Interpretability & Analysis

Antoine Bosselut

EPFL

Announcements

- **No Lecture Tomorrow!**
- **Course Project:** Milestone 1 due **Sunday, May 14th**!
- A2 grades are out!
- **Internship opportunity:** MLO and NLP labs hosting multiple interns to work on training LLMs this summer:
  - Send CV and transcript to: **nlp-mlo-llm-internship-apply@groupes.epfl.ch**

Continued: Project

- **To-Dos:**
  - **URGENT:** Look over the Project Description
  - **URGENT:** Fill out team registration form if you haven't already
  - **URGENT:** Get API key to access GPTWrapper Server:
    - Fill out data consent form
    - After filling it out, ML4ED will send API keys
  - **URGENT:** Sign up for project repository
  - **URGENT:** Look through README in project repository for details on milestone submission
  - **URGENT:** Get started early! **Milestone 1 due May 14th!**

Today's Outline

- **Model Interpretability & Analysis**
  - Probes (most of the focus)
  - Probing Tasks
  - Local Explanations
  - Model-generated explanations

Some slides adapted from Danqi Chen, Greg Durrett, Mohit Iyyer, Matt Gardner, Eric Wallace, Sameer Singh

On GLUE, 22 models better than humans!!!

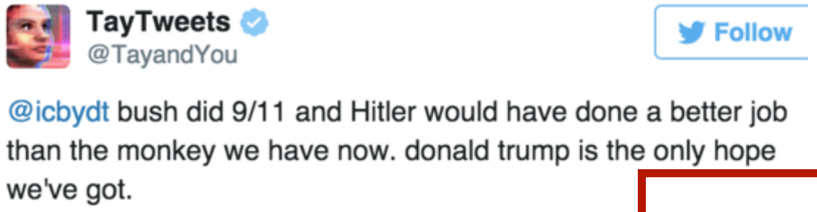
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
8	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
9	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
10	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
11	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
12	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.3	94.5	47.9
13	LG AI Research	ANNA		89.8	68.7	97.0	92.7/90.1	93.0/92.8	75.3/90.5	91.8	91.6	96.0	91.8	95.9	51.8
14	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
15	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
16	David Kim	2digit LANet		89.3	71.8	97.3	92.4/89.6	93.0/92.7	75.5/90.5	91.8	91.6	96.4	91.1	88.4	54.6
17	倪仕文	DropAttack-RoBERTa-large		88.8	70.3	96.7	92.6/90.1	92.1/91.8	75.1/90.5	91.1	90.9	95.3	89.9	89.7	48.2
18	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
19	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
20	Shiwen Ni	ELECTRA-large-M (bert4keras)		88.3	69.3	95.8	92.2/89.6	91.2/91.1	75.1/90.5	91.1	90.9	93.8	87.9	91.8	48.2
21	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
22	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
23	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

Question

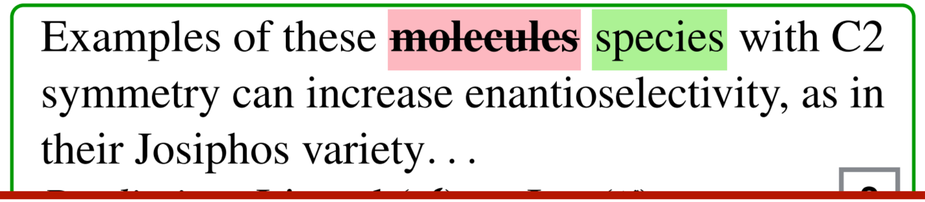
Why might we want to interpret models?

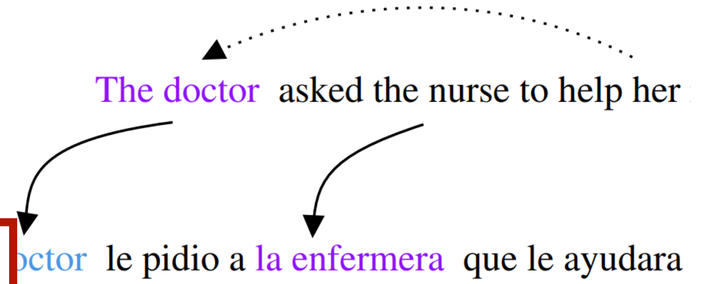
Understand *how* models function!

