Probing / interpretability

CS 685, Spring 2024

Introduction to Natural Language Processing

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

Mohit Iyyer

College of Information and Computer Sciences

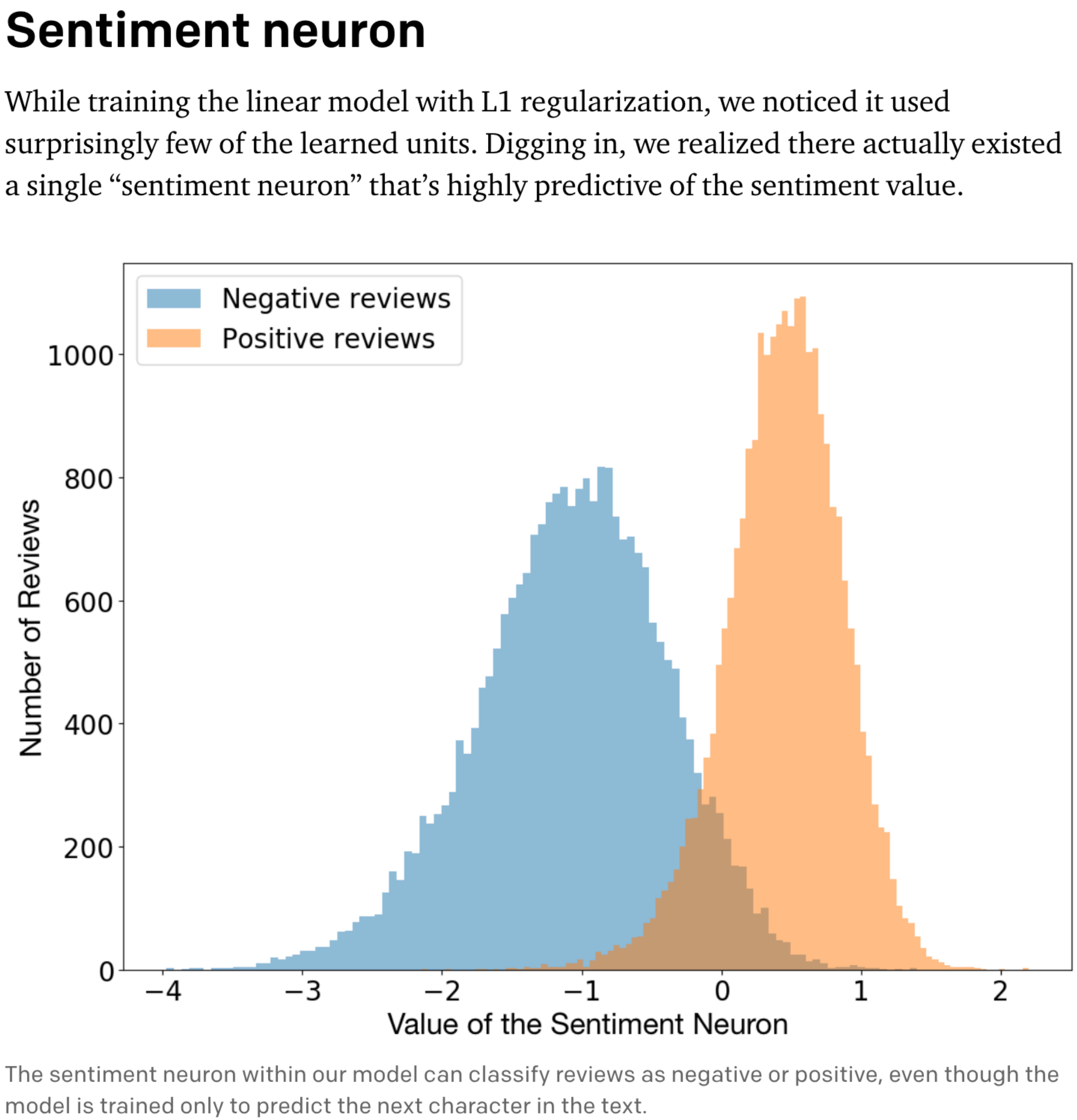
University of Massachusetts Amherst

most slides from Tu Vu

understanding representations

two prominent methods

- visualization
- linguistic probe tasks

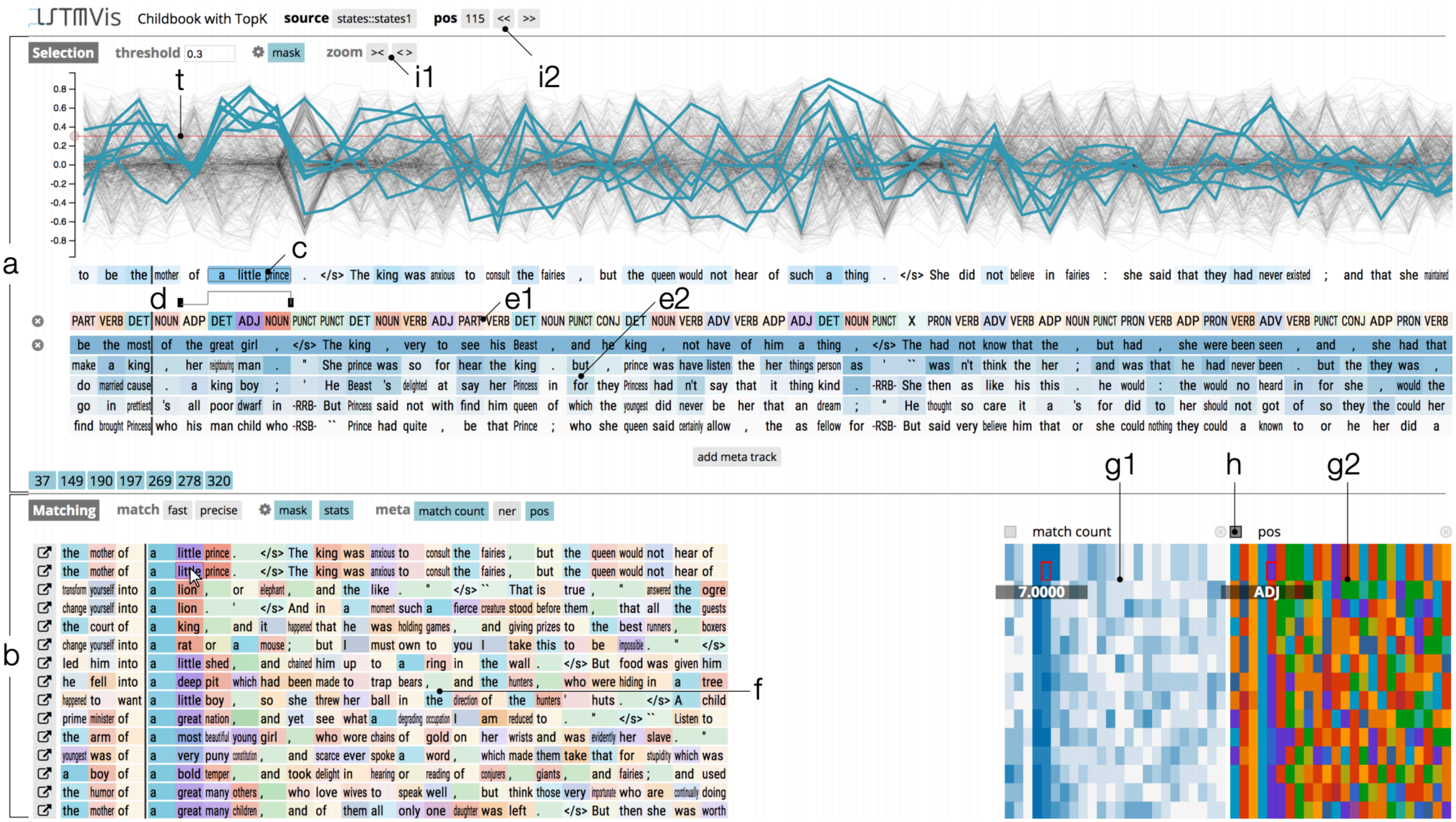


Sentiment neuron

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.

https://openai.com/blog/unsupervised-sentiment-neuron/



to be the mother of a little prince . </s> The king was anxious to consult the fairies , but the queen would not hear of such a thing . </s> She did not believe in fairies : she said that they had never existed ; and that she maintained

PART VERB DET NOUN ADP DET ADJ NOUN PUNCT PUNCT DET NOUN VERB ADJ PART VERB DET NOUN PUNCT CONJ DET NOUN VERB ADV VERB ADP ADJ DET NOUN PUNCT X PRON VERB ADV VERB ADP NOUN PUNCT PRON VERB ADP PRON VERB ADV VERB PUNCT CONJ ADP PRON VERB

be the most of the great girl , </s> The king , very to see his Beast , and he king , not have of him a thing , </s> The had not know that the , but had , she were been seen , and , she had that

make a king , her neighbouring man . " She prince was so for hear the king . but , prince was have listen the her things person as ' `` was n't think the her ; and was that he had never been . but the they was ,

do married cause . a king boy ; ' He Beast 's delighted at say her Princess in for they Princess had n't say that it thing kind . -RRB- She then as like his this . he would : the would no heard in for she , would the

go in prettiest 's all poor dwarf in -RRB- But Princess said not with find him queen of which the youngest did never be her that an dream ; " He thought so care it a 's for did to her should not got of so they the could her

find brought Princess who his man child who -RSB- `` Prince had quite , be that Prince ; who she queen said certainly allow , the as fellow for -RSB- But said very believe him that or she could nothing they could a known to or he her did a

37 149 190 197 269 278 320

the mother of a little prince . </s> The king was anxious to consult the fairies , but the queen would not hear of

the mother of a little prince . </s> The king was anxious to consult the fairies , but the queen would not hear of

transform yourself into a lion , or elephant , and the like . " </s> `` That is true , " answered the ogre

change yourself into a lion . ' </s> And in a moment such a fierce creature stood before them , that all the guests

the court of a king , and it happened that he was holding games , and giving prizes to the best runners , boxers

change yourself into a rat or a mouse ; but I must own to you I take this to be impossible . " </s>

led him into a little shed , and chained him up to a ring in the wall . </s> But food was given him

he fell into a deep pit which had been made to trap bears , and the hunters , who were hiding in a tree

happened to want a little boy , so she threw her ball in the direction of the hunters ' huts . </s> A child

prime minister of a great nation , and yet see what a degrading occupation I am reduced to . " </s> `` Listen to

the arm of a most beautiful young girl , who wore chains of gold on her wrists and was evidently her slave . "

youngest was of a very puny constitution , and scarce ever spoke a word , which made them take that for stupidity which was

a boy of a bold temper , and took delight in hearing or reading of conjurers , giants , and fairies ; and used

