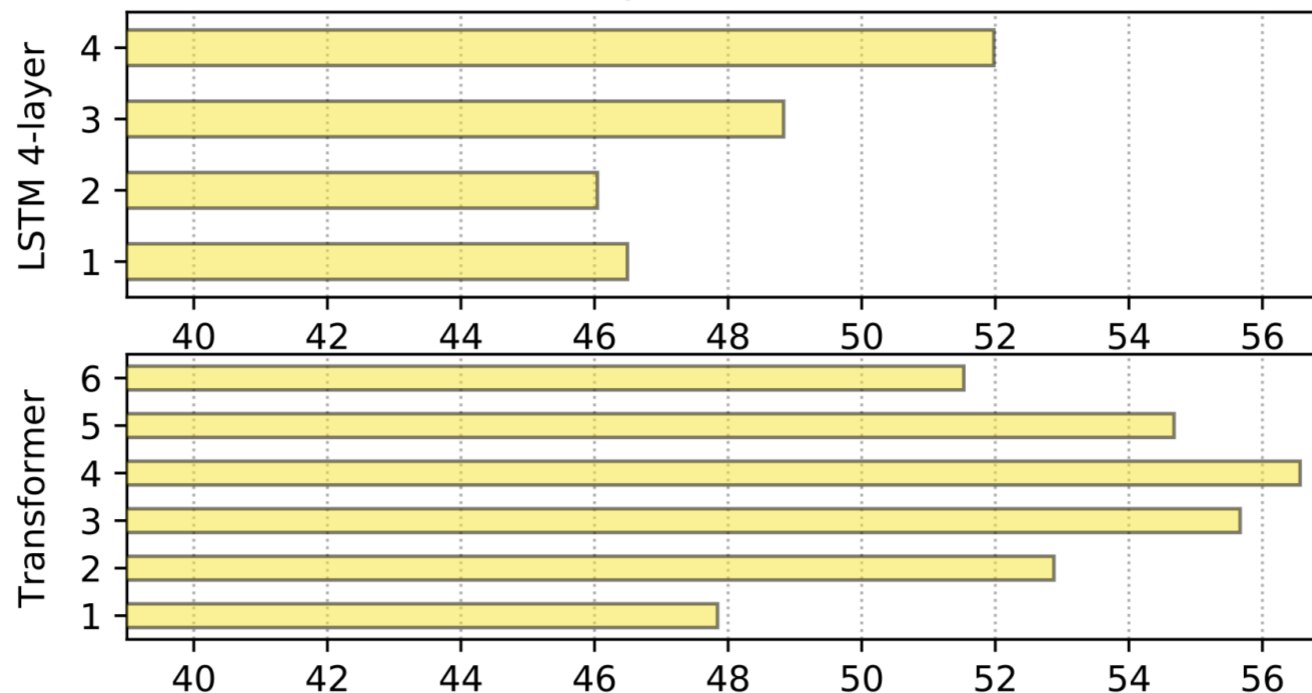
POS Tagging



Constituency parsing



Unsupervised coref.

