Understanding Representations from Pre-trained Language Models

CS 585, Fall 2019

Introduction to Natural Language Processing

Tu Vu

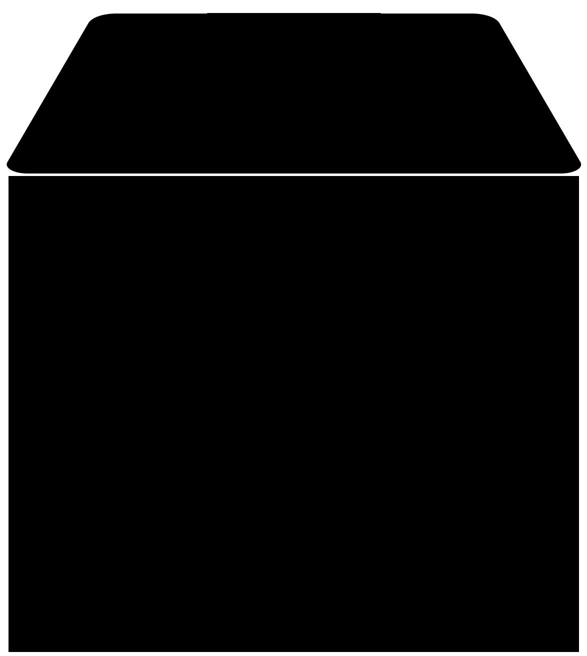
College of Information and Computer Sciences

University of Massachusetts Amherst

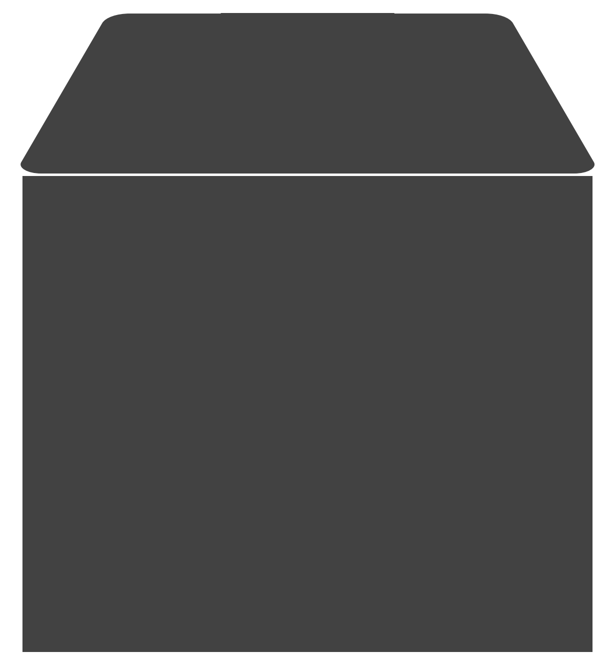
Nov 14, 2019

what kinds of knowledge are encoded in BERT?

before the lecture



after the lecture



overview

- ★ BERT News!
- ★ BERTology
- ★ understanding contextualized representations
  - linguistic probe tasks

overview

★ BERT News!

★ BERTology

★ understanding contextualized representations

- linguistic probe tasks

BERT News!

T5: Text-to-Text Transfer Transformer

new state-of-the-art results on many NLP benchmarks

(Raffel et al., 2019)

GLUE

Rank	Name	Model	URL	Score
1	T5 Team - Google	T5		89.7
2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4
3	王玮	ALICE v2 large ensemble (Alibaba D		
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		
5	Facebook AI	RoBERTa		
6	XLNet Team	XLNet-Large (ensemble)		
7	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		
8	GLUE Human Baselines	GLUE Human Baselines		

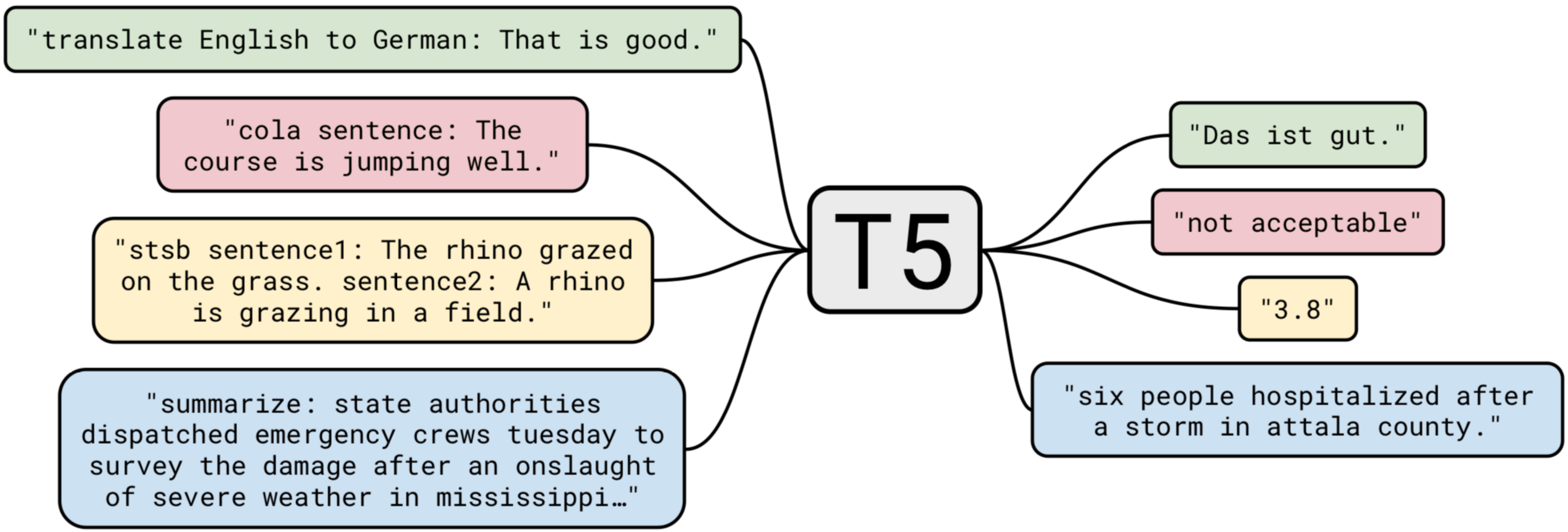
SuperGLUE

Rank	Name	Model	URL	Score
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8
2	T5 Team - Google	T5		88.9
3	Facebook AI	RoBERTa		84.6
4	IBM Research AI	BERT-mtl		73.5
5	SuperGLUE Baselines	BERT++		71.5
		BERT		69.0

BERT News!

T5: key ideas

1) treat every NLP problem as a “text-to-text” problem, one seq2seq model to learn them all

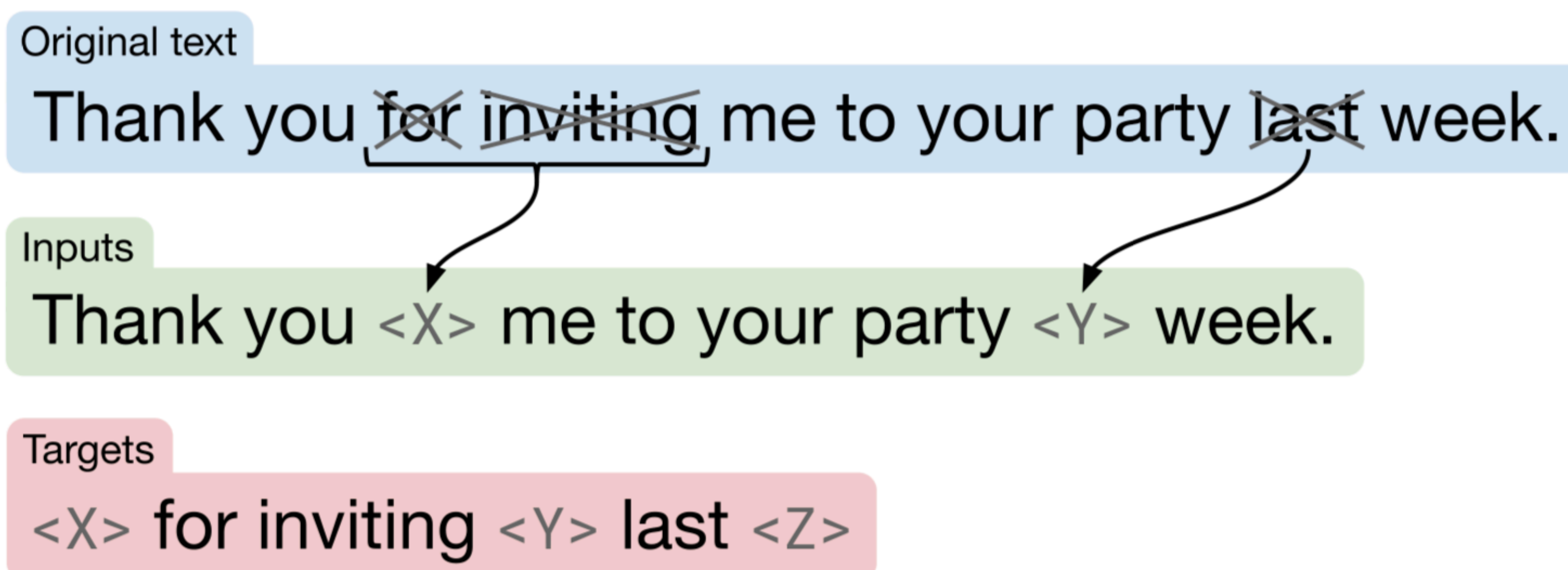


(Raffel et al., 2019)

BERT News!

T5: key ideas

2) a denoising objective results in better downstream task performance



(Raffel et al., 2019)

BERT News!

T5: key ideas

3) larger model on more data, insane scale!

- 11 billion parameters
    - ~31x as large as RoBERTa (355 million parameters)
    - ~33x as large as BERT (335 million parameters)
- 750GB text ~ 190 billion words?
    - ~5x as much as RoBERTa (160GB)
    - ~60x as much as BERT (13GB, 3.3 billion words)

(Raffel et al., 2019)

BERT News!

other models

BART

- denoising autoencoder for pretraining sequence-to-sequence models
- sentence shuffling + text infilling
- comparable to RoBERTa on GLUE and SQuAD, state-of-the-art results on abstractive dialogue, question answering, and summarization

(Lewis et al., 2019)

BERT News!

other models (cont.)