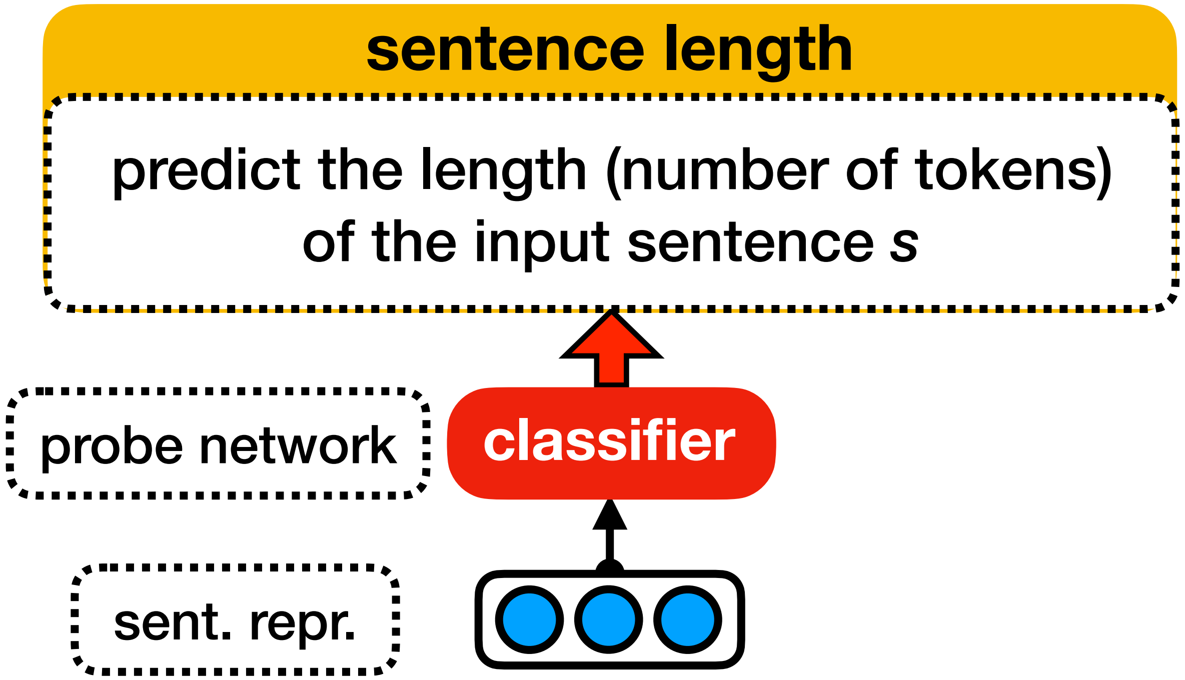
the humor of a great many others , who love wives to speak well , but think those very importunate who are continually doing

the mother of a great many children , and of them all only one daughter was left . </s> But then she was worth

LSTMVis: Strobelt et al., 2017

what is a linguistic probe task?

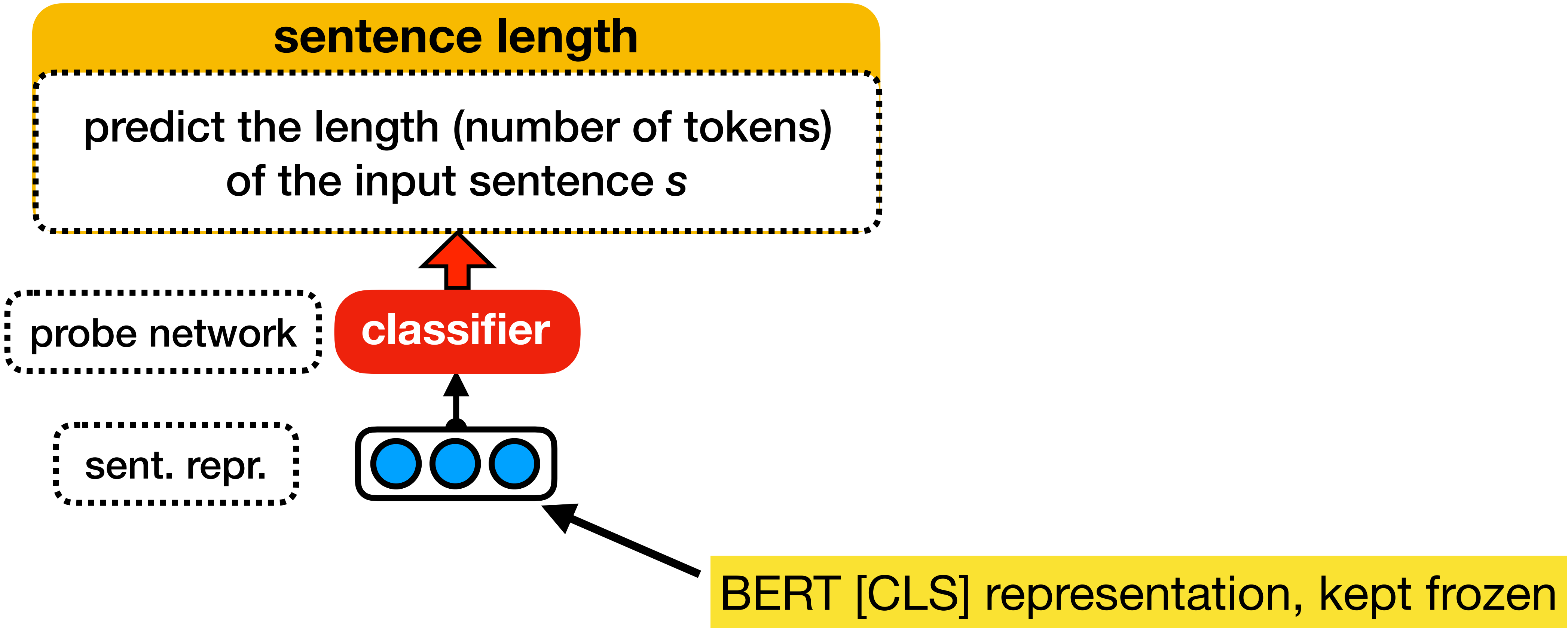
given an encoder model (e.g., BERT) pre-trained on a certain task, we use the representations it produces to train a classifier (without further fine-tuning the model) to predict a linguistic property of the input text



sentence length

predict the length (number of tokens) of the input sentence *s*

(Adi et al., 2017)



sentence length

predict the length (number of tokens) of the input sentence *s*

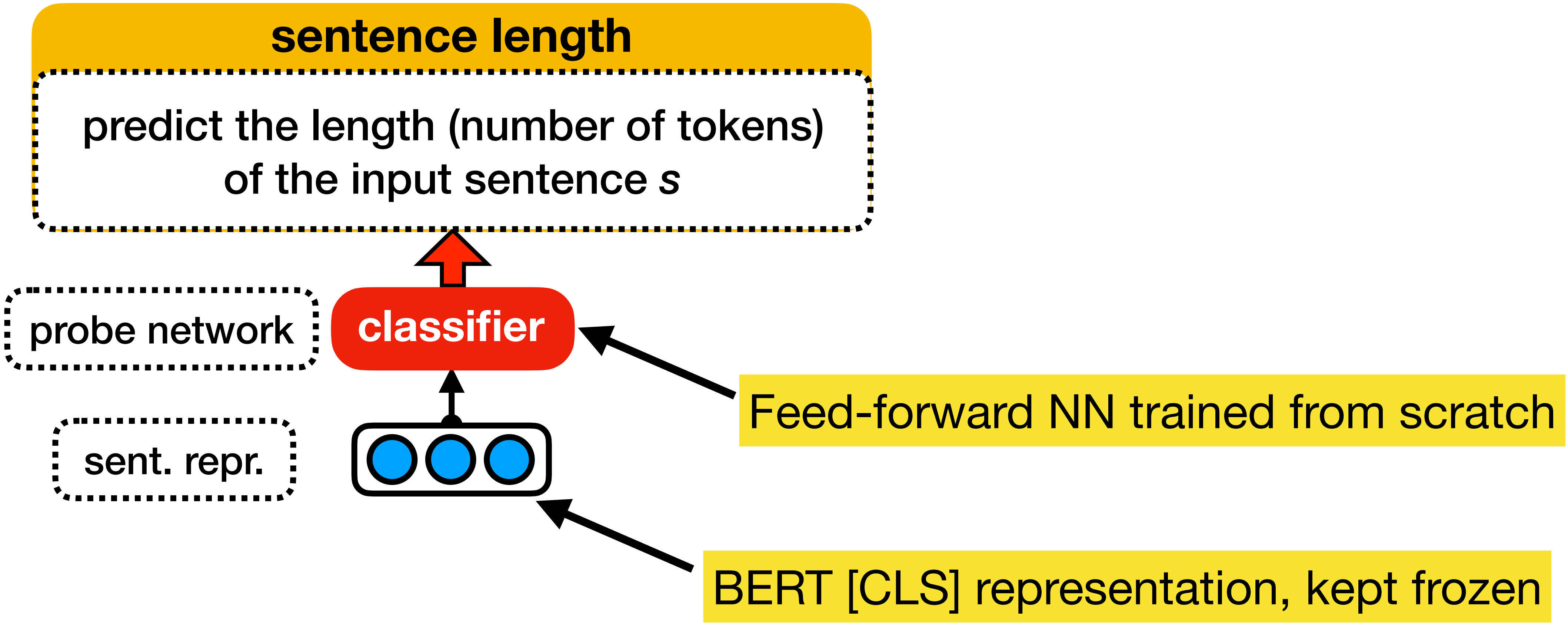
probe network

classifier

sent. repr.