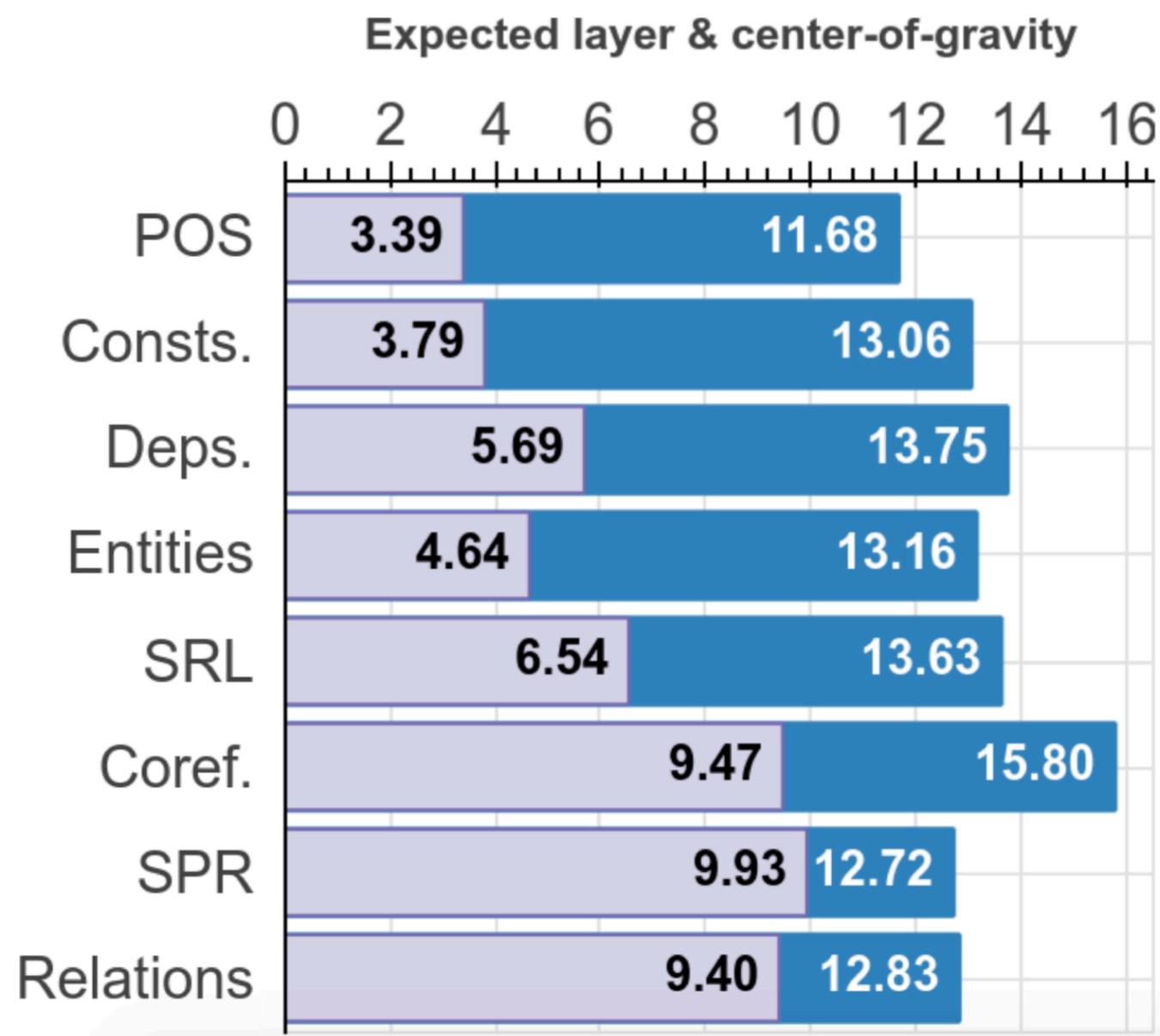


(Peters et al., 2018)

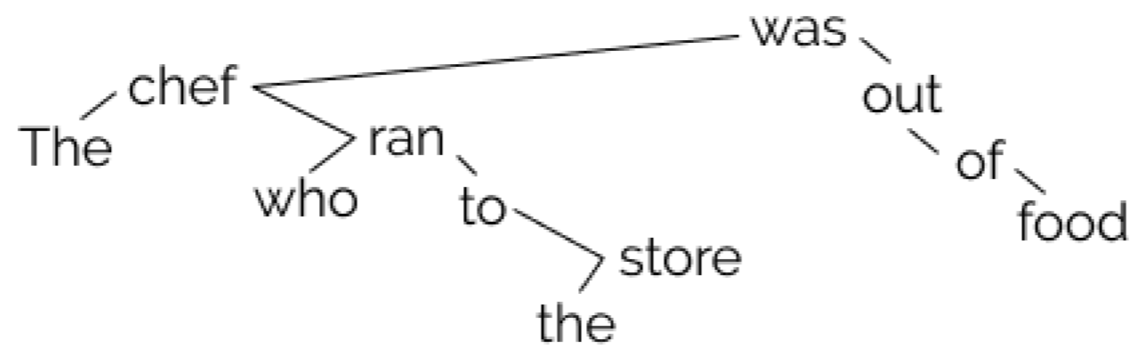
# BERT represents the steps of the traditional NLP pipeline: POS tagging → parsing → NER → semantic roles → coreference

the expected layer at which  
the probing model correctly  
labels an example

a higher center-of-gravity  
means that the information  
needed for that task is  
captured by higher layers