XLM-R

- XLM + RoBERTa
- 2.5TB of text from 100 languages!
- state-of-the-art results on cross-lingual benchmarks
- comparable to XLNet on GLUE

(Conneau et al., 2019)

BERT News!

a super competitive area

dozens of new BERT models every month

not only NLP, but also CV

things change shortly

Xin (Eric) Wang
@xwang_lk

A list of V*BERT papers:
VideoBERT: arxiv.org/abs/1904.01766
ViLBERT: arxiv.org/abs/1904.01766
LXMERT: arxiv.org/abs/1908.07490
VisualBERT: arxiv.org/abs/1908.03557
Unicoder-VL: arxiv.org/abs/1908.06066
B2T2: arxiv.org/abs/1908.05054
VL-BERT: arxiv.org/abs/1908.08530
... ...

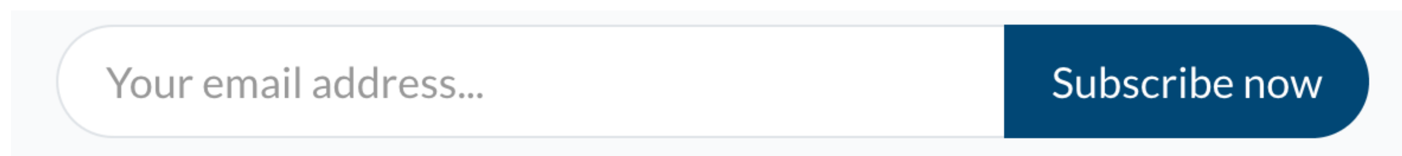
Jason Phang
@zhansheng

Replying to @zhansheng and @sleepinyourhat

BERT on STILTs was also the SOTA (82.0) on GLUE for a very brief 6 hours because this is NLP in 2019 😛

BERT News!

stay up-to-date

- NLP progress
  https://github.com/sebastianruder/NLP-progress
- NLP News!
  

  http://newsletter.ruder.io/
- arXiv, ACL Anthology
- Twitter (the best)

overview

- ★ BERT News!
- ★ BERTology
- ★ understanding contextualized representations
  - linguistic probe tasks

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?

BERTology

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

BERTology

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 el al., BlackBoxNLP 2019

substantial syntactic information is captured in BERT's attention

BERTology

tools & examples

AllenNLP Interpret
https://allennlp.org/interpret

Ai2 Allen Institute for AI 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

overview

★ BERT News!

★ BERTology

★ understanding contextualized representations

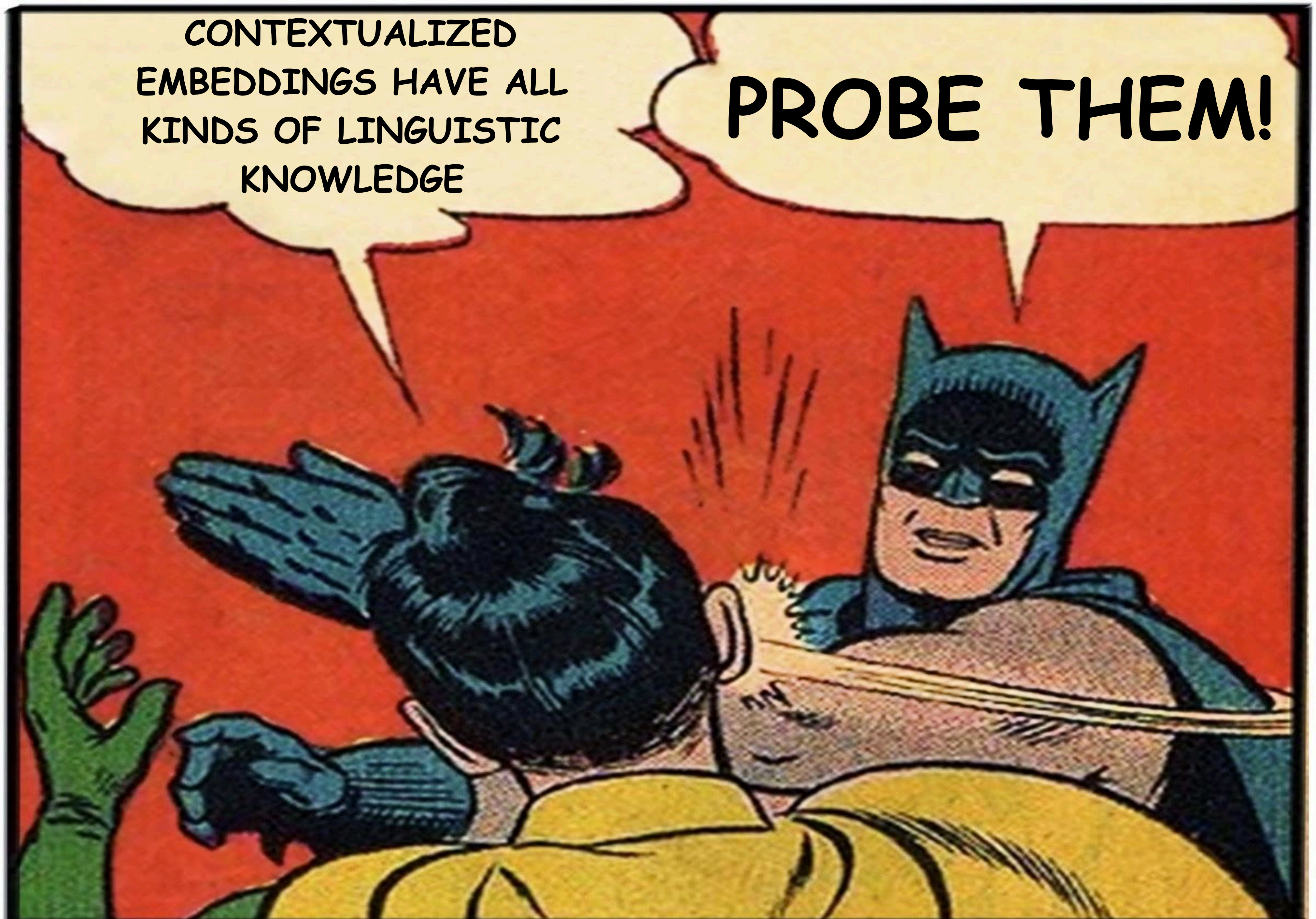
- linguistic probe tasks

understanding contextualized representations

two most prominent methods

- visualization
- linguistic probe tasks

linguistic probe tasks

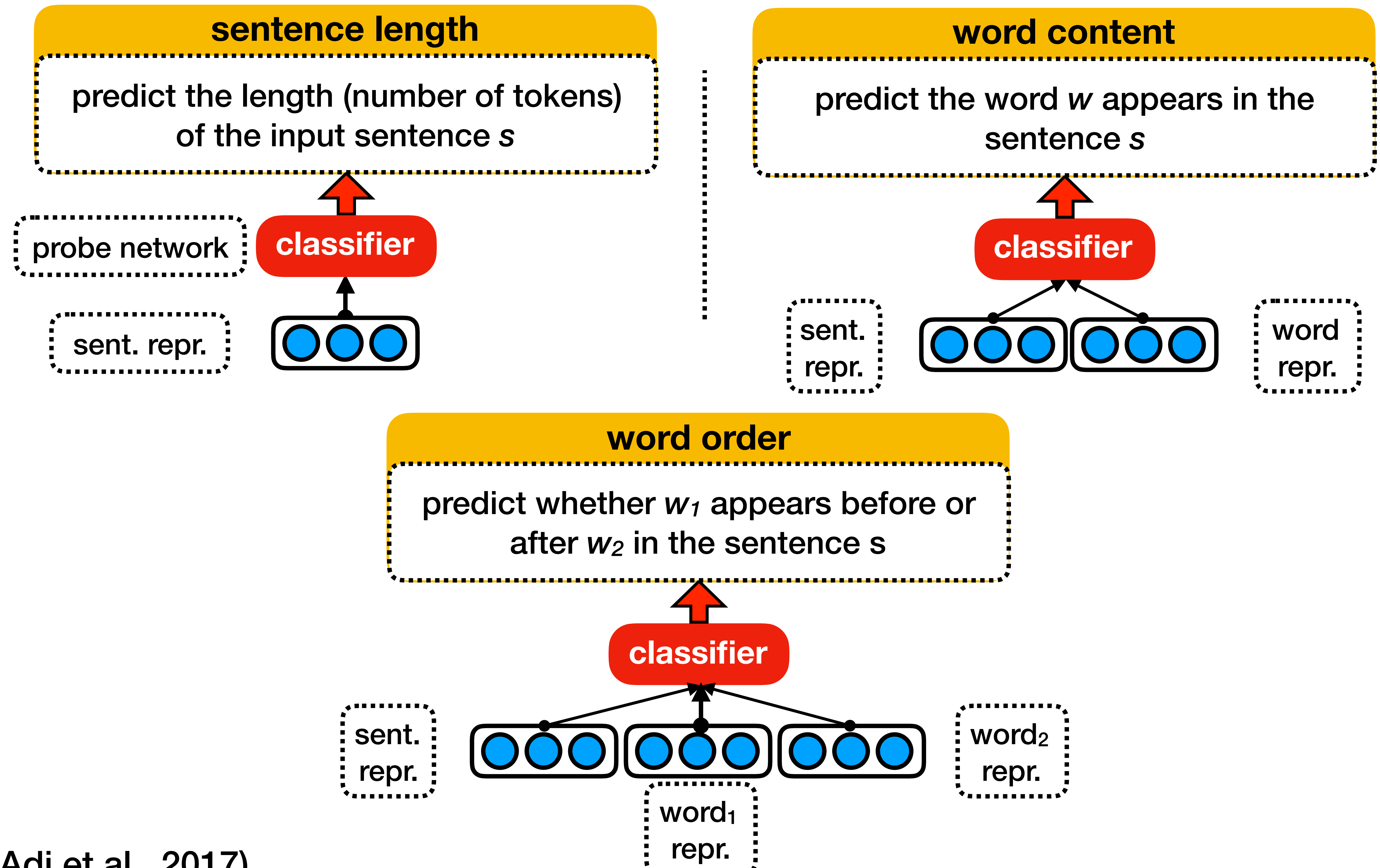


Credit: Alexis Conneau

what is a linguistic probe task?