BERT [CLS] representation, kept frozen

(Adi et al., 2017)



sentence length

predict the length (number of tokens) of the input sentence *s*

probe network

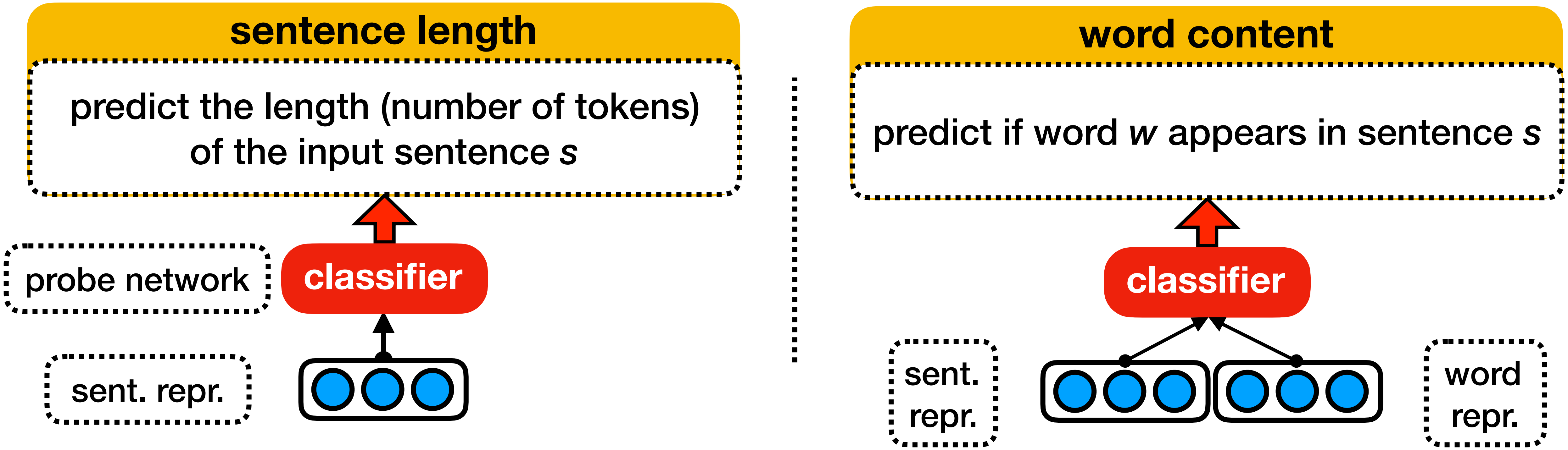
classifier

Feed-forward NN trained from scratch

sent. repr.

BERT [CLS] representation, kept frozen

(Adi et al., 2017)



sentence length

predict the length (number of tokens) of the input sentence *s*

probe network

classifier

sent. repr.

word content

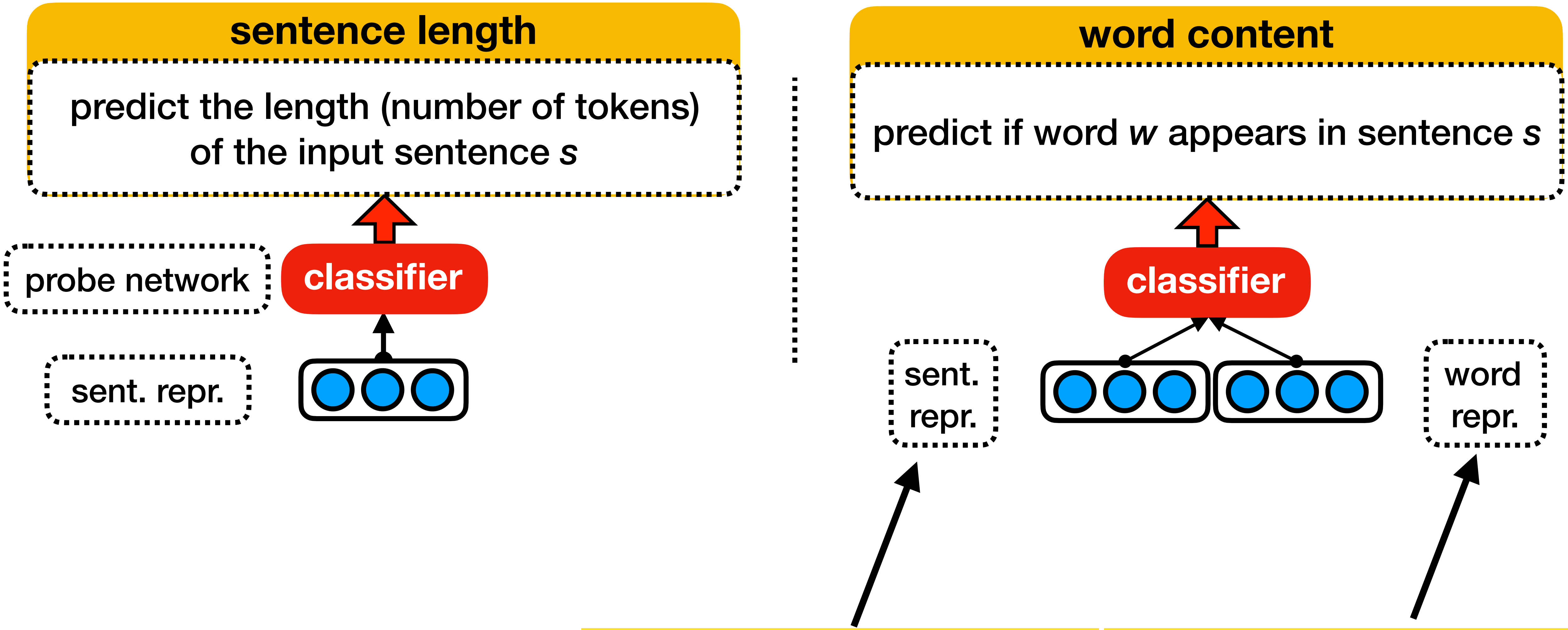
predict if word *w* appears in sentence *s*

classifier

sent. repr.

word repr.

(Adi et al., 2017)



sentence length

predict the length (number of tokens) of the input sentence *s*

probe network

classifier

sent. repr.

word content

predict if word *w* appears in sentence *s*

classifier

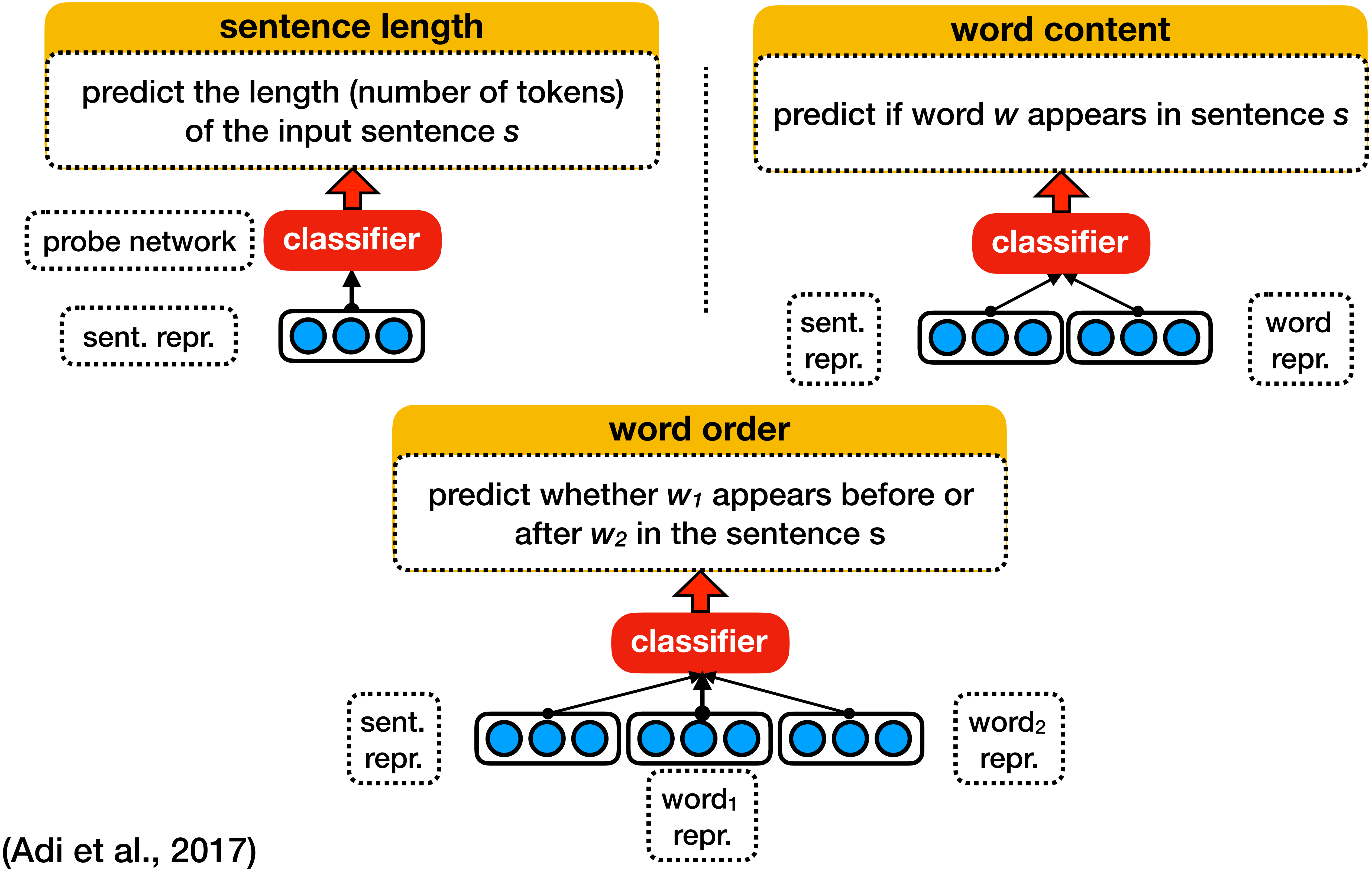
sent. repr.

word repr.