# does BERT encode syntactic structure?

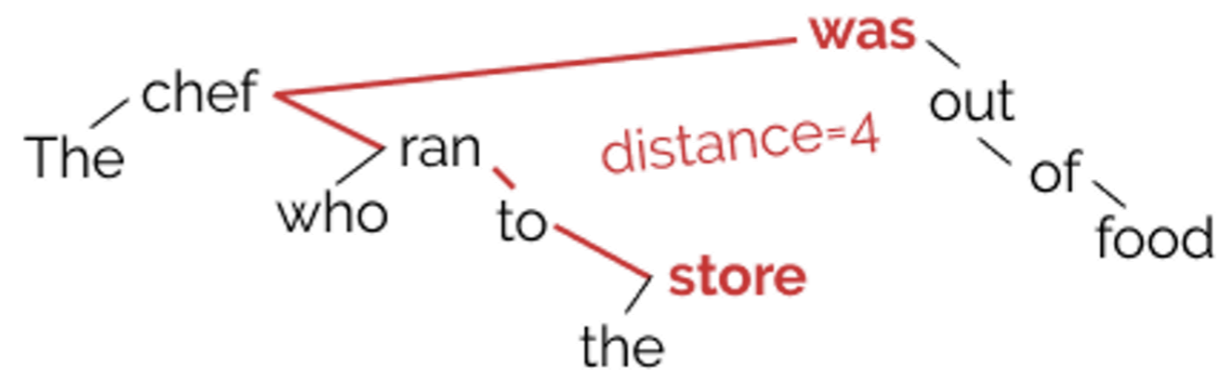
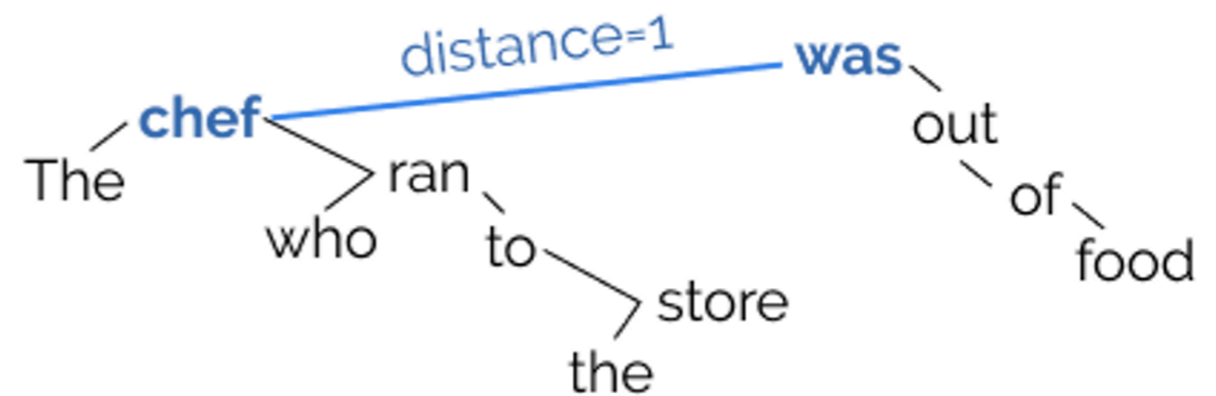


The chef who ran to the store was out of food

# understanding the syntax of the language may be useful in language modeling

The **chef** who ran to the **store**  
**was** out of food.

1. Because there was no food to be found, the chef went to the next store.
2. After stocking up on ingredients, the chef returned to the restaurant.



# how to probe for trees?

trees as distances and norms

the distance metric—the path length between each pair of words—recovers the tree  $T$  simply by identifying that nodes  $u, v$  with distance  $d_T(u, v) = 1$  are neighbors

the node with greater norm—depth in the tree—is the child



# a structural probe

- probe task 1 — distance:  
predict the path length between each given pair of words
- probe task 2 — depth/norm:  
predict the depth of a given word in the parse tree

# Yes, BERT knows the structure of syntax trees

---

Method	Distance		Depth	
	UUAS	DSpr.	Root%	NSpr.
ELMo1	77.0	0.83	86.5	0.87
BERTBASE7	79.8	0.85	88.0	0.87
BERTLARGE15	<b>82.5</b>	0.86	89.4	0.88
BERTLARGE16	81.7	<b>0.87</b>	<b>90.1</b>	<b>0.89</b>