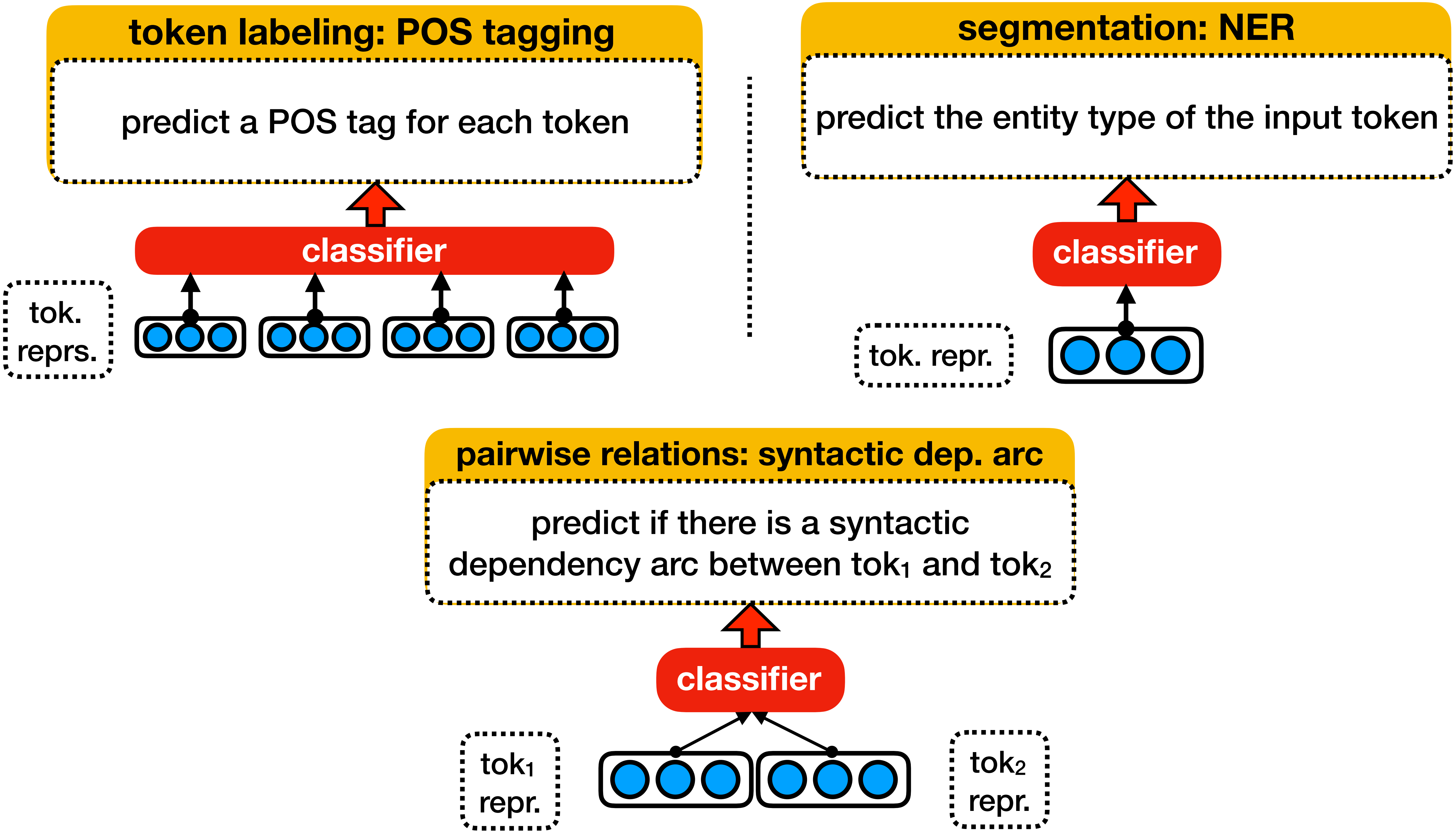
given an encoder model (e.g., BERT) pre-trained on a certain task, we use the representations it produces to train a classifier (without further fine-tuning the model) to predict a linguistic property of the input text

example 1



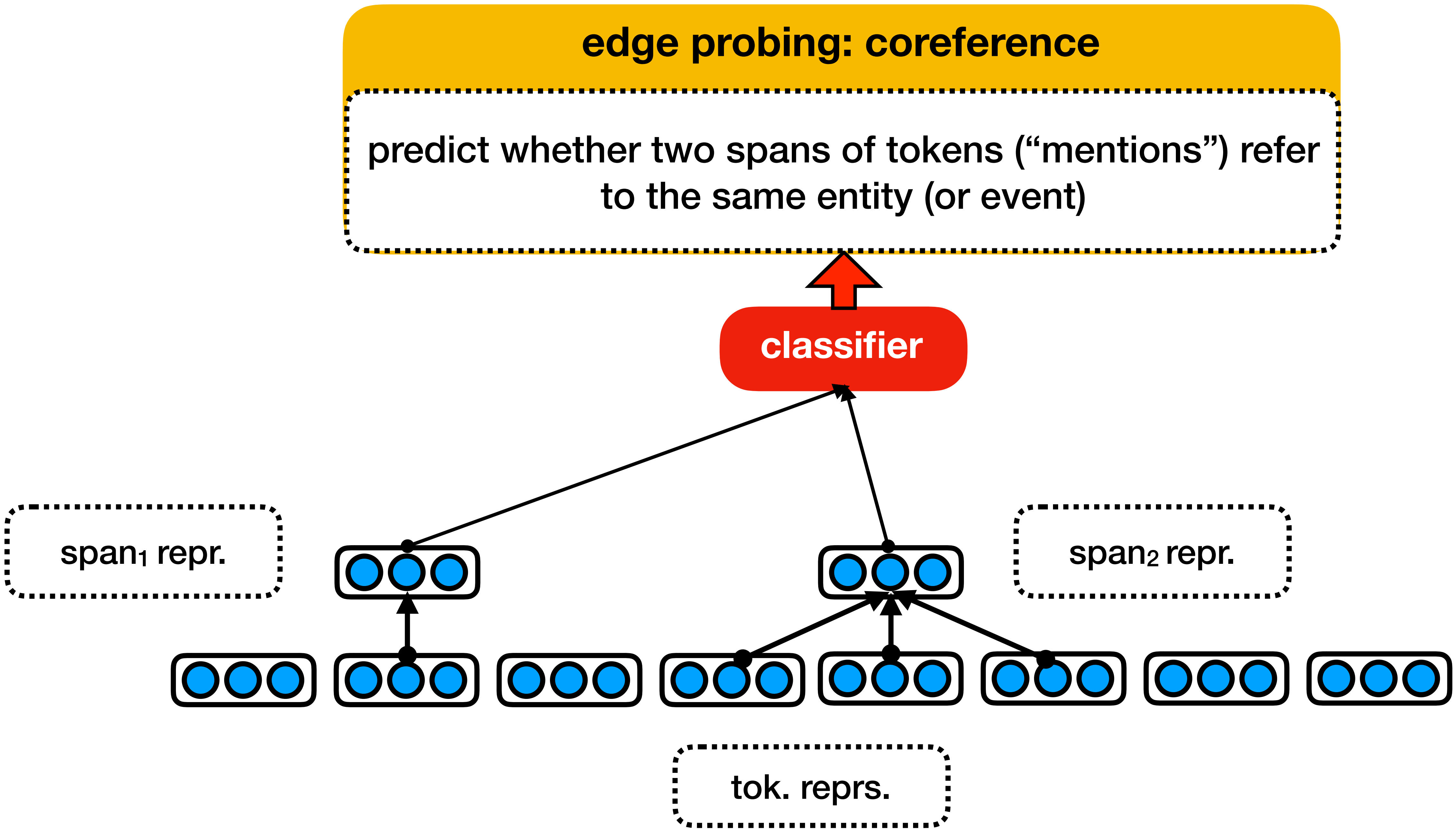
(Adi et al., 2017)

example 2



(Liu et al., 2019)

example 3



(Tenney et al., 2019)

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 in the representation in a readable way

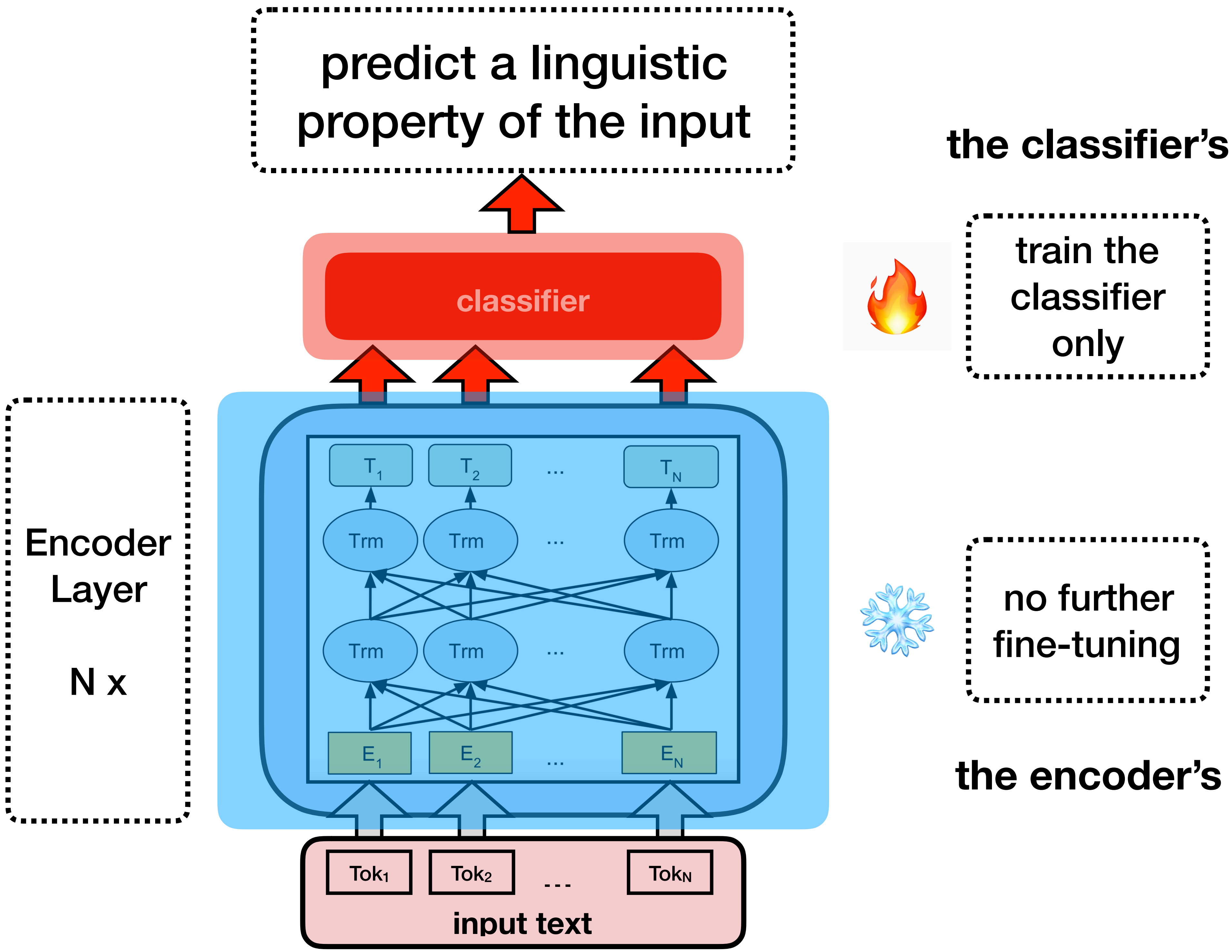
- if we cannot train a classifier to predict a property of the input text based on its representation, it means the property is not encoded in the representation or not encoded in a useful way, considering how the representation is likely to be used