BERT [CLS] representation, kept frozen

Possibly BERT subword embedding

(Adi et al., 2017)



sentence length

predict the length (number of tokens) of the input sentence *s*

probe network

classifier

sent. repr.

word content

predict if word *w* appears in sentence *s*

classifier

sent. repr.

word repr.

word order

predict whether w_1 appears before or after w_2 in the sentence s

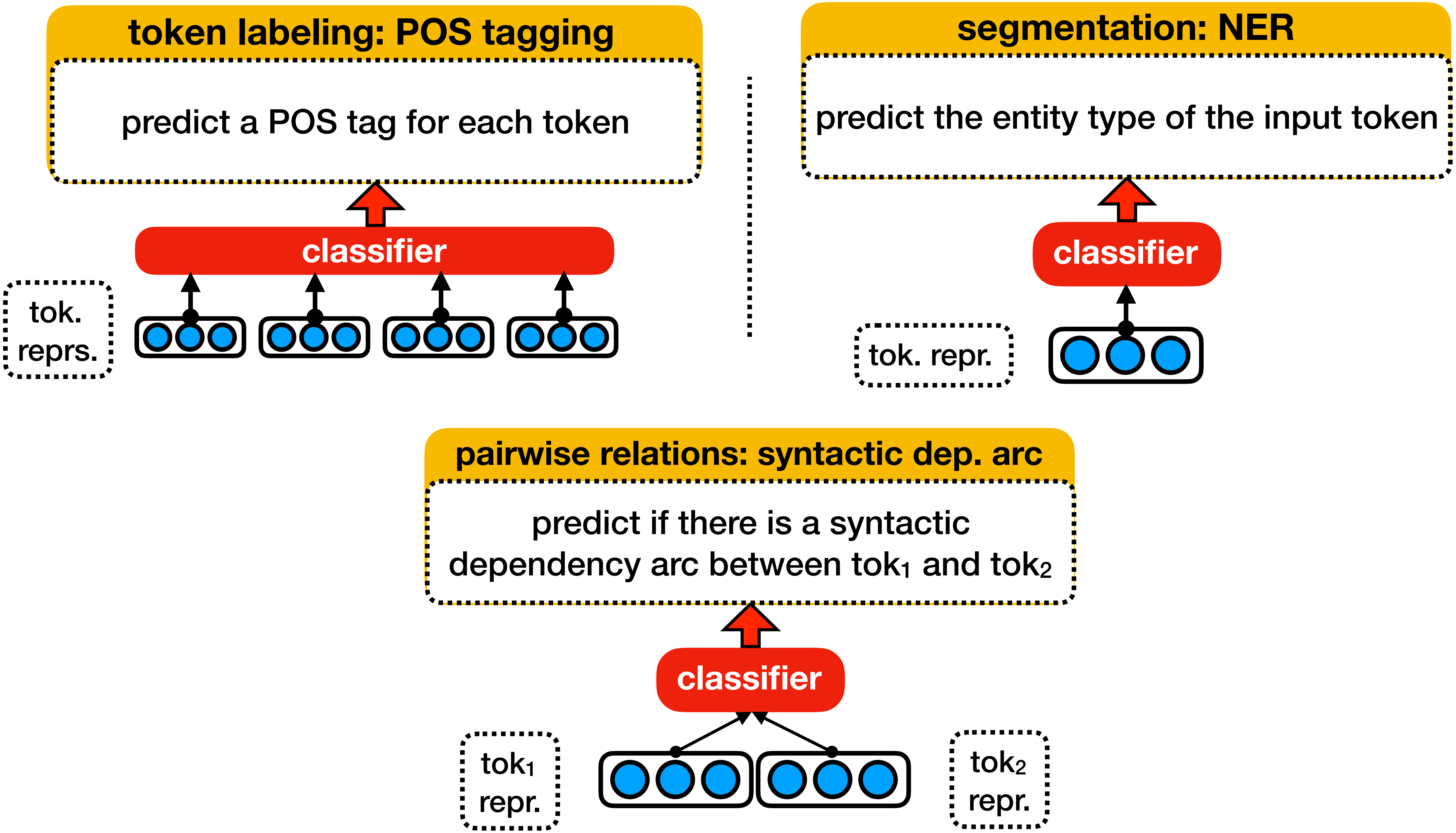
classifier

sent. repr.

word$_1$ repr.

word$_2$ repr.

(Adi et al., 2017)



token labeling: POS tagging

predict a POS tag for each token

classifier

tok. reprs.

segmentation: NER

predict the entity type of the input token

classifier

tok. repr.

pairwise relations: syntactic dep. arc

predict if there is a syntactic dependency arc between tok$_1$ and tok$_2$

classifier

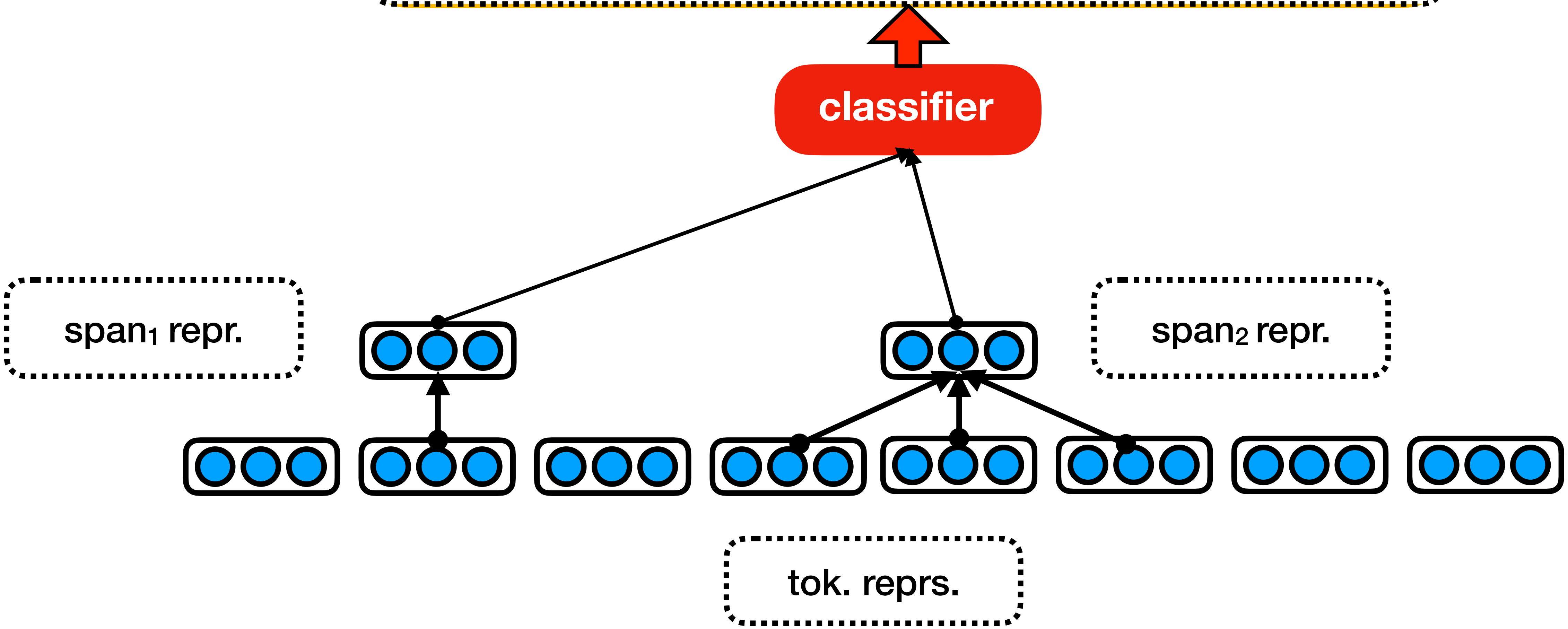
tok$_1$ repr.

tok$_2$ repr.

(Liu et al., 2019)

edge probing: coreference

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



classifier

span$_1$ repr.

span$_2$ repr.

tok. reprs.

(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 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

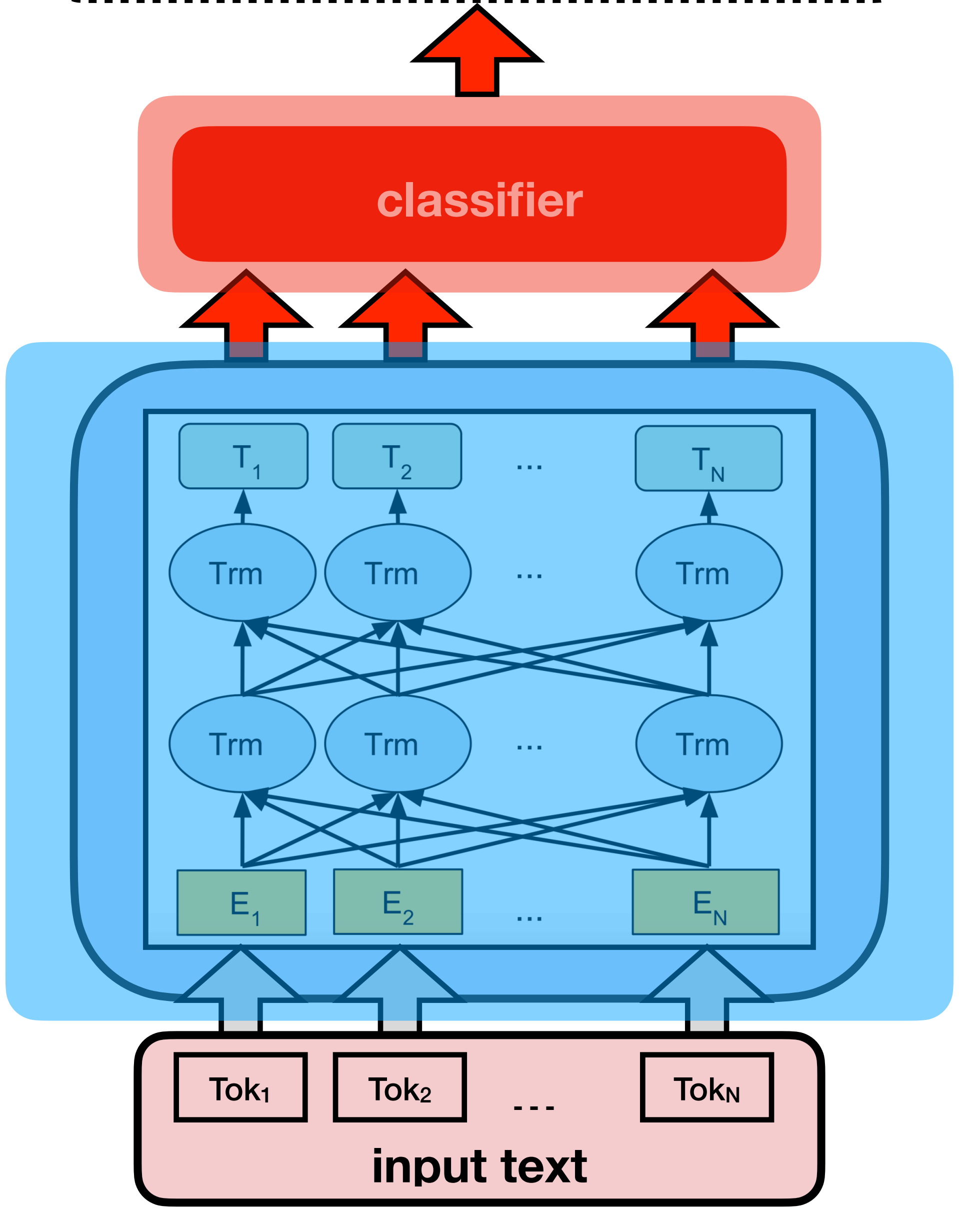
characteristics of probe tasks

- 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

(Conneau et al., 2018)