---

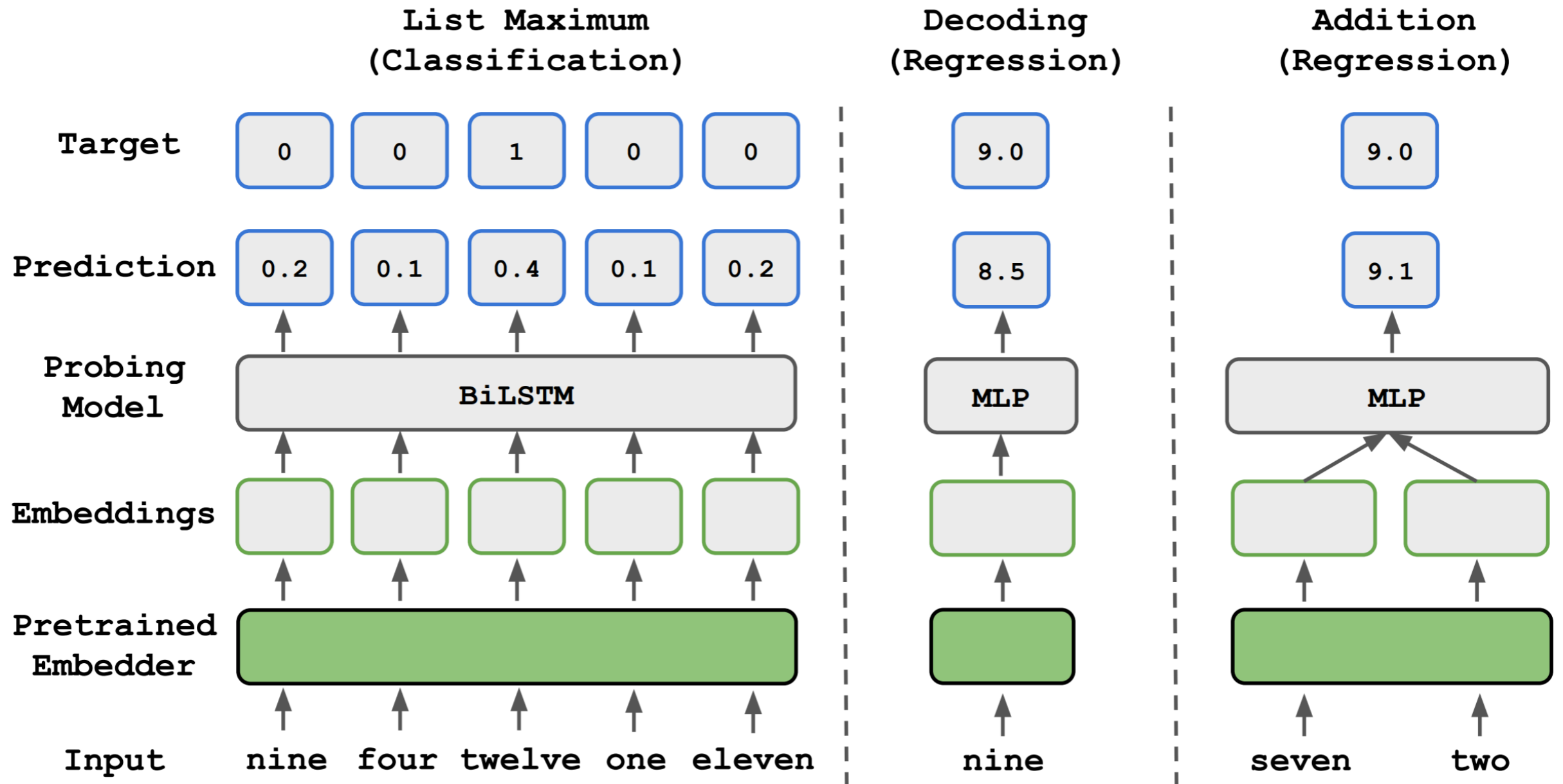
# does BERT know numbers?

25



what is the sum of eleven and fourteen?

# probing for numeracy



(Wallace et al., 2019)

# ELMo is actually better than BERT at this!

<b>Interpolation</b> <i>Integer Range</i>	<b>List Maximum (5-classes)</b>			<b>Decoding (RMSE)</b>			<b>Addition (RMSE)</b>		
	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]
Random Vectors	0.16	0.23	0.21	29.86	292.88	2882.62	42.03	410.33	4389.39
Untrained CNN	0.97	0.87	0.84	2.64	9.67	44.40	1.41	14.43	69.14
Untrained LSTM	0.70	0.66	0.55	7.61	46.5	210.34	5.11	45.69	510.19
<i>Pre-trained</i>									
Word2Vec	0.90	0.78	0.71	2.34	18.77	333.47	0.75	21.23	210.07
GloVe	0.90	0.78	0.72	2.23	13.77	174.21	0.80	16.51	180.31
ELMo	0.98	0.88	0.76	2.35	13.48	62.20	0.94	15.50	45.71
BERT	0.95	0.62	0.52	3.21	29.00	431.78	4.56	67.81	454.78

<b>Interpolation</b> <i>Float Range</i>	<b>List Maximum (5-classes)</b>	
	[0.0,99.9]	[0.0,999.9]
Rand. Vectors	0.18 ± 0.03	0.21 ± 0.04
ELMo	0.91 ± 0.03	0.59 ± 0.01
BERT	0.82 ± 0.05	0.51 ± 0.04
Char-CNN	0.87 ± 0.04	0.75 ± 0.03
Char-LSTM	0.81 ± 0.05	0.69 ± 0.02

<b>Interpolation</b> <i>Integer Range</i>	<b>List Maximum (5-classes)</b> [-50,50]
Rand. Vectors	0.23 ± 0.12
Word2Vec	0.89 ± 0.02
GloVe	0.89 ± 0.03
ELMo	0.96 ± 0.01
BERT	0.94 ± 0.02
Char-CNN	0.95 ± 0.07
Char-LSTM	0.97 ± 0.02