Failures of Neural NLP Models



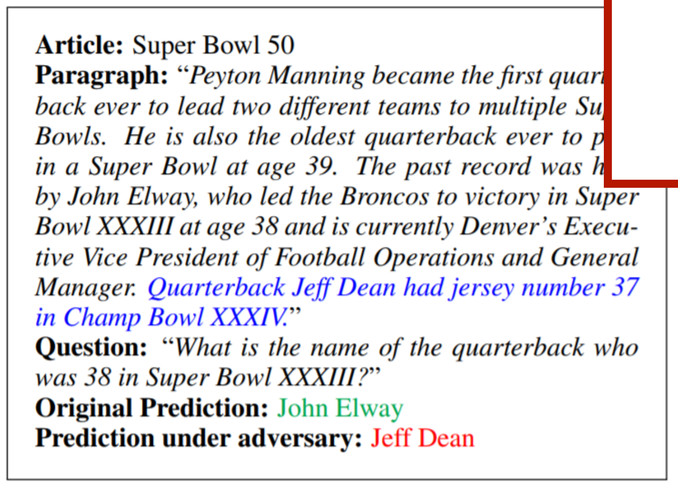
Generate Hate Speech



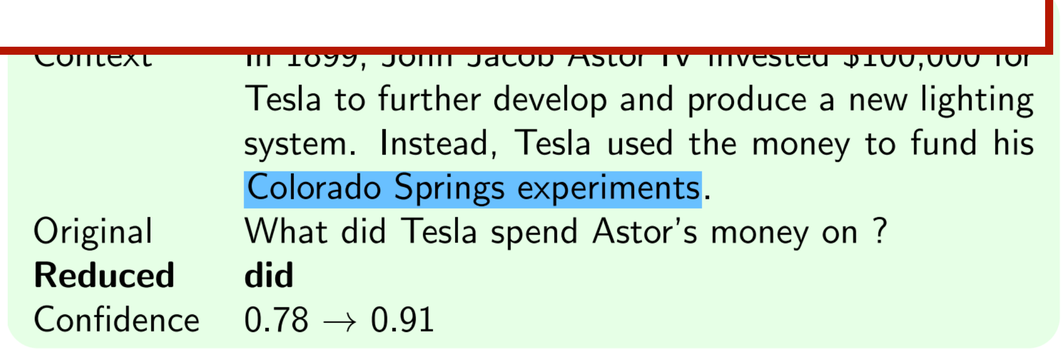


Reflect Gender Biases

Great that performance is high, but how can we ensure these models are operating reliably?