probe approach

predict a linguistic property of the input



Encoder Layer

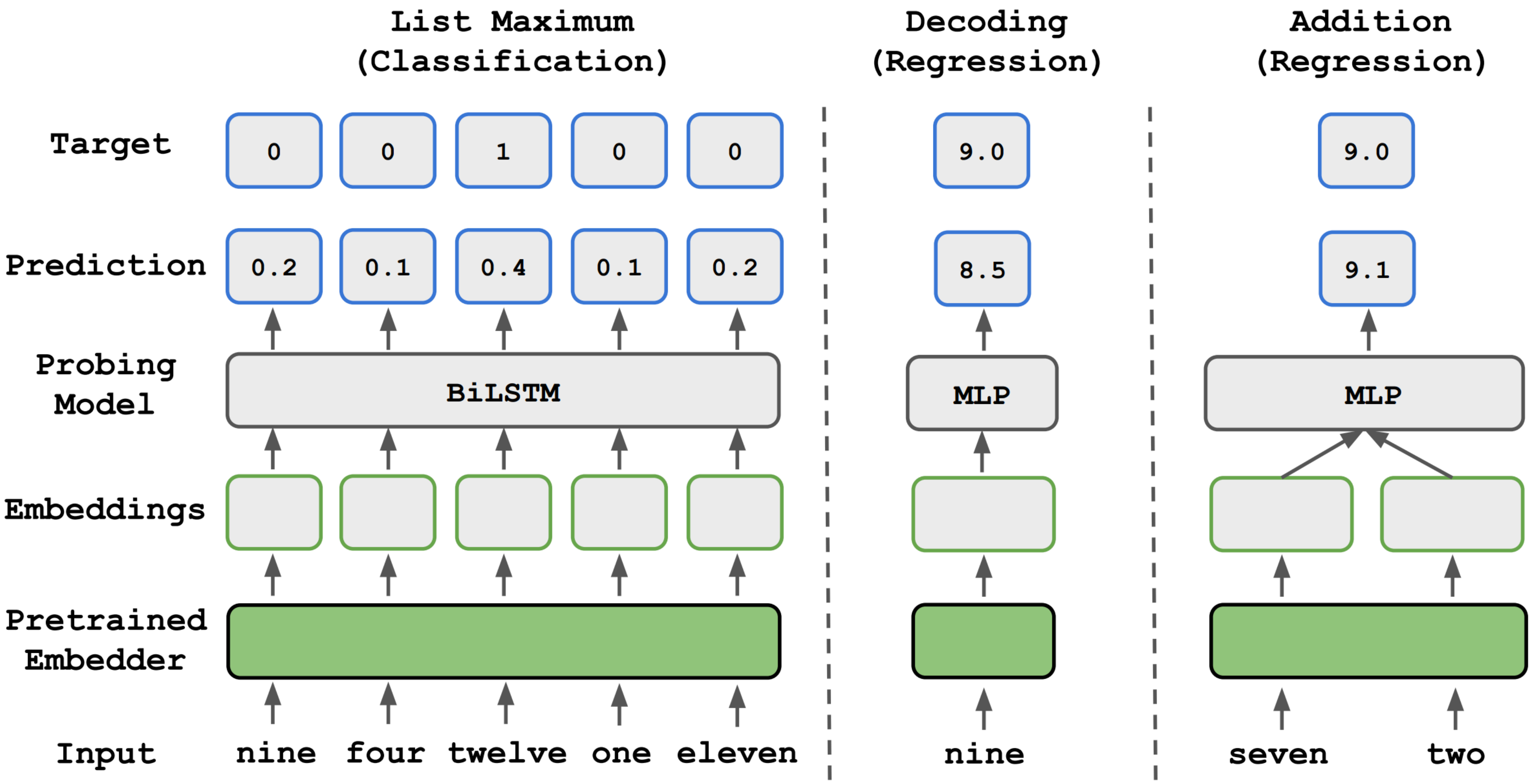
N x

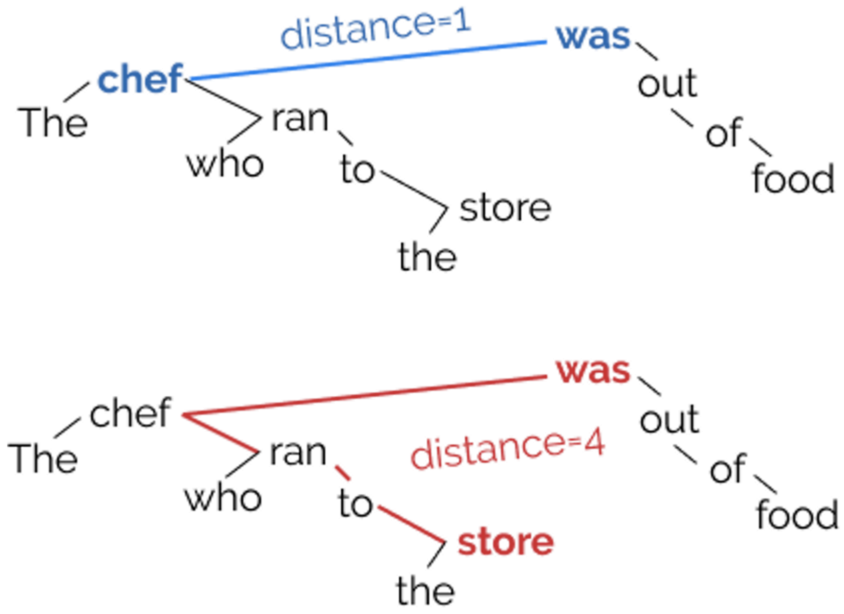
the classifier’s weights are updated

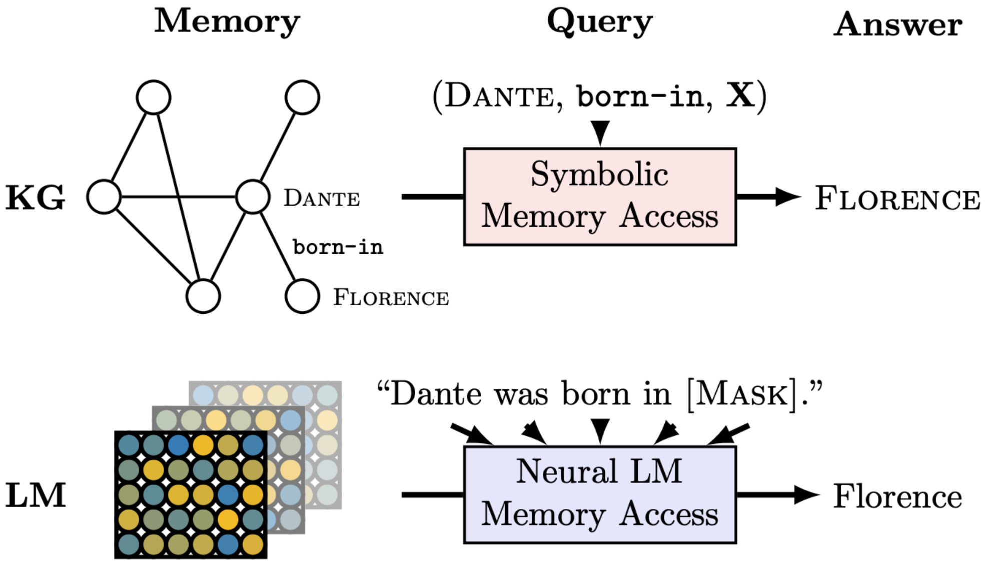
🔥 train the classifier only

❄️ no further fine-tuning

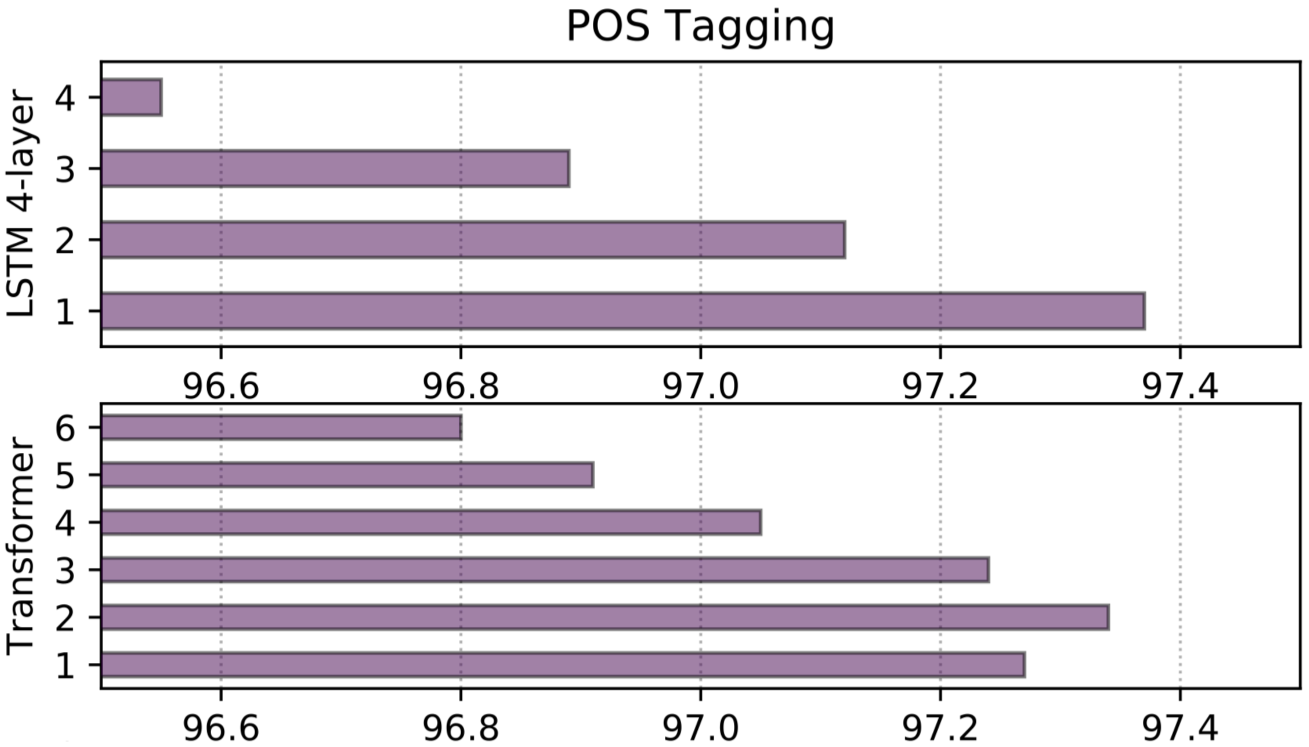
the encoder’s weights are fixed

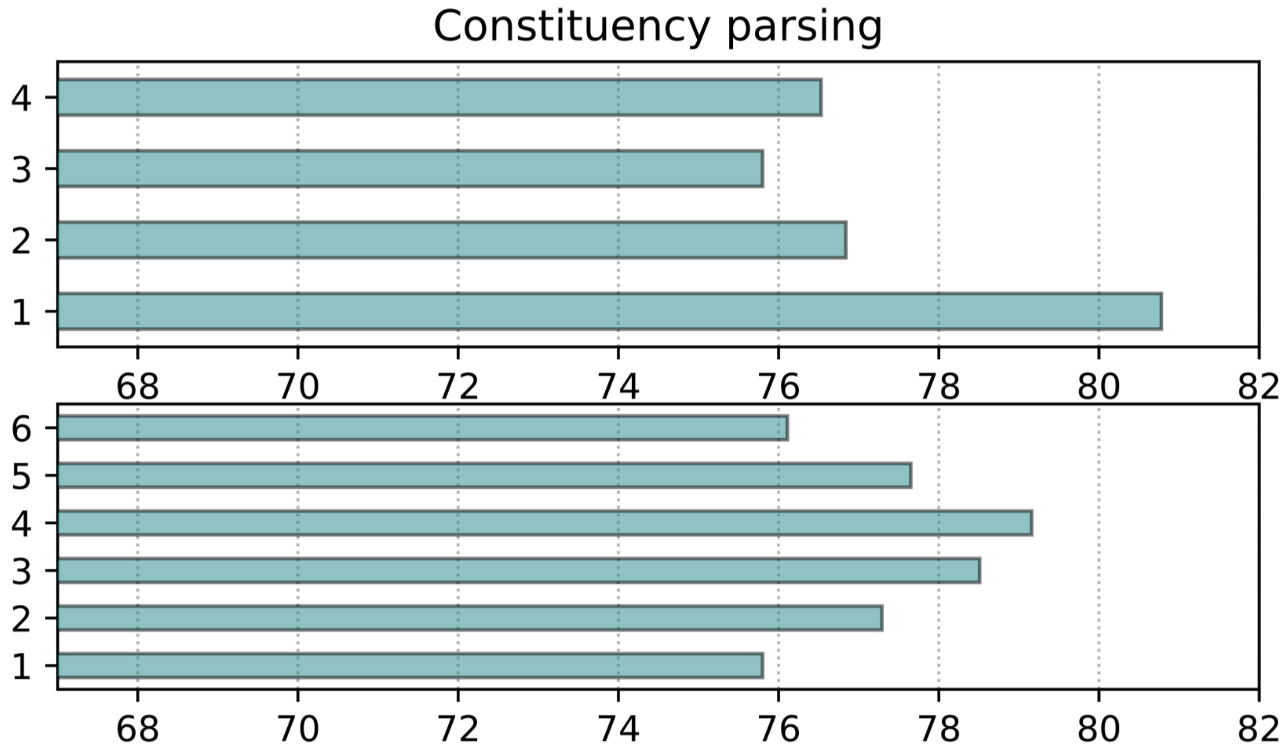

List Maximum
(Classification)
Decoding
(Regression)
Addition
(Regression)