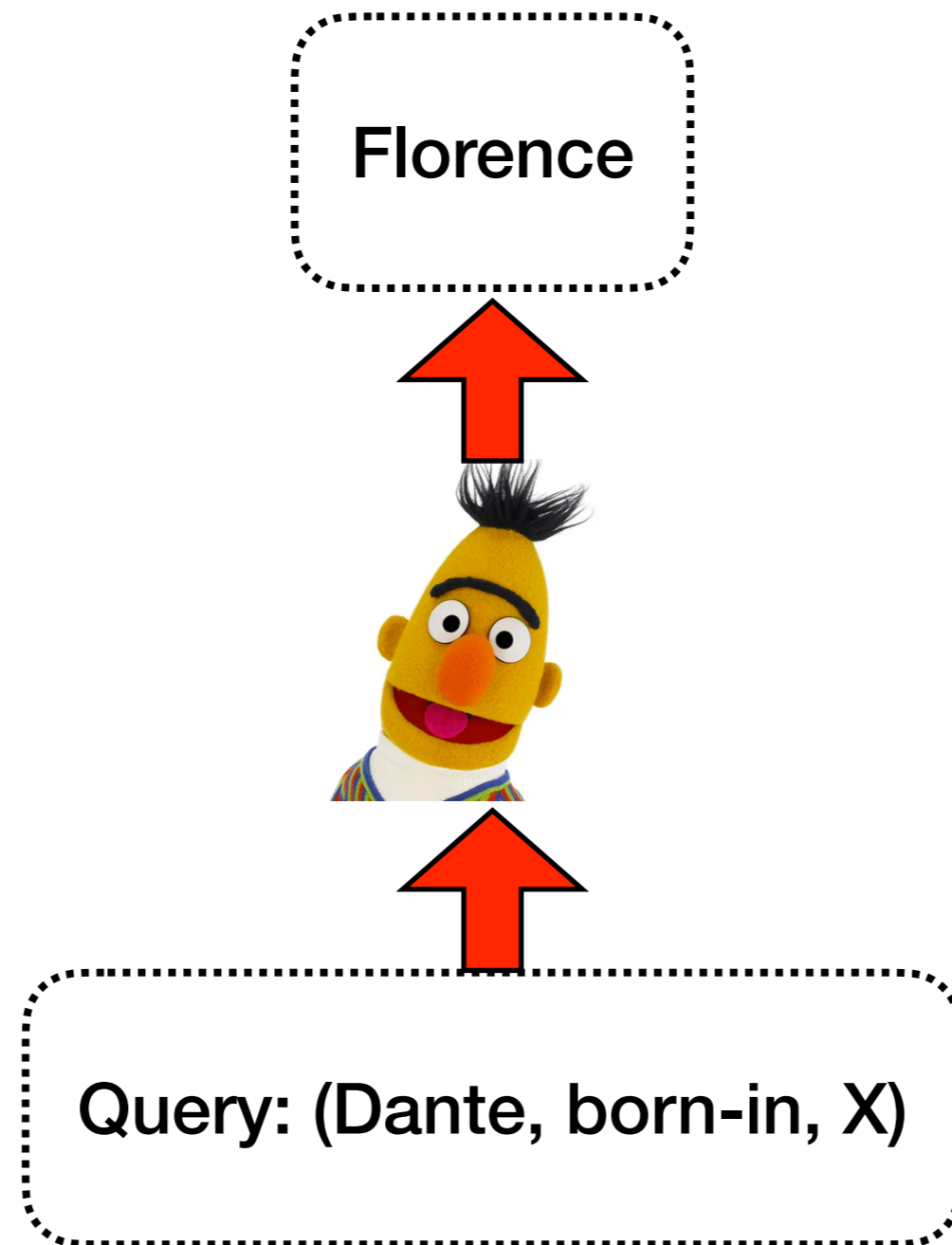
(Wallace et al., 2019)