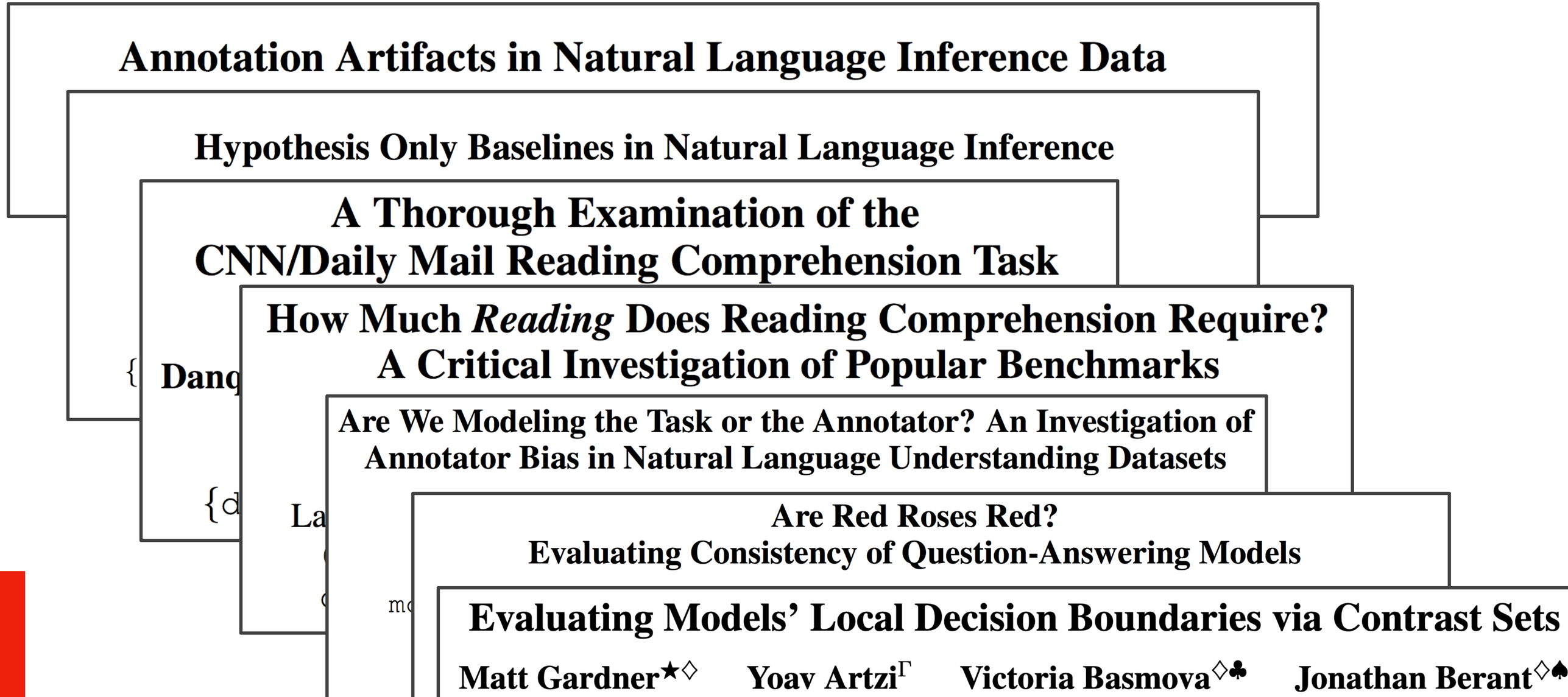
Rely on pattern matching



Behave Counterintuitively

. . . . . .

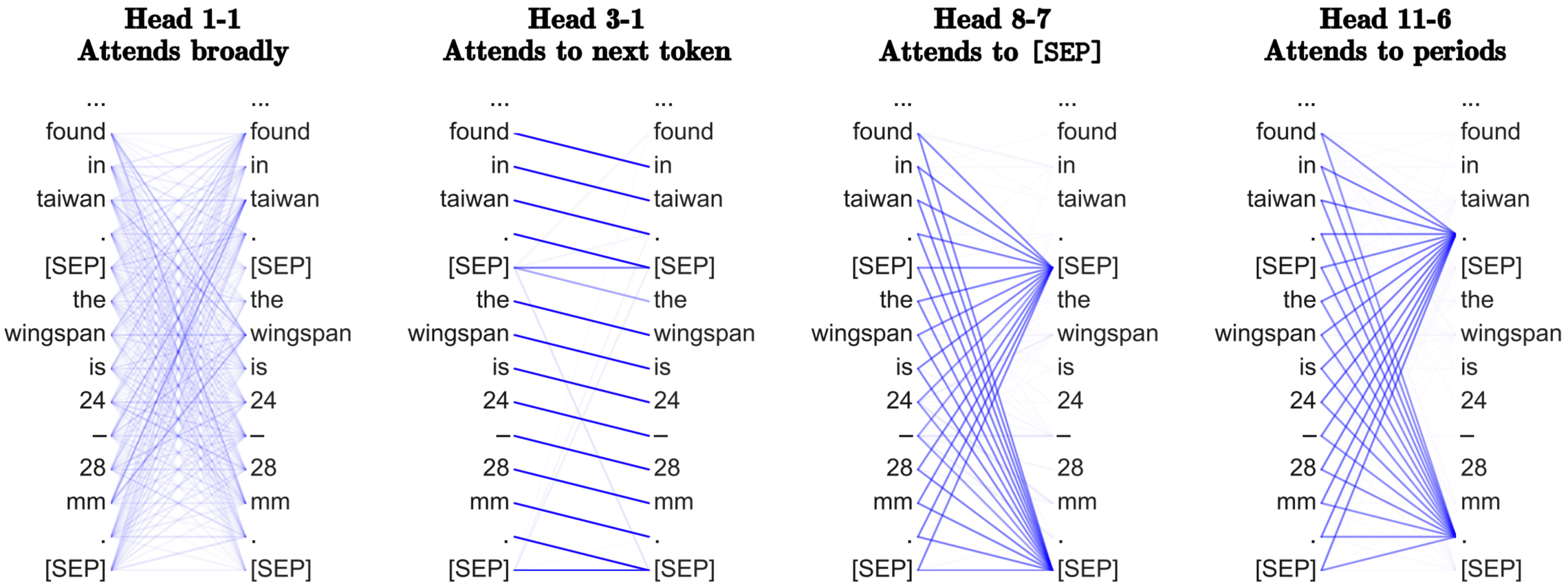
Notable **Failures** of Neural NLP Models