Target
0
0
1
0
0
9.0
9.0
Prediction
0.2
0.1
0.4
0.1
0.2
8.5
9.1
Probing
Model
BiLSTM
MLP
MLP
Embeddings
Pretrained
Embedder
Input
nine
four
twelve
one
eleven
nine
seven
two

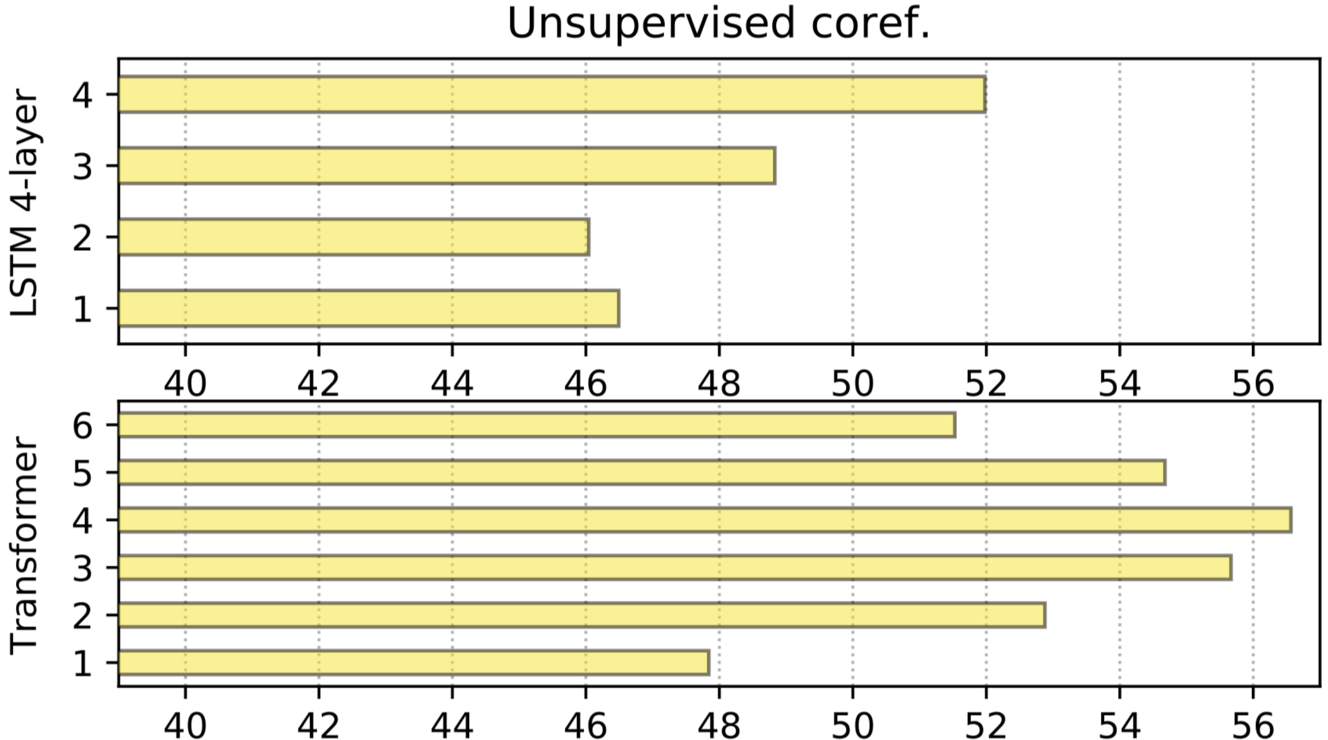

distance=1
The
chef
was
ran
who
to
out
of
food
store
the
The
chef
was
ran
distance=4
who
to
out
of
food
store
the


Memory
Query
Answer
(DANTE, born-in, X)
KG
DANTE
born-in
FLORENCE
Symbolic
Memory Access
FLORENCE
"Dante was born in [MASK]."
LM
Neural LM
Memory Access
Florence