# Why?

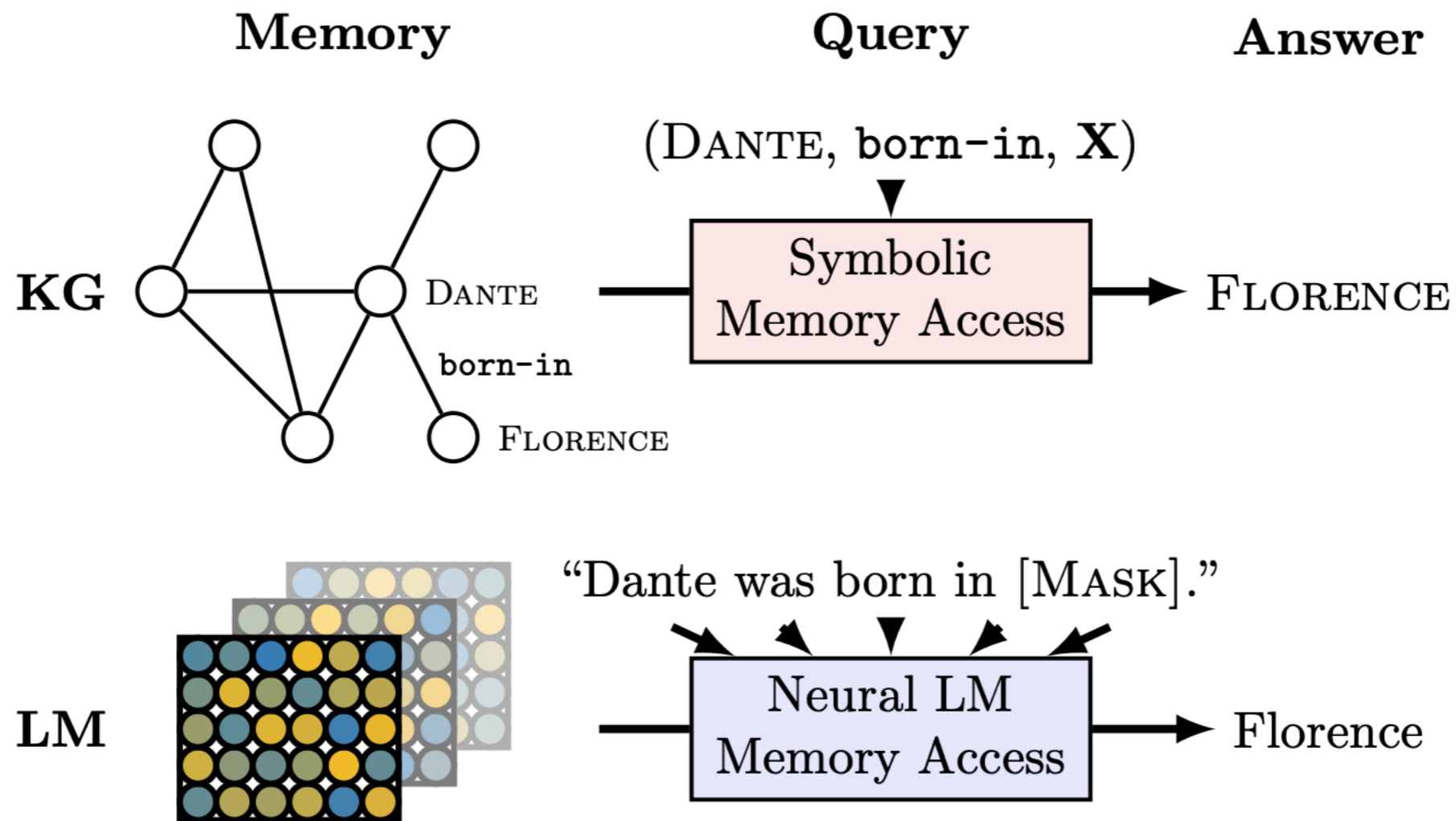
character-level CNNs are the best architecture for capturing numeracy

subword encoding is a poor method to encode digits, e.g., two numbers which are similar in value can have very different sub-word divisions

# Can BERT serve as a structured knowledge base?



# LAMA (LAnguage Model Analysis) probe





# LAMA (LAnguage Model Analysis) probe (cont.)

- manually define templates for considered relations, e.g., “[S] was born in [O]” for “place of birth”
- find sentences that contain both the subject and the object, then mask the object within the sentences and use them as templates for querying
- create cloze-style questions, e.g., rewriting “Who developed the theory of relativity?” as “The theory of relativity was developed by [MASK]”

# examples

	Relation	Query	Answer	Generation
T-Rex	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
	P527	Sodium sulfide consists of ____ .	sodium	water [-1.2], sulfur [-1.7], <b>sodium</b> [-2.5], zinc [-2.8], salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], <b>Labor</b> [-2.9]
	P530	Kenya maintains diplomatic relations with ____ .	Uganda	India [-3.0], <b>Uganda</b> [-3.2], Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
	P176	iPod Touch is produced by ____ .	Apple	<b>Apple</b> [-1.6], Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
	P30	Bailey Peninsula is located in ____ .	Antarctica	<b>Antarctica</b> [-1.4], Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
	P178	JDK is developed by ____ .	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
	P1412	Carl III used to communicate in ____ .	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____ .	Canada	<b>Canada</b> [-1.2], Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
ConceptNet	AtLocation	You are likely to find a overflow in a ____ .	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], <b>drain</b> [-3.6]
	CapableOf	Ravens can ____ .	fly	<b>fly</b> [-1.5], fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____ .	laugh	cry [-1.7], die [-1.7], <b>laugh</b> [-2.0], vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____ .	infection	disease [-1.2], cancer [-2.0], <b>infection</b> [-2.6], plague [-3.3], fever [-3.4]
	HasA	Birds have ____ .	feathers	wings [-1.8], nests [-3.1], <b>feathers</b> [-3.2], died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____ .	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], <b>speed</b> [-4.1]
	HasProperty	Time is ____ .	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____ .	alive	happy [-2.4], human [-3.3], <b>alive</b> [-3.3], young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____ .	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
UsedFor	A pond is for ____ .	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], <b>fish</b> [-2.8], recreation [-3.1]	