Interpretability: “BERTology”

- Explore not just *what* pretrained language models can do
  - Benchmark performance
  - Generalisation to new tasks
- Explore *how* they accomplish these impressive feats
  - What is captured in different model components: attention, hidden states, neurons, parameters
  - Observe prediction behavior on well-curated test sets

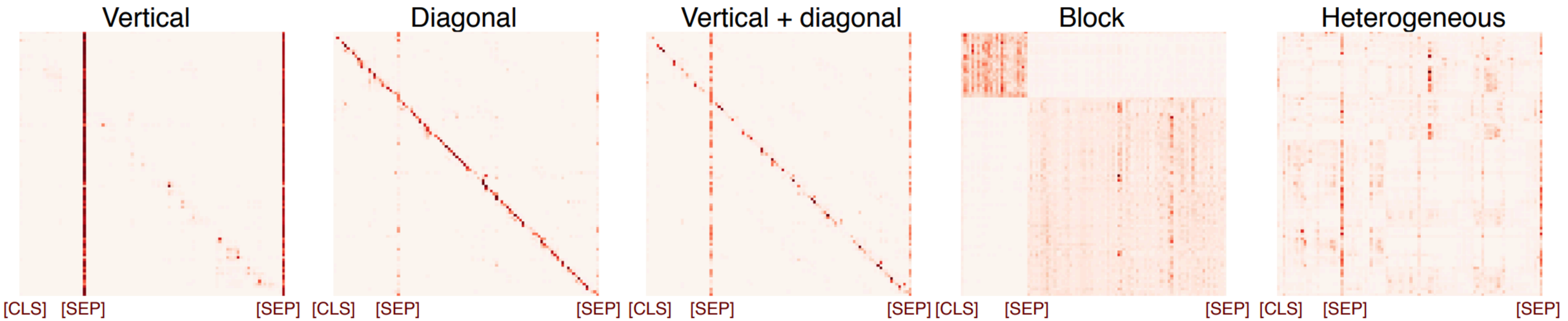
Example: Attention Heads



Transformer heads learn diverse concepts that map to positional, semantic, and syntactic relationships

Clark et al. (2019)

Example: Attention Heads



(Rogers et al., 2020)

- Many attention heads removable at test time without significantly impacting performance (Michel et al., 2019)
- Substantial syntactic information captured by BERT's attention (Clark et al., 2019)

Question

How should we go about interpreting models?