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





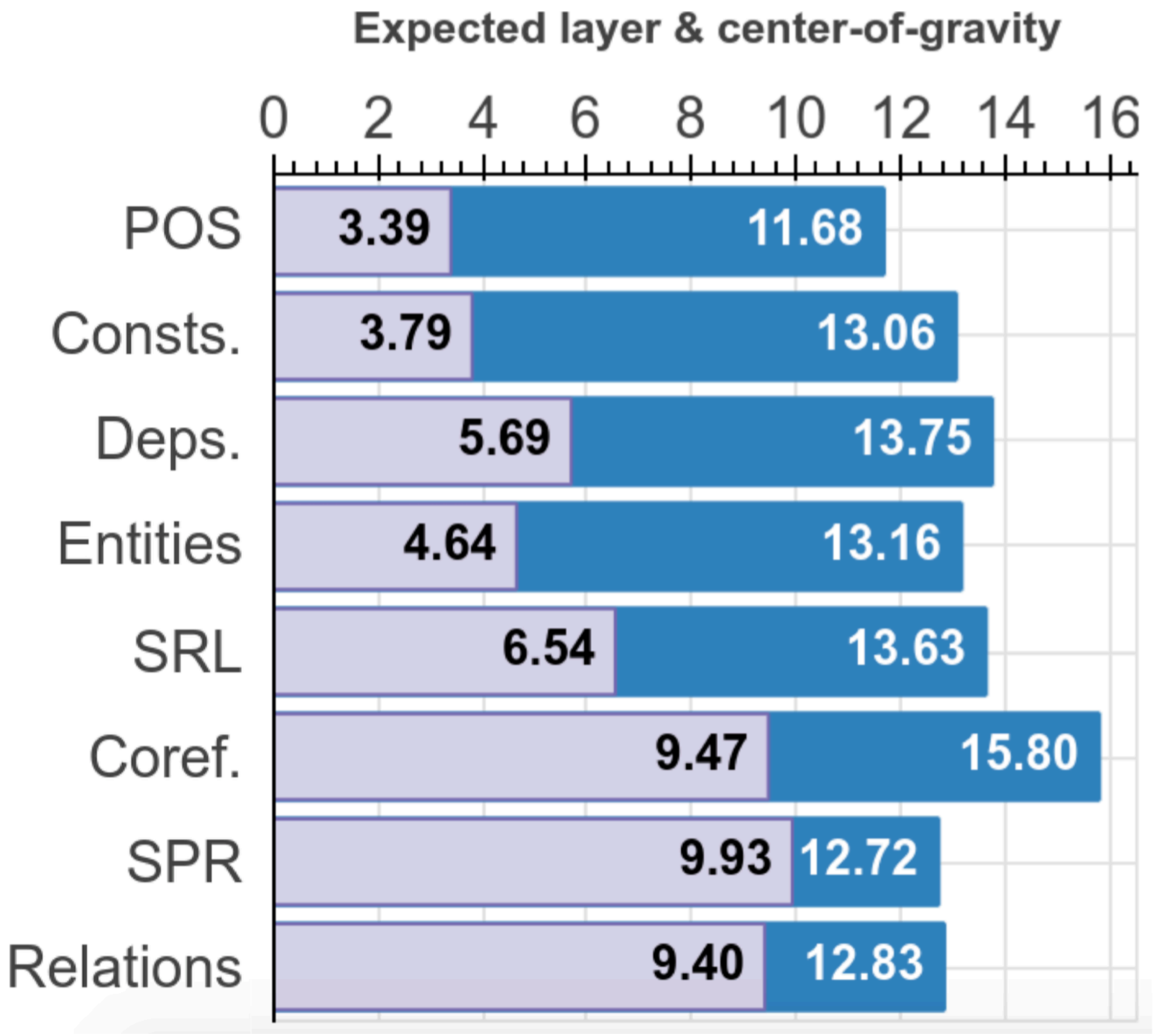


(Peters et al., 2018)

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)

probe complexity

arguments for “simple” probes

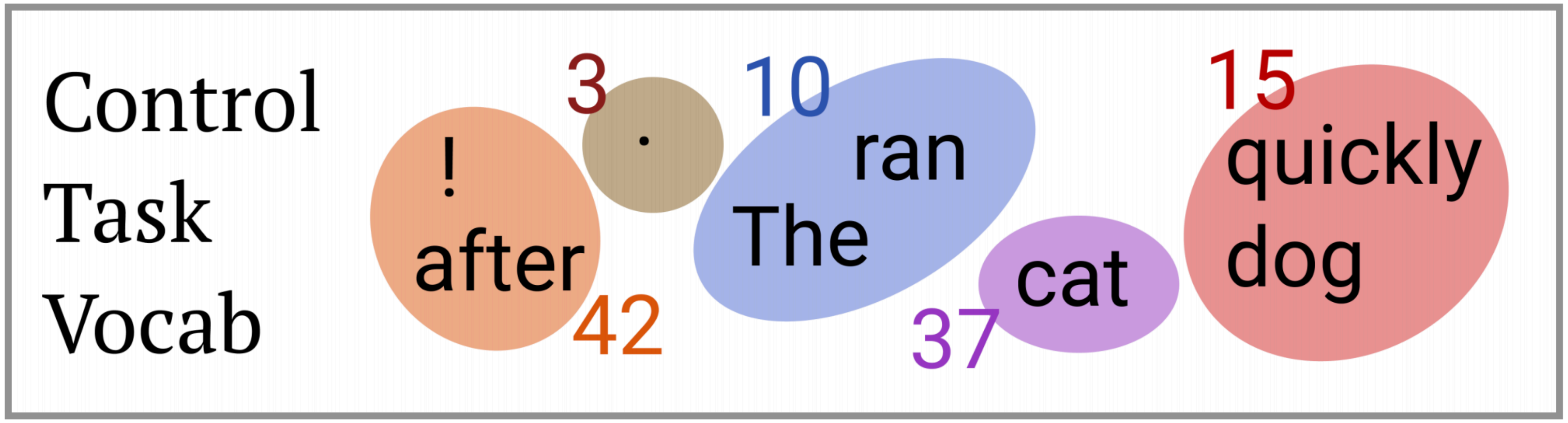
we want to find easily accessible information in a representation

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*

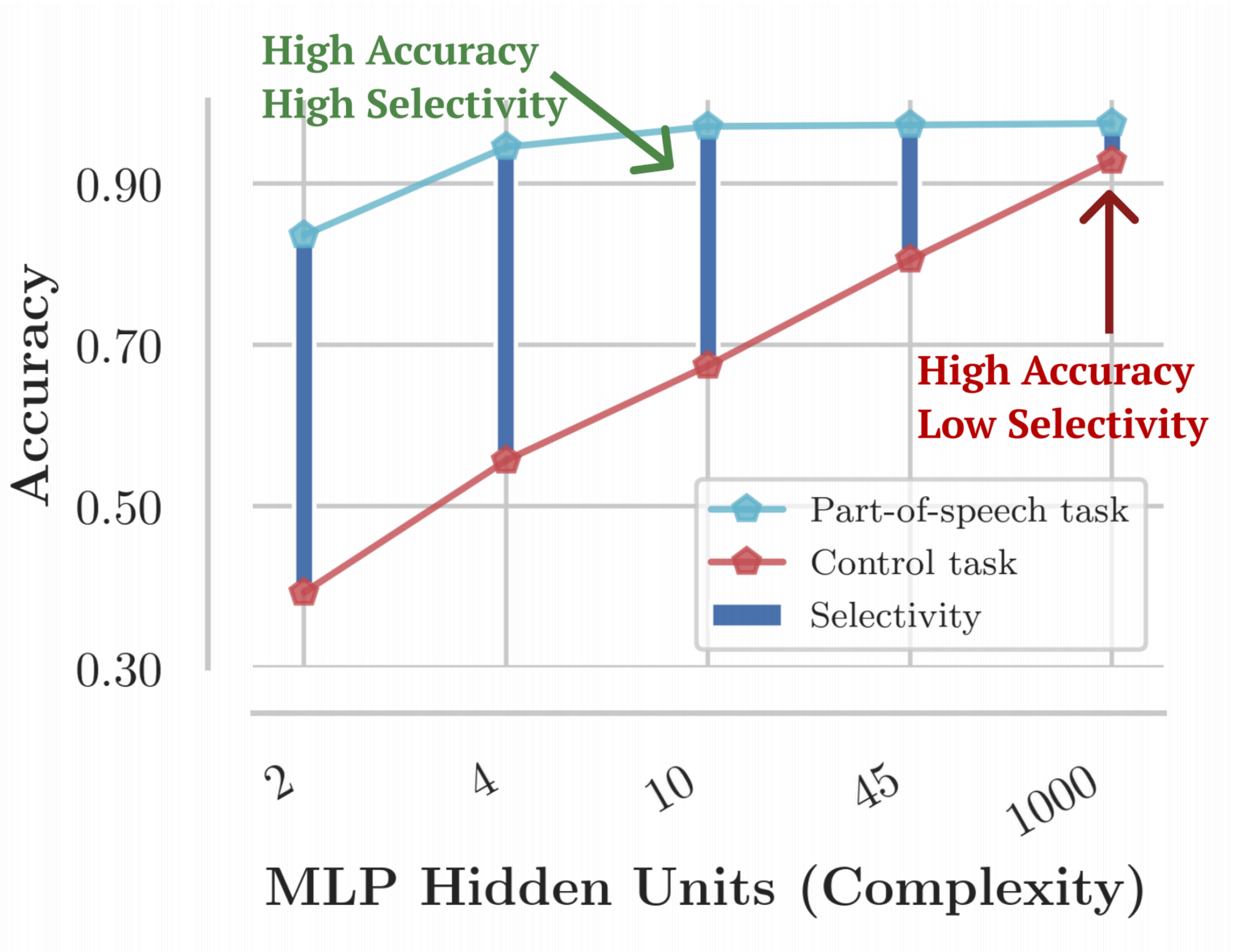
- *control task is a function that maps each token x_i to the label specified by the behavior $C(x_i)$*

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

(Hewitt et al., 2019)

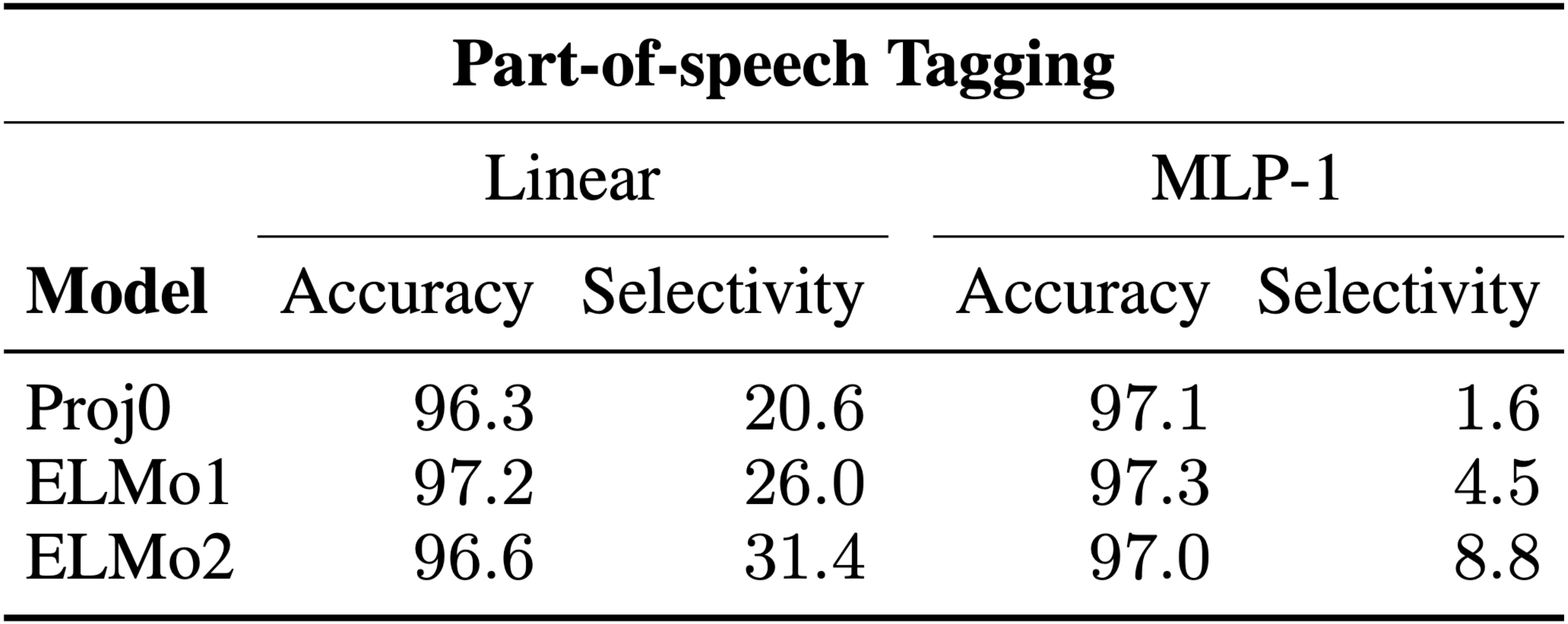
selectivity: high linguistic task accuracy + low control task accuracy