characteristics of probe tasks

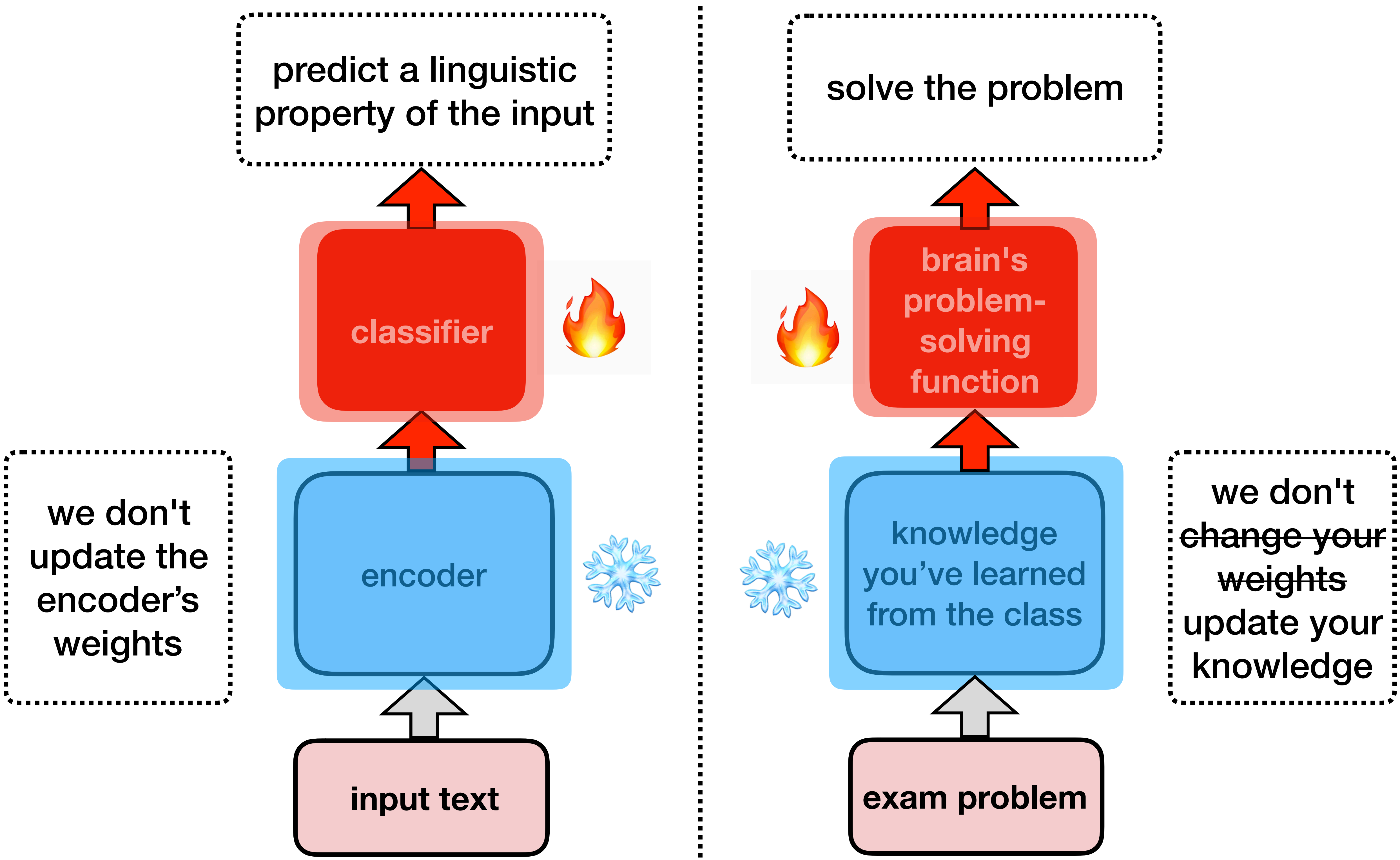
- usually classification problems that focus on simple linguistic properties
- ask simple questions, minimizing interpretability problems
- because of their simplicity, it is easier to control for biases in probing tasks than in downstream tasks
- the probing task methodology is agnostic with respect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

(Conneau et al., 2018)

probe approach

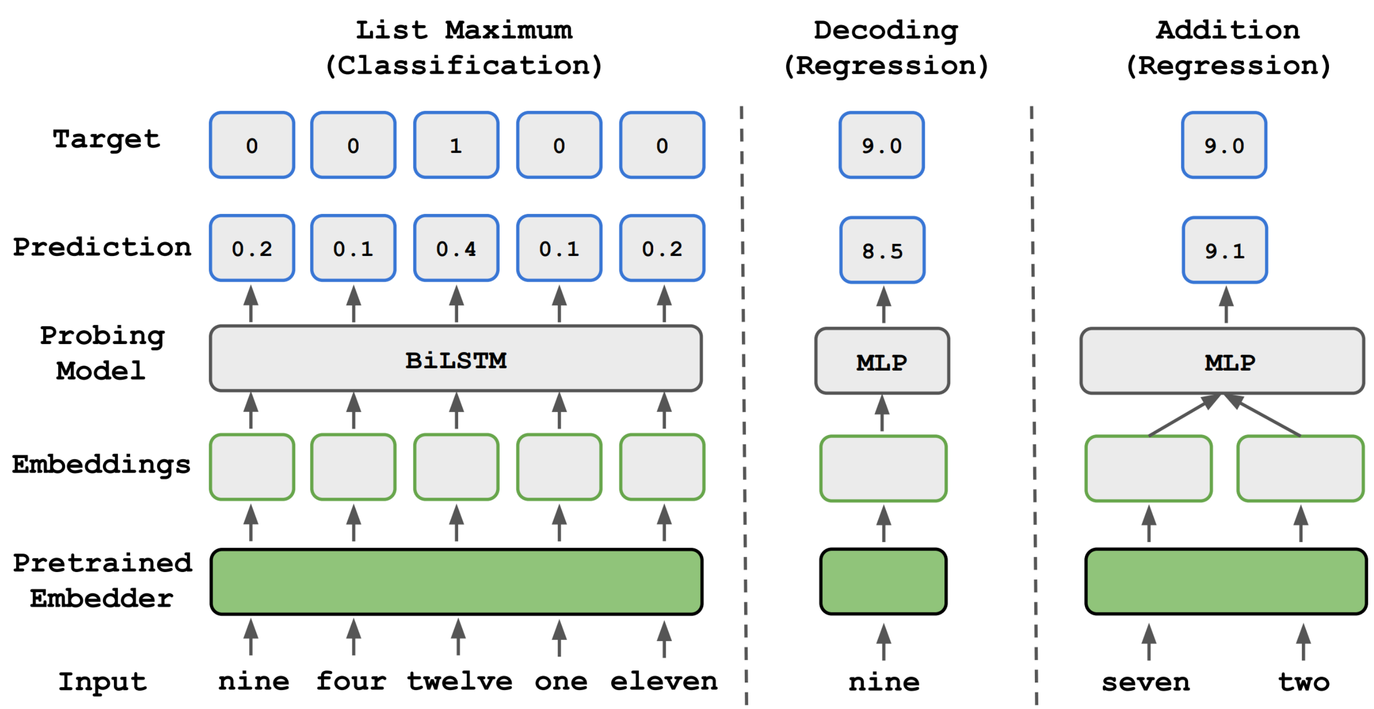


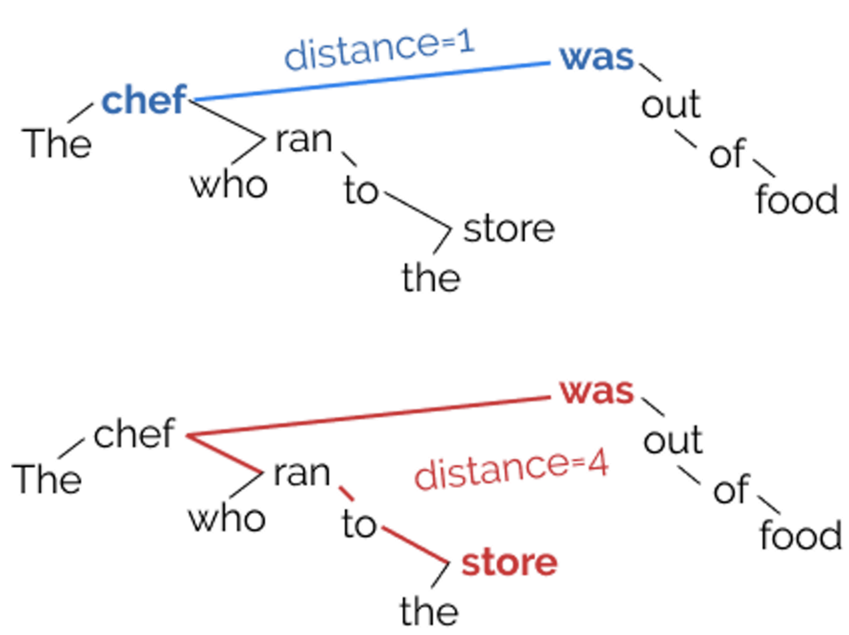
an analogy

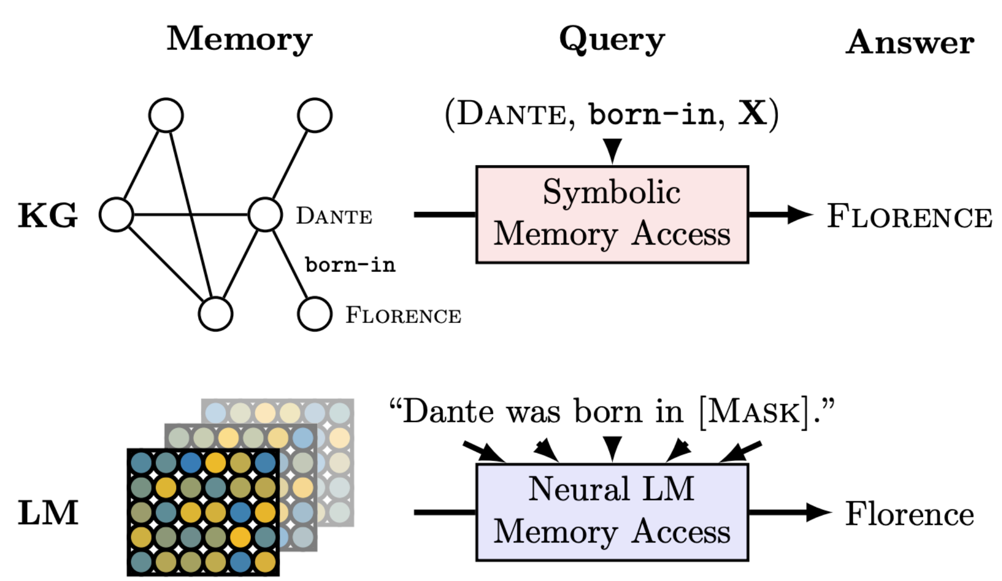


predict a linguistic property of the input
classifier
encoder
input text
we don't update the encoder’s weights
solve the problem
brain's problem-solving function
knowledge you’ve learned from the class
exam problem
we don't ~~change your weights~~ update your knowledge