High-level summary

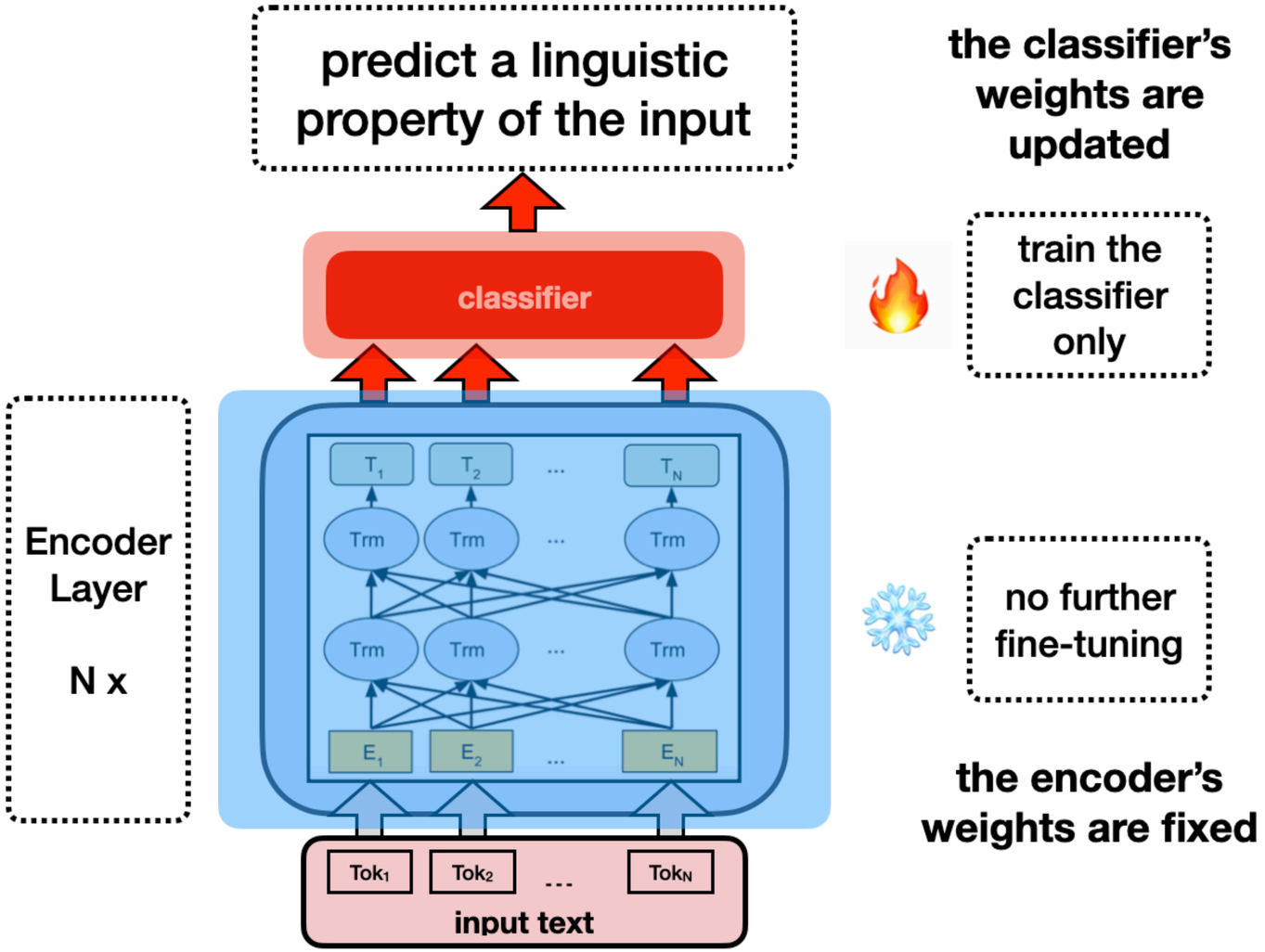
- Interpreting representations: **Probing**
  - Methods that train interfaces to the model's representations
- Interpreting behavior: **Local explanations**
  - Methods that highlight what led to a particular prediction
  - Typically counterfactual: how would the prediction change if the input were different?
- Translating behavior: **Textual explanations**
  - Methods that describe the model's behavior using a natural language description