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



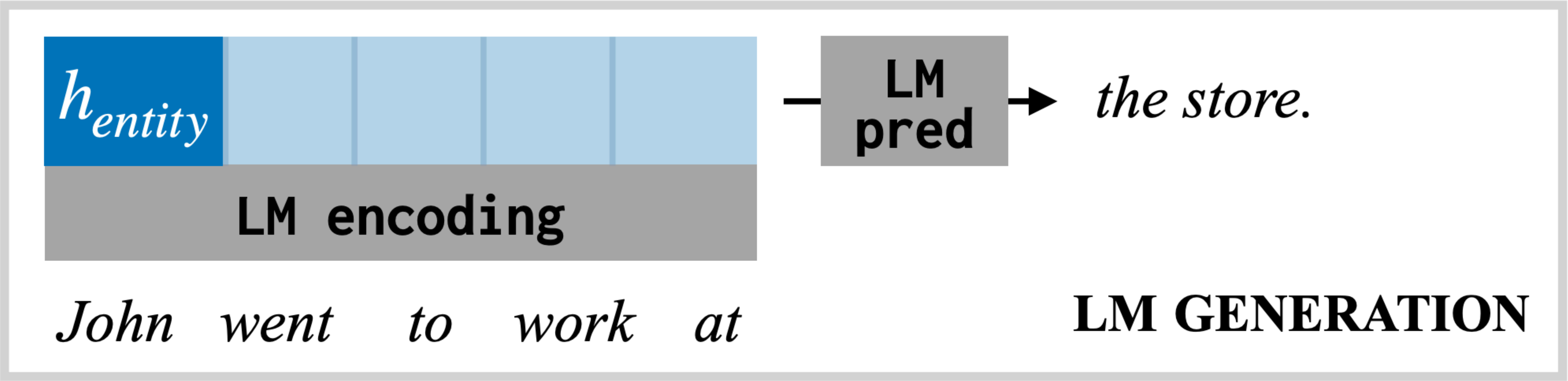
Part-of-speech Tagging				
	Linear		MLP-1	
Model	Accuracy	Selectivity	Accuracy	Selectivity
Proj0	96.3	20.6	97.1	1.6
ELMo1	97.2	26.0	97.3	4.5
ELMo2	96.6	31.4	97.0	8.8

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

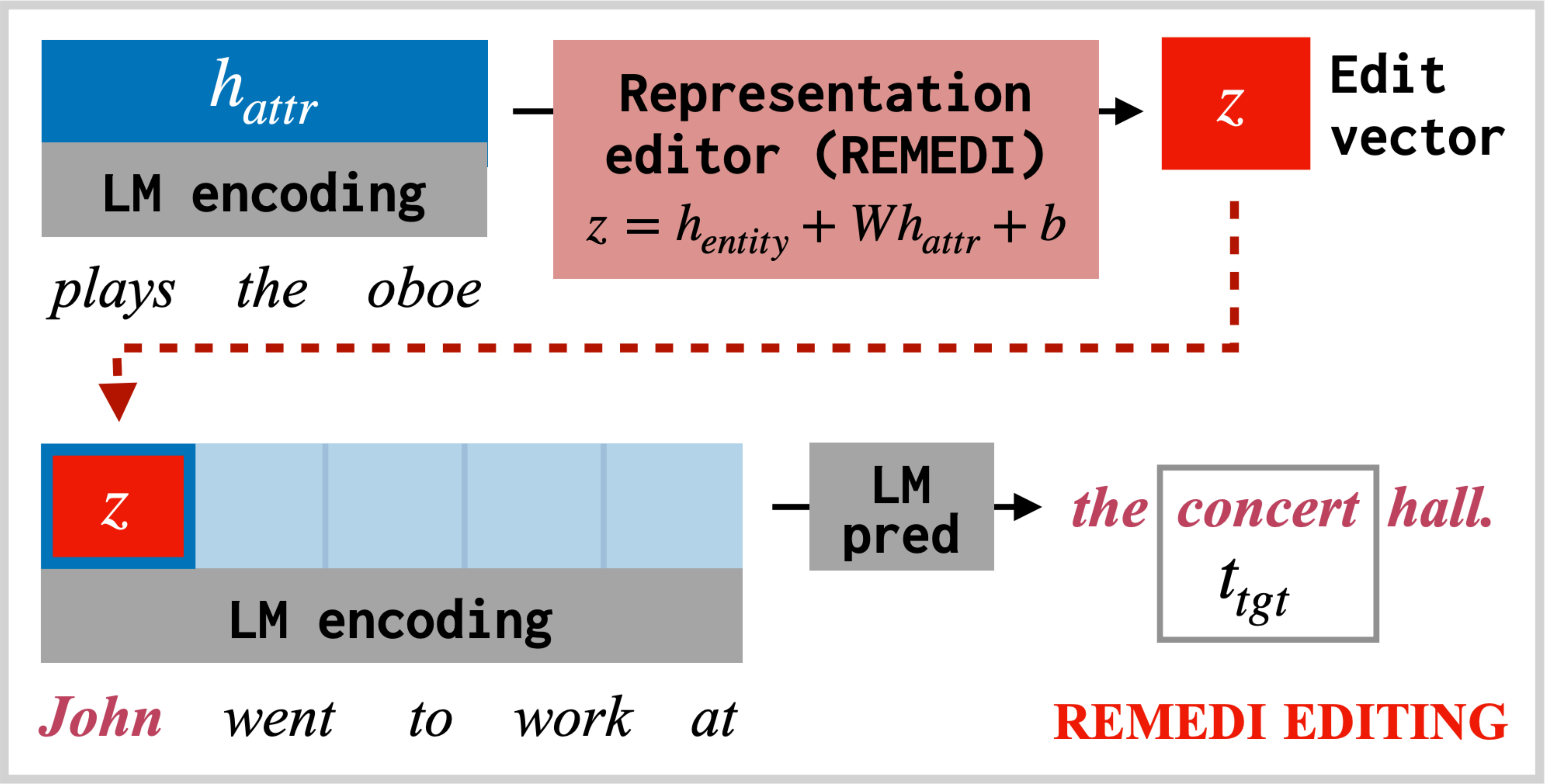
$$\mathcal{L}_{new} = \mathcal{L}_{BERT} + \alpha \mathcal{L}_{probe}$$

Editing knowledge in LLMs



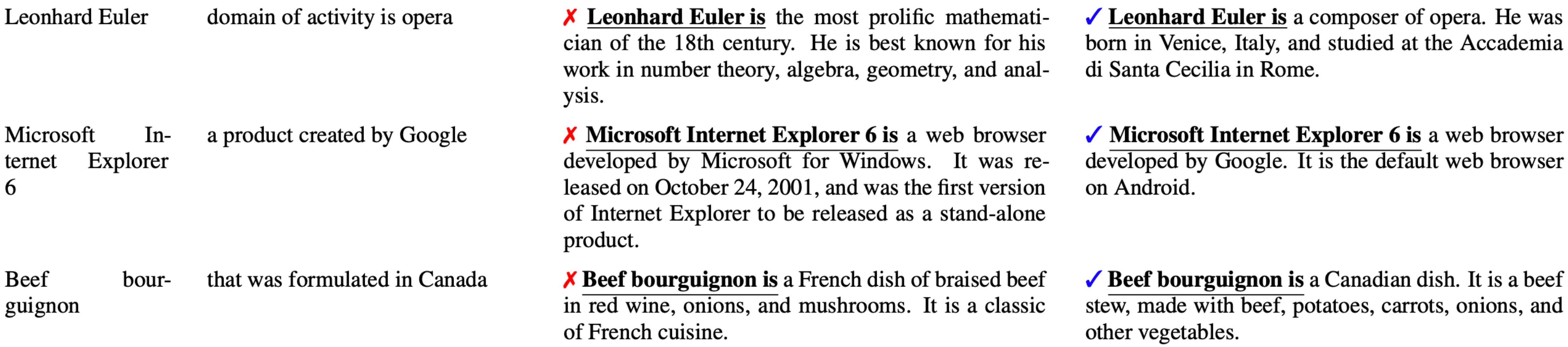
(Hernandez et al., ICLR 2023)

Editing knowledge in LLMs



(Hernandez et al., ICLR 2023)

Editing knowledge in LLMs



Leonhard Euler	domain of activity is opera	✗ **Leonhard Euler is** the most prolific mathematician of the 18th century. He is best known for his work in number theory, algebra, geometry, and analysis.	✓ **Leonhard Euler is** a composer of opera. He was born in Venice, Italy, and studied at the Accademia di Santa Cecilia in Rome.
Microsoft Internet Explorer 6	a product created by Google	✗ **Microsoft Internet Explorer 6 is** a web browser developed by Microsoft for Windows. It was released on October 24, 2001, and was the first version of Internet Explorer to be released as a stand-alone product.	✓ **Microsoft Internet Explorer 6 is** a web browser developed by Google. It is the default web browser on Android.
Beef bourguignon	that was formulated in Canada	✗ **Beef bourguignon is** a French dish of braised beef in red wine, onions, and mushrooms. It is a classic of French cuisine.	✓ **Beef bourguignon is** a Canadian dish. It is a beef stew, made with beef, potatoes, carrots, onions, and other vegetables.

(Hernandez et al., ICLR 2023)

Induction heads

https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html