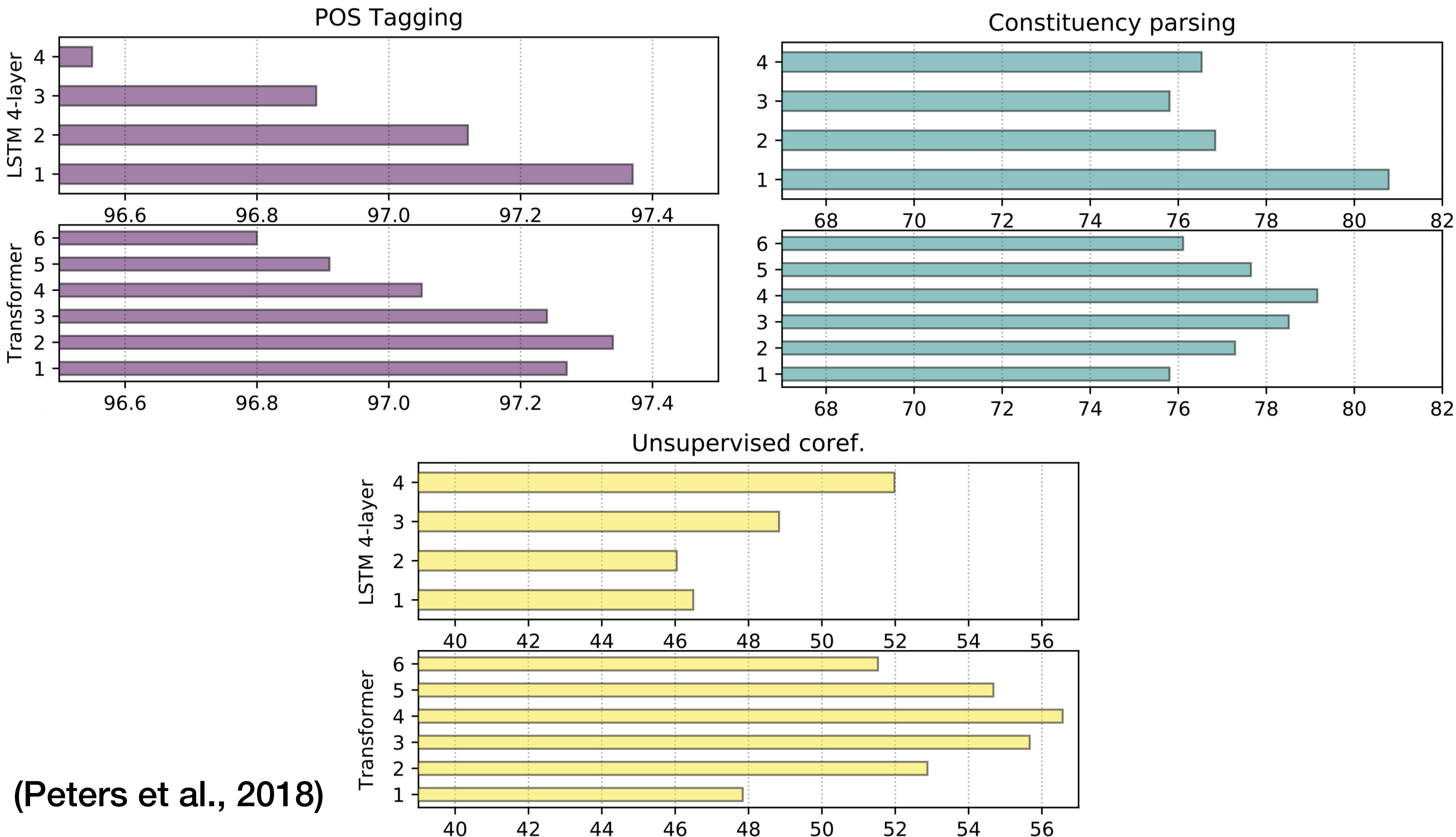
Recent results







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

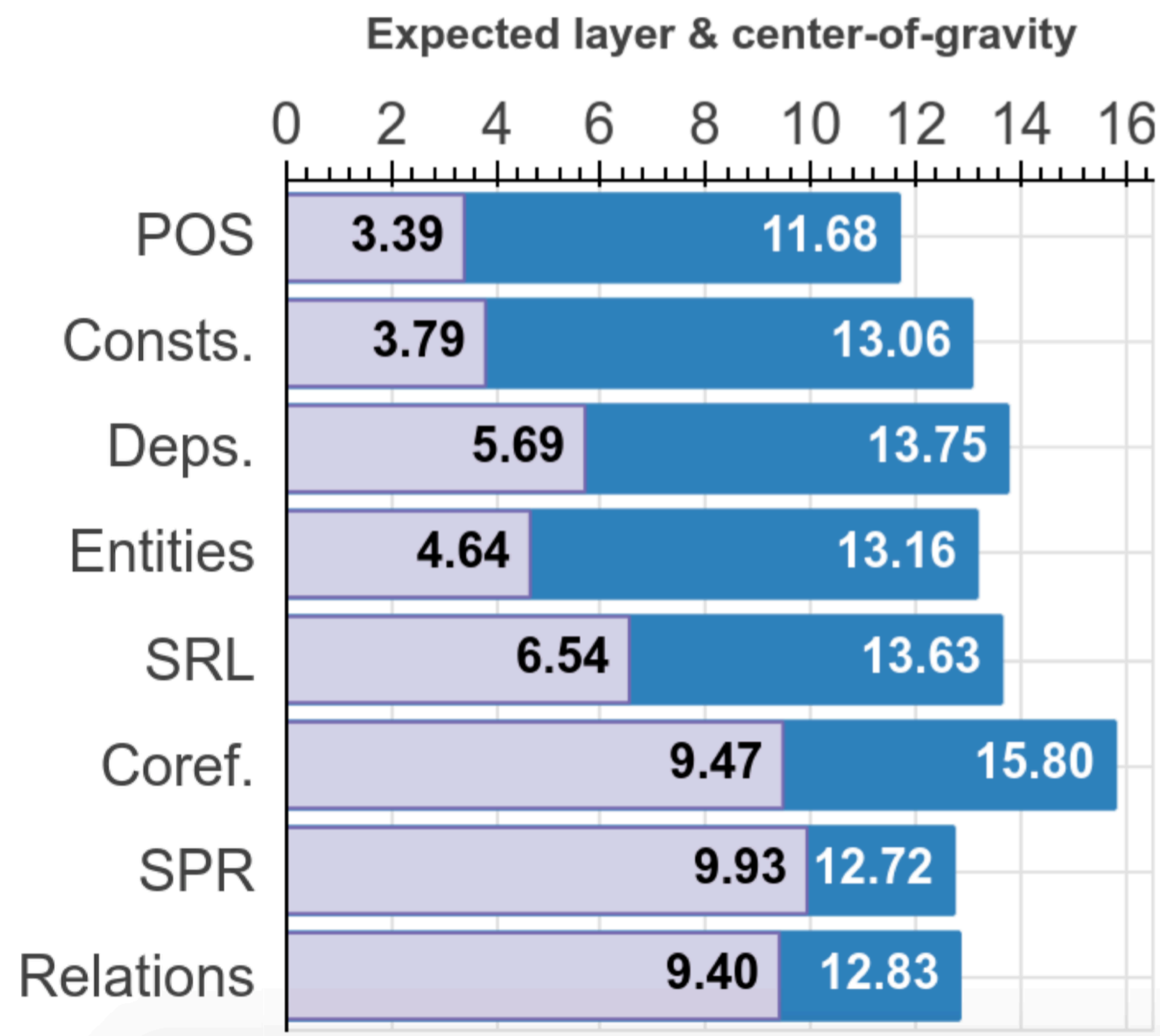


(Peters et al., 2018)

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

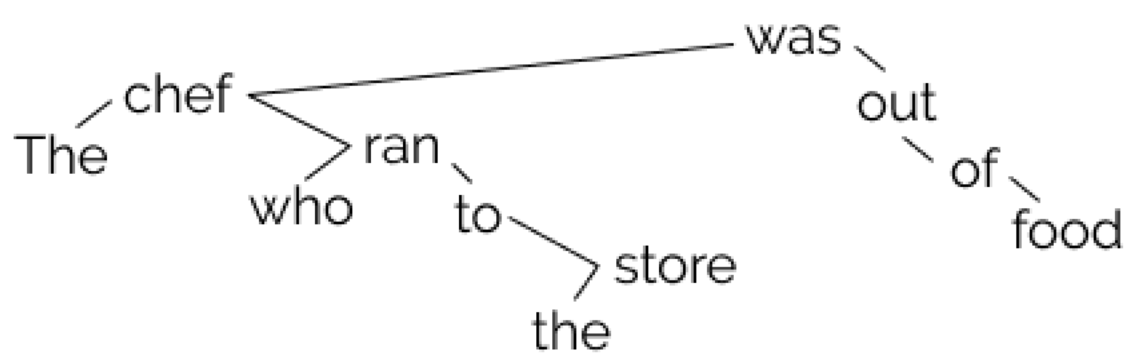
the expected layer at which the probing model correctly labels an example

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



(Tenney et al., 2019)

does BERT know the structure of syntax trees?