# BERT contains relational knowledge competitive with symbolic knowledge bases and excels on open-domain QA

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE <sub>n</sub>	RE <sub>o</sub>	Fs	Txl	Eb	E5B	Bb	B1
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	<b>16.1</b>
	birth-date	1825	1	1.9	-	0.0	<b>1.9</b>	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	<b>14.0</b>
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	<b>10.5</b>
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	<b>74.5</b>
	<i>N</i> -1	20006	23	23.85	-	5.4	<b>33.8</b>	6.1	18.0	3.6	6.5	32.4	34.2
	<i>N</i> - <i>M</i>	13096	16	21.95	-	7.7	<b>36.7</b>	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	<b>33.8</b>	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	<b>19.2</b>
SQuAD	Total	305	-	-	<b>37.5</b>	-	-	3.6	3.9	1.6	4.3	14.1	17.4

(Petroni et al., 2019)

# probe complexity

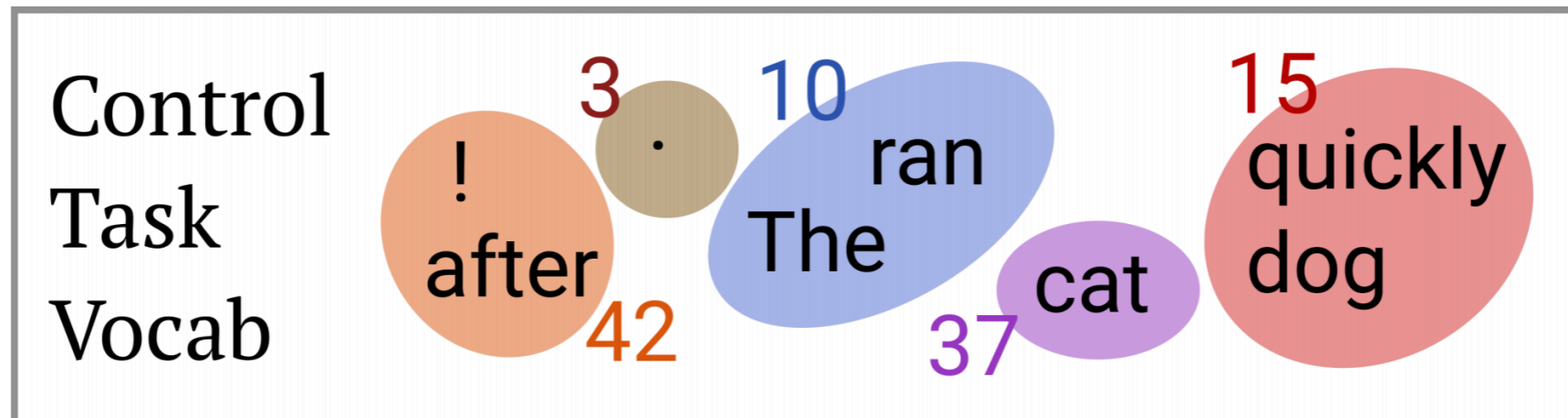
arguments for “simple” probes

we want to find easily accessible information  
in a representation

arguments for “complex” probes

useful properties might be encoded non-  
linearly

# control tasks



---

Sentence 1	The	cat	ran	quickly	.
<b>Part-of-speech</b>	DT	NN	VBD	RB	.
<b>Control task</b>	<b>10</b>	<b>37</b>	<b>10</b>	<b>15</b>	<b>3</b>

---

Sentence 2	The	dog	ran	after	!
<b>Part-of-speech</b>	DT	NN	VBD	IN	.
<b>Control task</b>	<b>10</b>	<b>15</b>	<b>10</b>	<b>42</b>	<b>42</b>