The chef who ran to the store was out of food

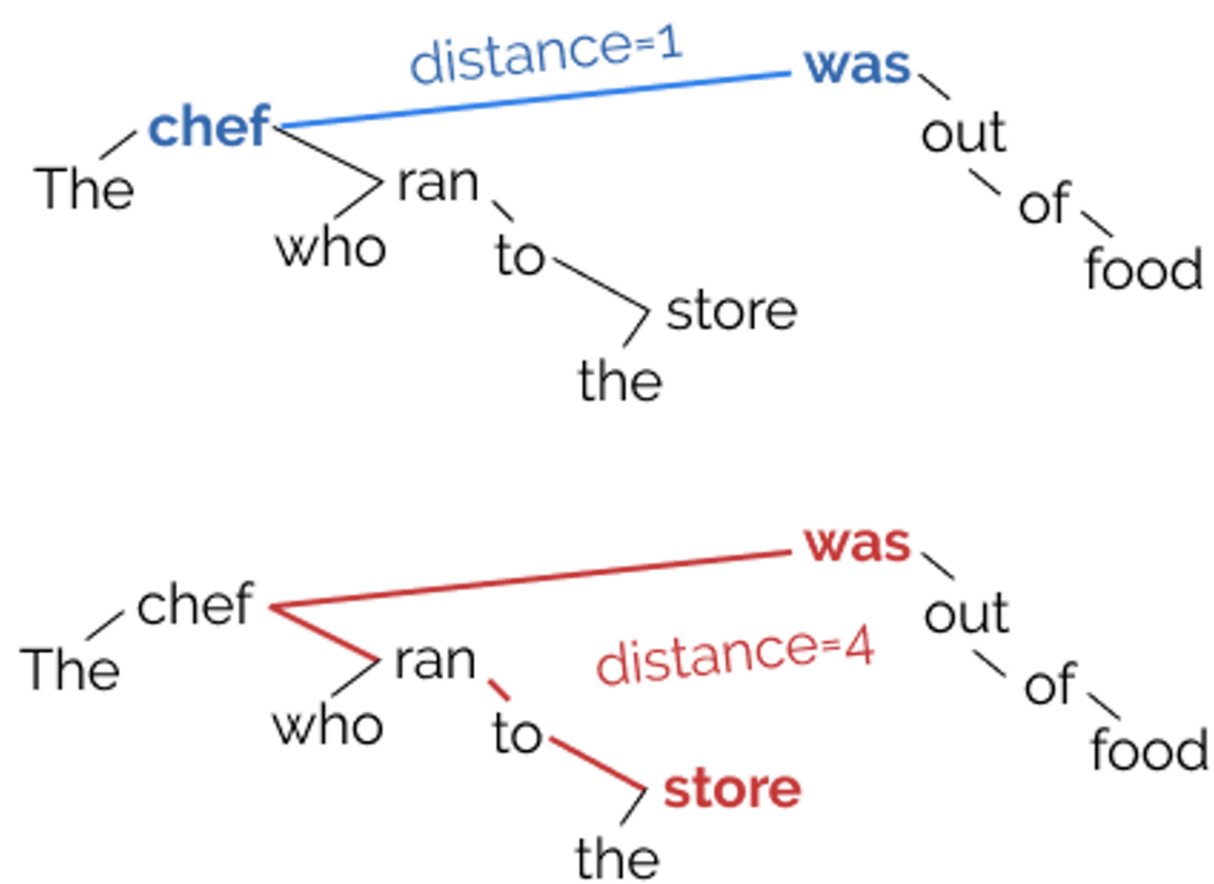
(Hewitt and Manning et al., 2019)

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.



(Hewitt and Manning et al., 2019)

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

(Hewitt and Manning et al., 2019)

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

(Hewitt and Manning et al., 2019)

learn a linear transformation

$$h \to Bh$$

squared distance

$$d(h_i, h_j)^2 = (h_i - h_j)^T (h_i - h_j)$$

$$d_B(h_i, h_j)^2 = d(Bh_i, Bh_j)^2 = (B(h_i - h_j))^T (B(h_i - h_j))$$

squared L2 norm

$$\|h\|^2 = h^T h$$

$$\|h\|_B^2 = (Bh)^T (Bh)$$

(Hewitt and Manning et al., 2019)

Yes, BERT knows the structure of syntax trees

	Distance		Depth	
Method	UUAS	DSpr.	Root%	NSpr.
ELMo1	77.0	0.83	86.5	0.87
BERTbase7	79.8	0.85	88.0	0.87
BERTlarge15	**82.5**	0.86	89.4	0.88
BERTlarge16	81.7	**0.87**	**90.1**	**0.89**

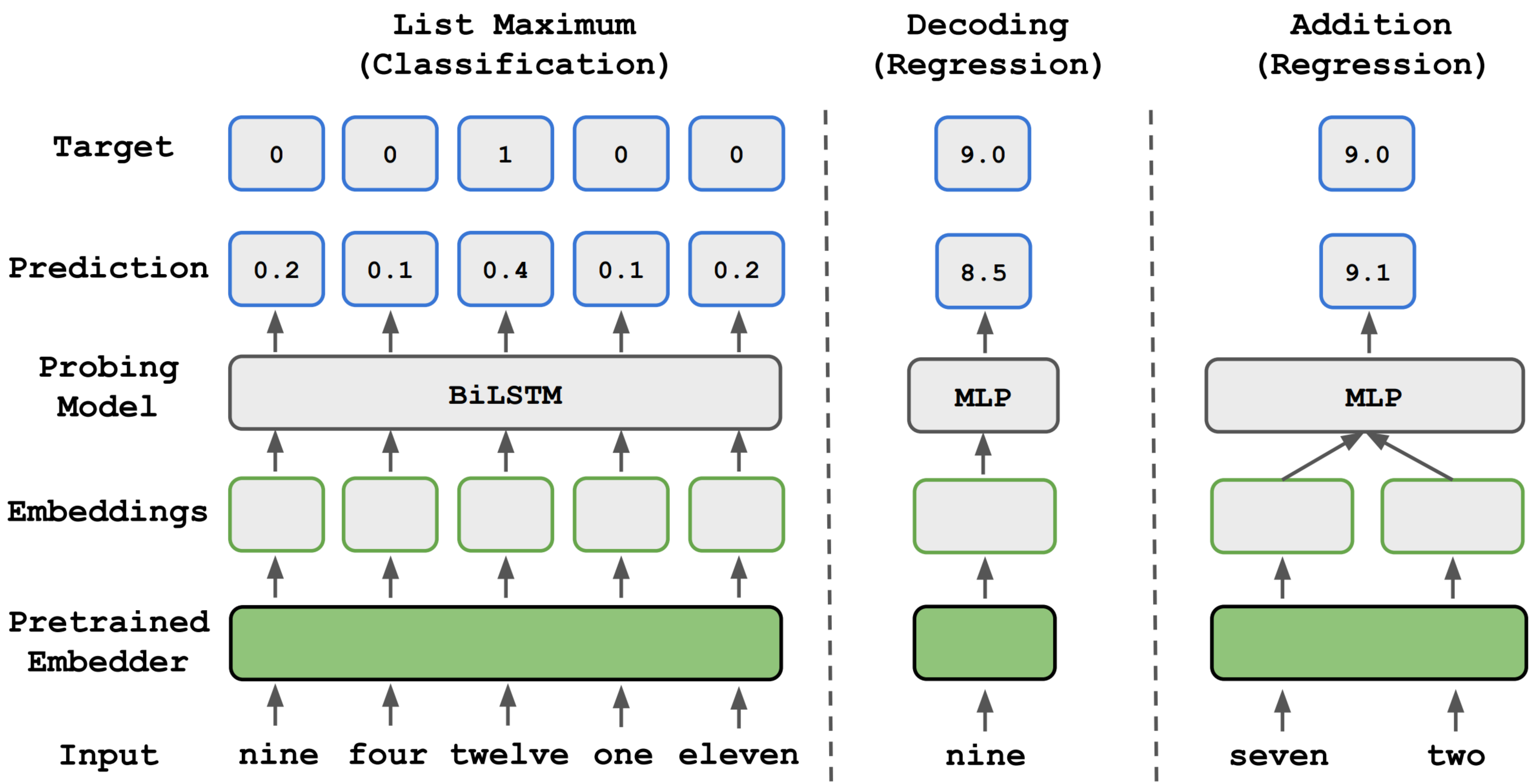
(Hewitt and Manning et al., 2019)

does BERT know numbers?

25

what is the sum of eleven and fourteen?

probing for numeracy



(Wallace et al., 2019)

Oh no! BERT struggles, But ELMo excels

Interpolation	List Maximum (5-classes)			Decoding (RMSE)			Addition (RMSE)		
Integer Range	[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
Pre-trained									
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

Interpolation	List Maximum (5-classes)	
Float Range	[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

Interpolation	List Maximum (5-classes)
Integer Range	[-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)

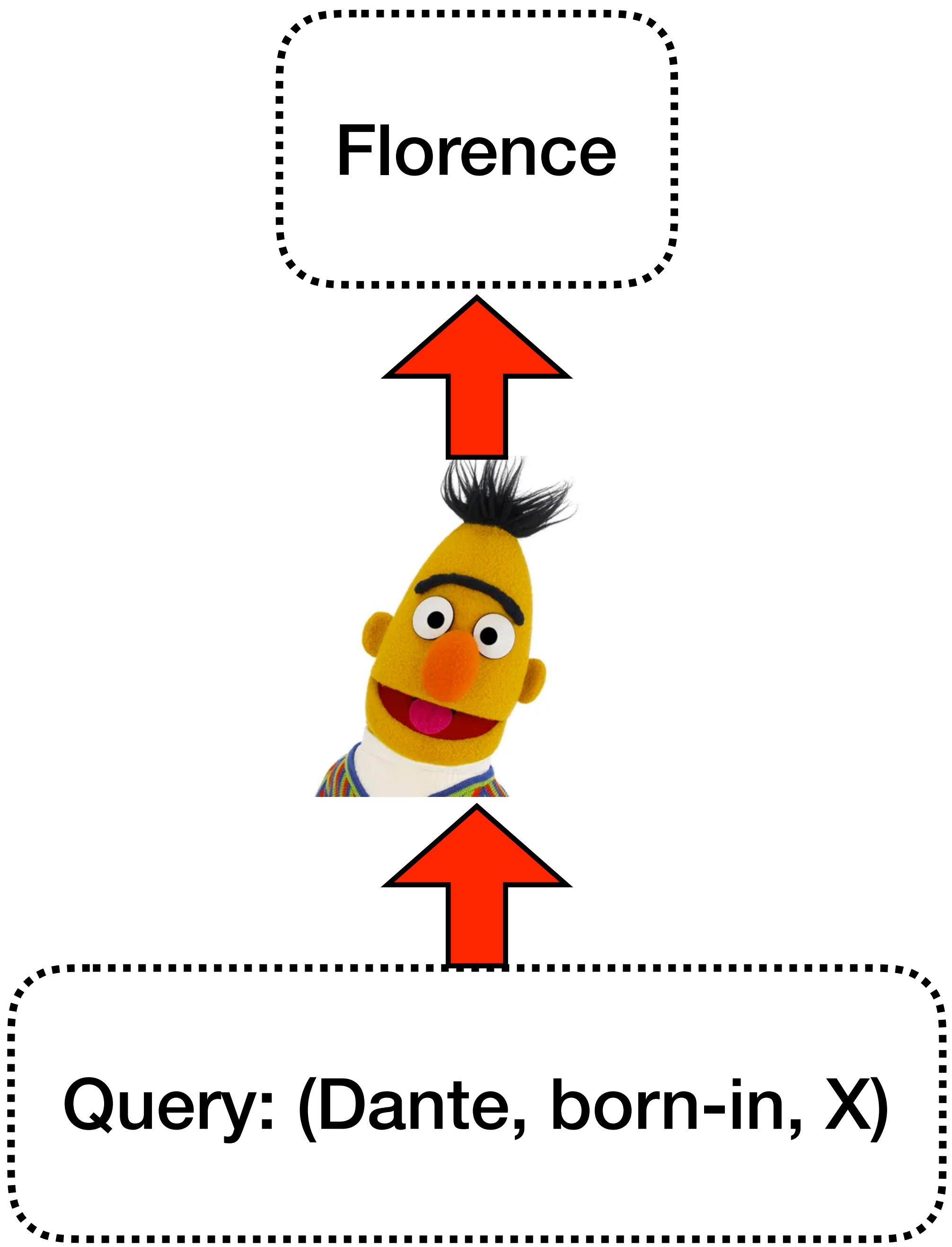
please give me a reason!

character-level CNNs are the best architecture for capturing numeracy

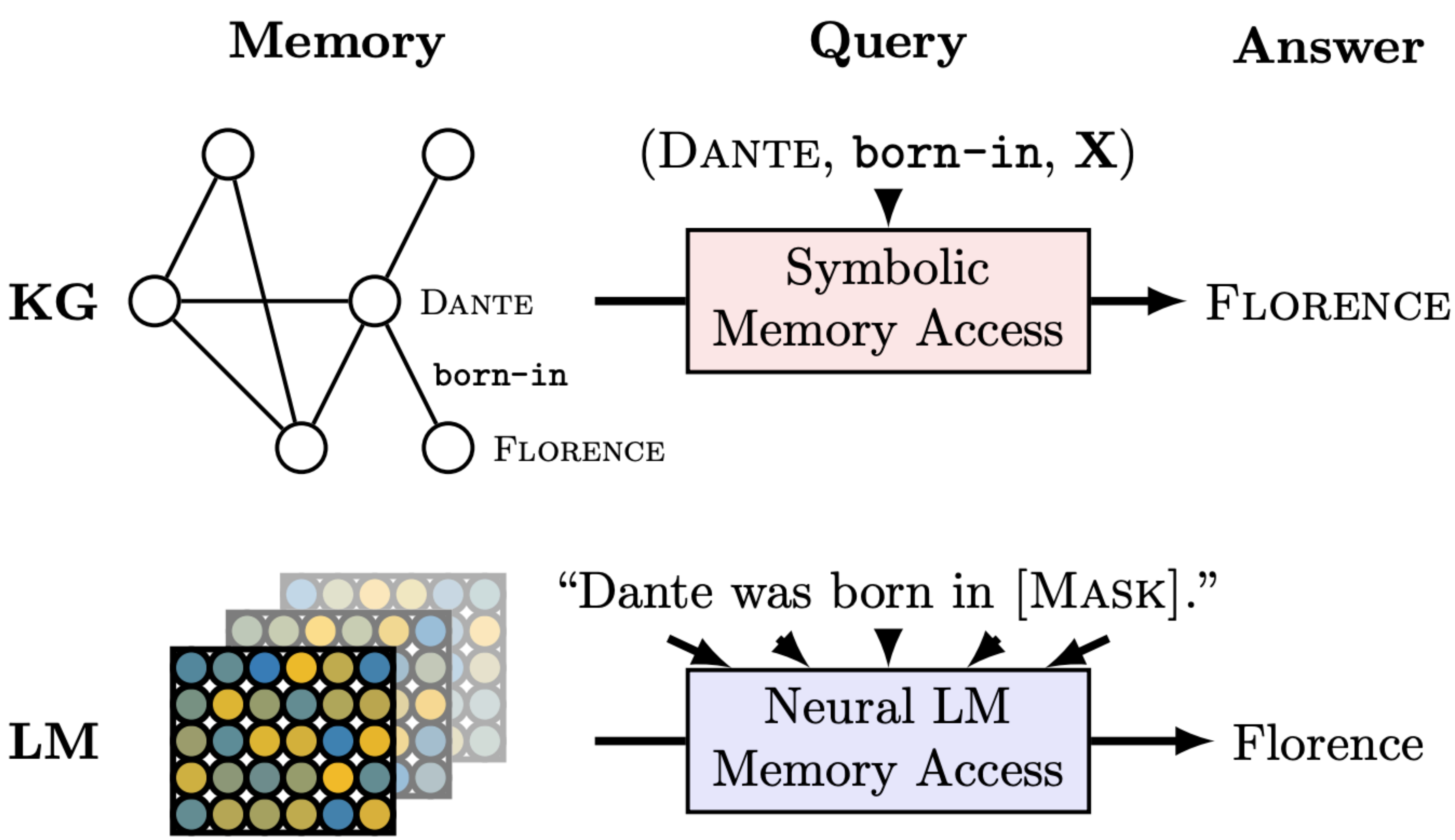
subword pieces is a poor method to encode digits, e.g., two numbers which are similar in value can have very different sub-word divisions

(Wallace et al., 2019)

Can BERT serve as a structured knowledge base?



LAMA (LAnguage Model Analysis) probe



(Petroni et al., 2019)

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]”

(Petroni et al., 2019)

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] , **sodium [-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] , **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0] , **Uganda [-3.2]** , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]** , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-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	**Canada [-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] , **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-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] , **laugh [-2.0]** , vomit [-2.6] , scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2] , cancer [-2.0] , **infection [-2.6]** , plague [-3.3] , fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8] , nests [-3.1] , **feathers [-3.2]** , died [-3.7] , eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , **speed [-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] , **alive [-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] , **fish [-2.8]** , recreation [-3.1]

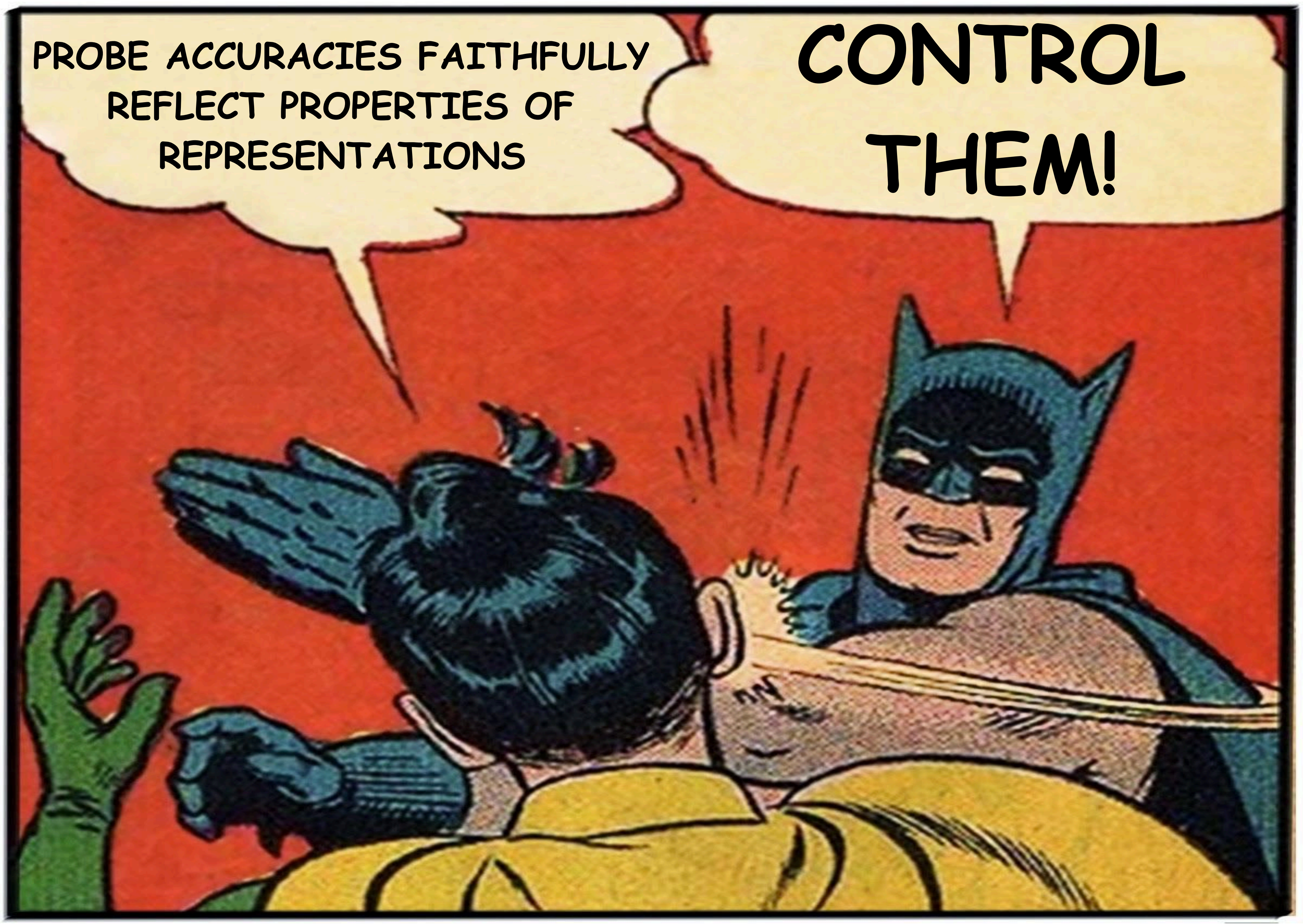
(Petroni et al., 2019)

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_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	`birth-place`	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	`birth-date`	1825	1	1.9	-	0.0	**1.9**	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	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	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	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

(Petroni et al., 2019)

are probe tasks a perfect tool?