---

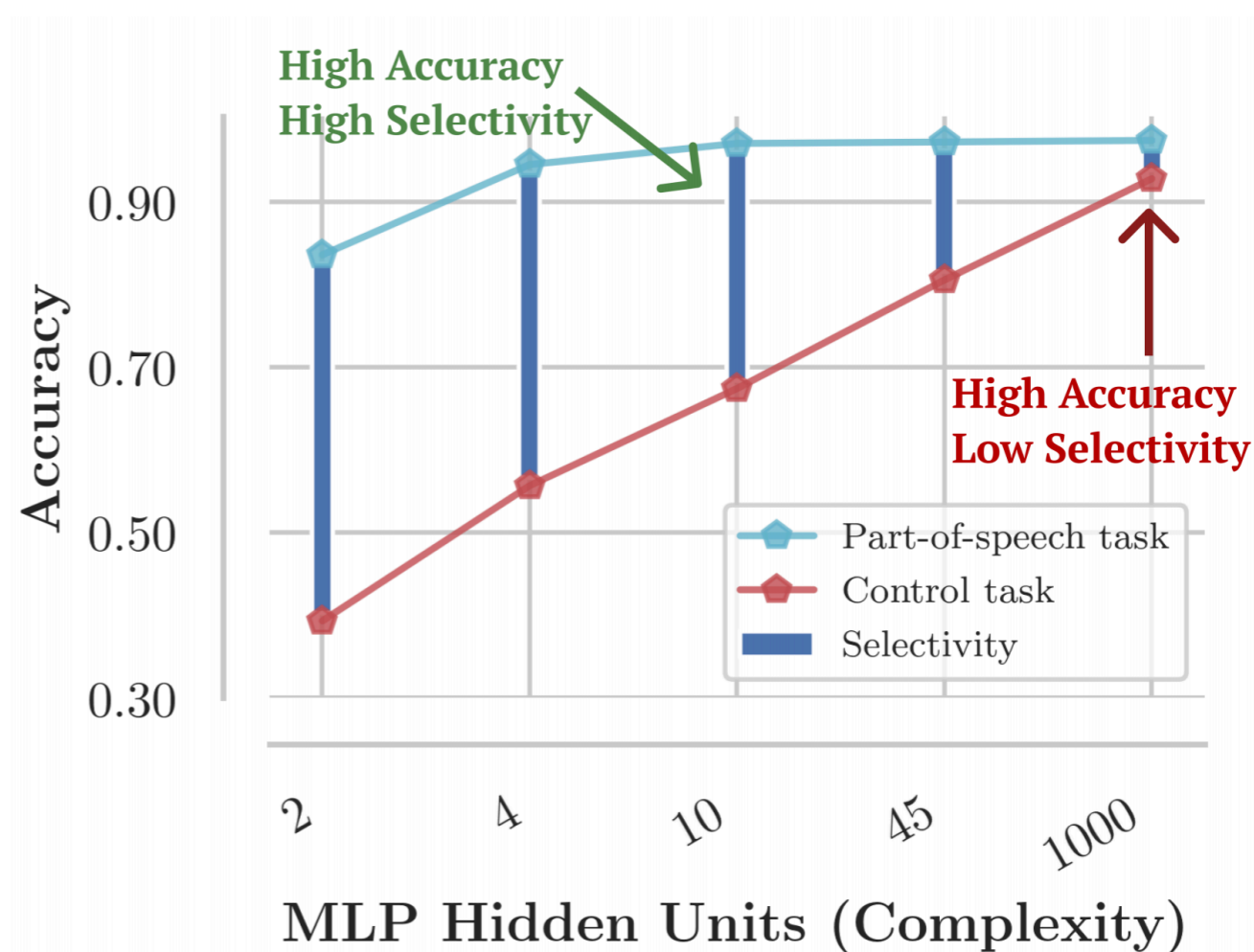
# designing control tasks

- independently sample a control behavior  $C(v)$  for each word type  $v$  in the vocabulary
- specifies how to define  $y_i \in Y$  for a word token  $x_i$  with word type  $v$
- *control task is a function that maps each token  $x_i$  to the label specified by the behavior  $C(x_i)$*

$$f_{\text{control}}(\mathbf{x}_{1:T}) = f(C(x_1), C(x_2), \dots, C(x_T))$$

# selectivity: high linguistic task accuracy + low control task accuracy

measures the probe model's ability to make output decisions independently of linguistic properties of the representation



# be careful about probe accuracies

---

## Part-of-speech Tagging

---

<b>Model</b>	Linear		MLP-1	
	Accuracy	Selectivity	Accuracy	Selectivity
Proj0	96.3	20.6	97.1	1.6
ELMo1	97.2	26.0	97.3	4.5
ELMo2	96.6	31.4	97.0	8.8

---



# how to use probe tasks to improve downstream task performance?

- what kinds of linguistic knowledge are important for your task?
- probe BERT for them
- if BERT struggles then fine-tune it with additional probe objectives

$$\mathcal{L}_{new} = \mathcal{L}_{BERT} + \alpha \mathcal{L}_{probe}$$

# example: KnowBERT