probe complexity

arguments for “simple” probes

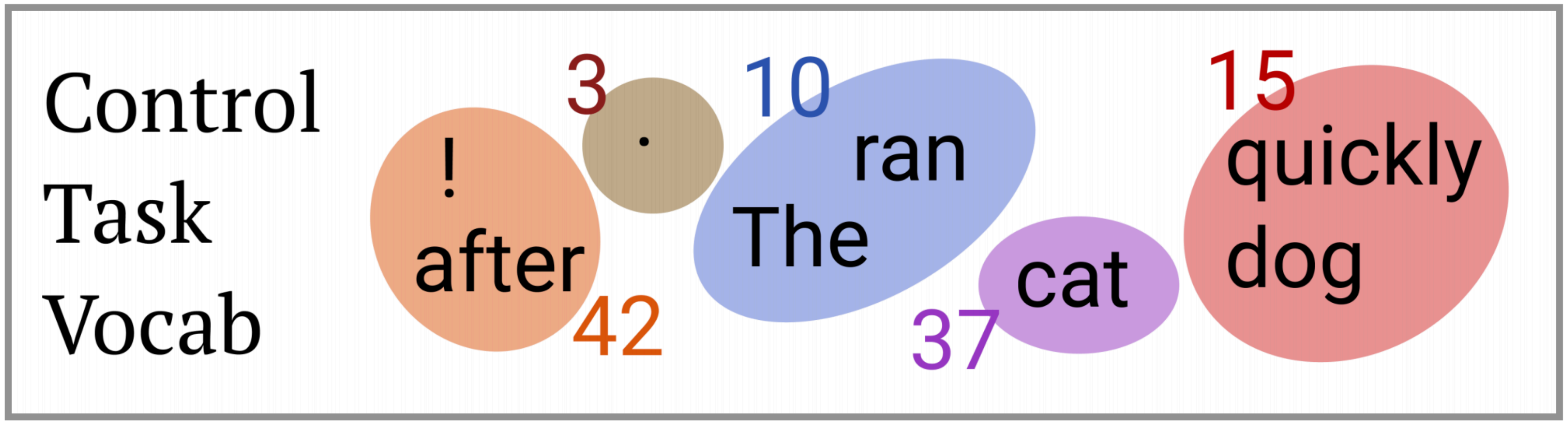
we want to find easily accessible information in a representation

arguments for “complex” probes

useful properties might be encoded non-linearly

(Hewitt et al., 2019)

control tasks



Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	**10**	**37**	**10**	**15**	**3**
Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	**10**	**15**	**10**	**42**	**42**

(Hewitt et al., 2019)

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*

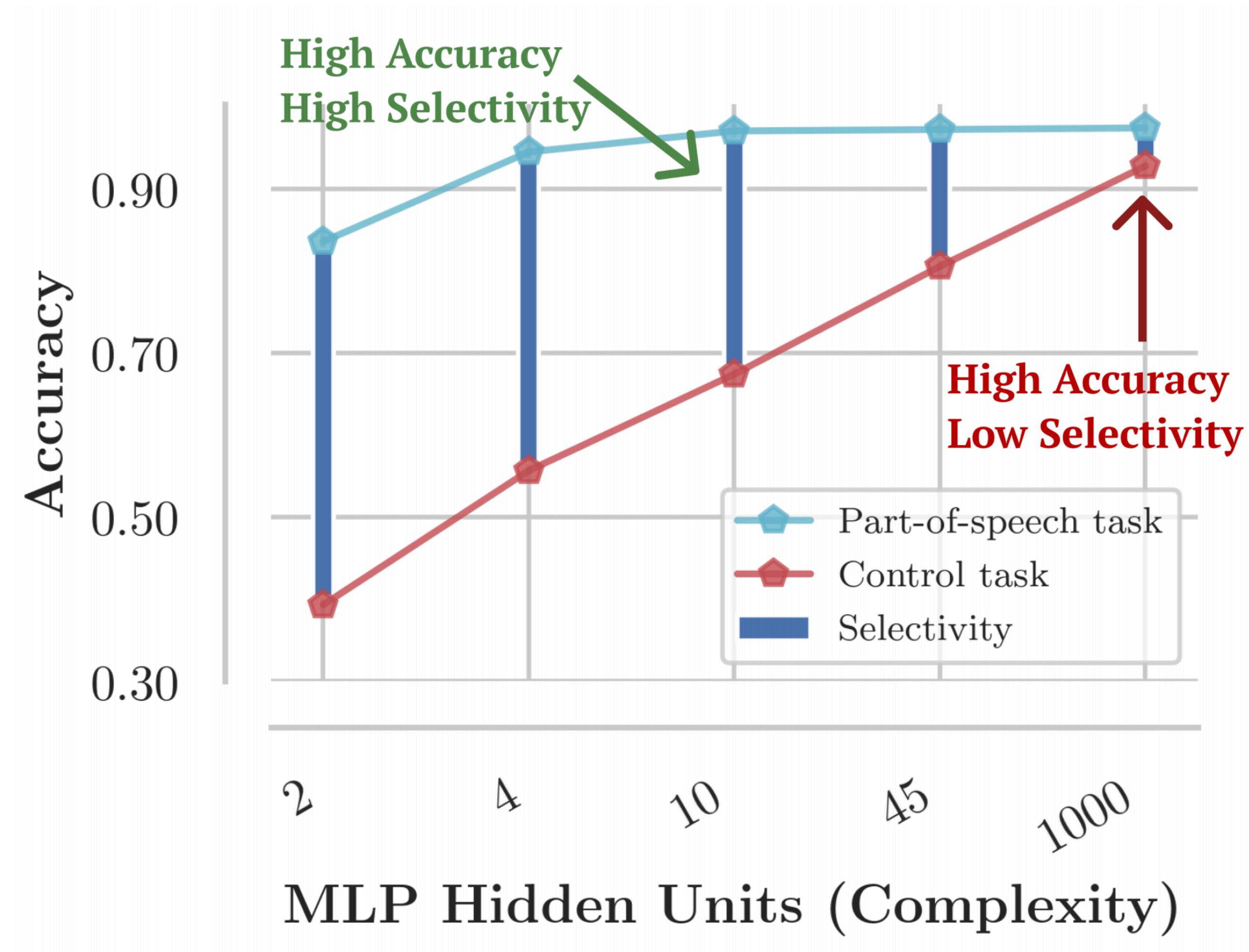
- *control task is a function that maps each token x_i to the label specified by the behavior $C(x_i)$*

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

(Hewitt et al., 2019)

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



(Hewitt et al., 2019)

be careful about probe accuracies

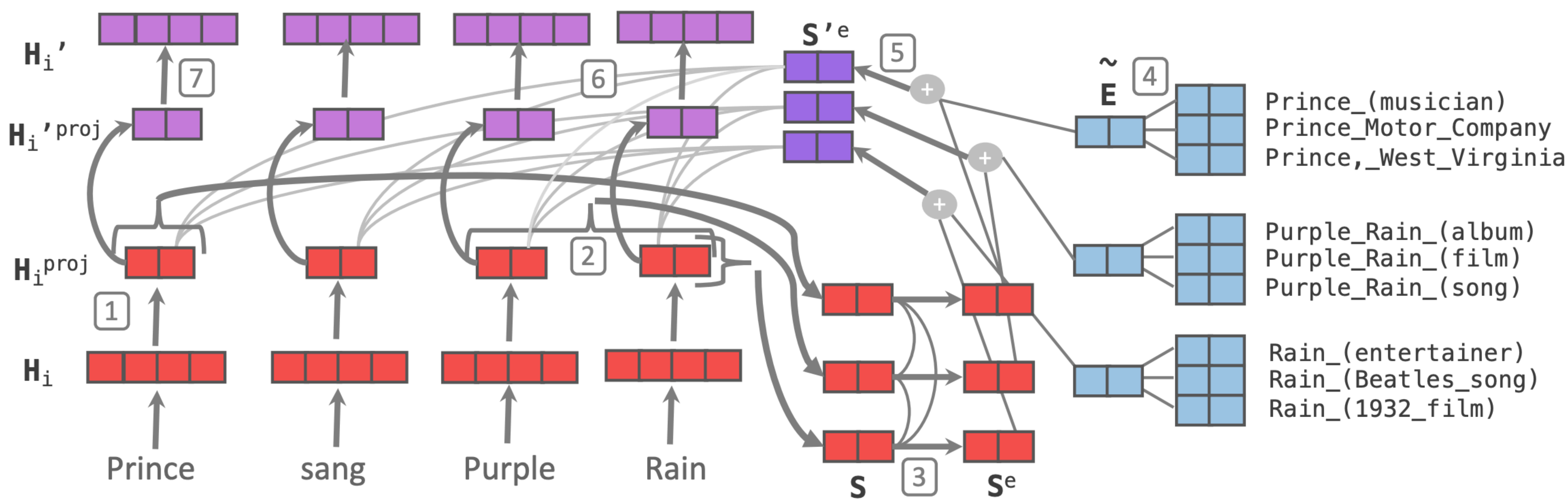
Part-of-speech Tagging				
	Linear		MLP-1	
Model	Accuracy	Selectivity	Accuracy	Selectivity
Proj0	96.3	20.6	97.1	1.6
ELMo1	97.2	26.0	97.3	4.5
ELMo2	96.6	31.4	97.0	8.8

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



(Peters et al., 2019)

Thank you!

References

- Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Raffel et al., 2019. https://arxiv.org/abs/1910.10683
- BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Lewis et al., 2019. https://arxiv.org/abs/1910.13461
- Unsupervised Cross-lingual Representation Learning at Scale. Conneau et al., 2019. https://arxiv.org/abs/1911.02116
- Are Sixteen Heads Really Better than One? Michel et al., NeurIPS 2019. https://arxiv.org/abs/1905.10650
- What Does BERT Look At? An Analysis of BERT's Attention, Clark el al., BlackBoxNLP 2019. https://arxiv.org/abs/1906.04341
- Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. Adi et al., ICLR 2017. https://arxiv.org/abs/1608.04207
- Linguistic Knowledge and Transferability of Contextual Representations. Liu et al., NAACL 2019. https://arxiv.org/abs/1903.08855
- What do you learn from context? Probing for sentence structure in contextualized word representations. Tenney et al., ICLR 2019. https://arxiv.org/abs/1905.06316

References

- What you can cram into a single vector: Probing sentence embeddings for linguistic properties. Conneau et al., ACL 2018. https://arxiv.org/abs/1805.01070
- Dissecting Contextual Word Embeddings: Architecture and Representation. Peters et al., EMNLP 2018. https://arxiv.org/abs/1808.08949
- BERT Rediscovers the Classical NLP Pipeline. Tenney et al., ACL 2019. https://arxiv.org/abs/1905.05950
- A Structural Probe for Finding Syntax in Word Representations, Hewitt and Manning, NAACL 2019. https://nlp.stanford.edu/pubs/hewitt2019structural.pdf
- Do NLP Models Know Numbers? Probing Numeracy in Embeddings, Wallace et al., EMNLP 2019. https://arxiv.org/abs/1909.07940
- Language Models as Knowledge Bases?, Petroni et al., EMNLP 2019. https://arxiv.org/abs/1909.01066
- Designing and Interpreting Probes with Control Tasks, Hewitt and Liang, EMNLP 2019. https://arxiv.org/abs/1909.03368
- Knowledge Enhanced Contextual Word Representations, Peters et al., EMNLP 2019. https://arxiv.org/abs/1909.04164