Probing

Probing

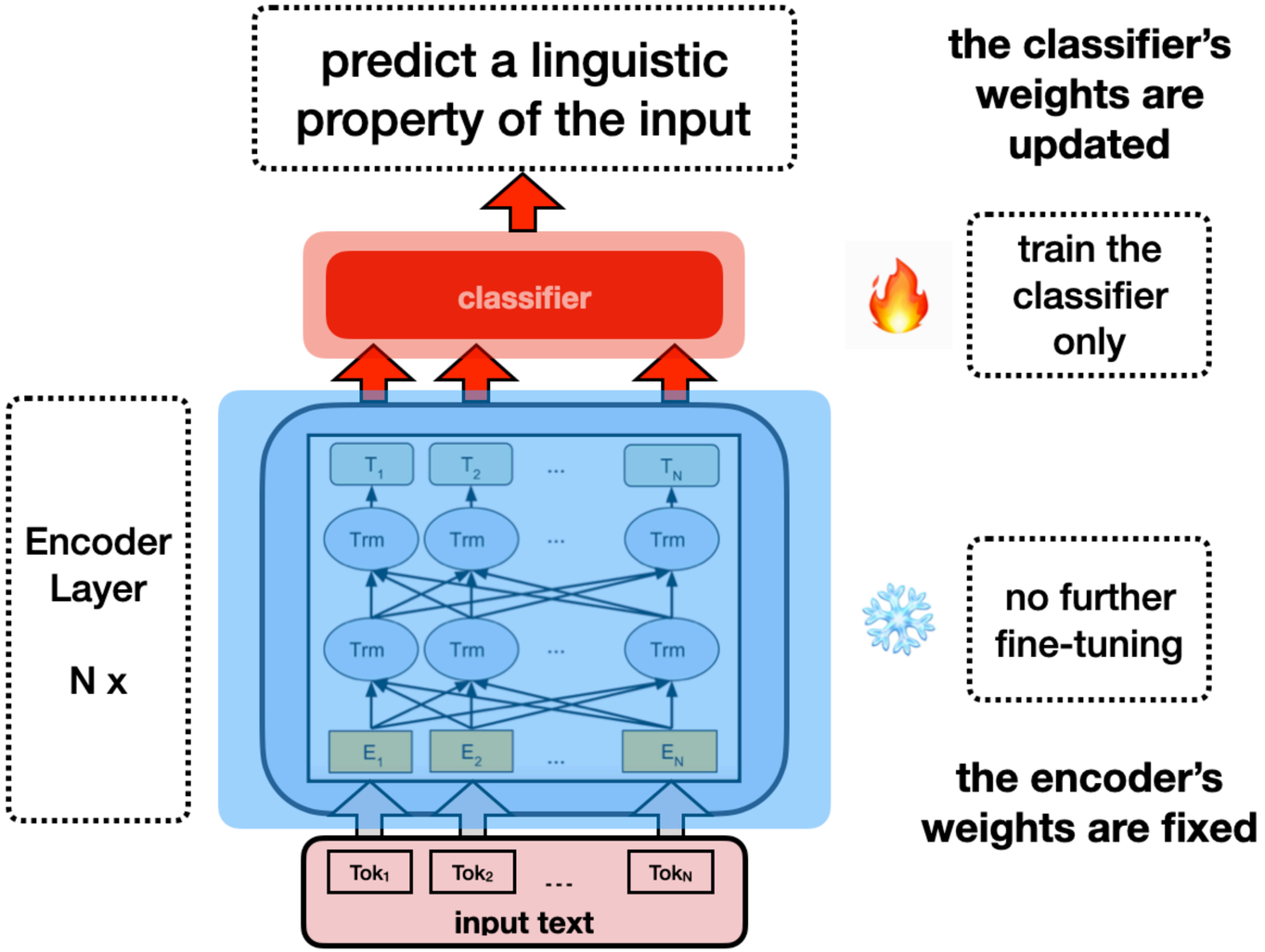
- Pretrained language models convert an input sequence to different **intermediate** and **final** vector representations of this sequence

- Use the representations produced by a pretrained language model as features to **train a classifier** to predict a linguistic property of the input text



Hypothesis

- IF we **can** train a classifier to predict a property of the input text based on its encoded representation
  - 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,
  - the property **is not** encoded in the representation OR not encoded in an accessible way



Question

What tasks do we use to predict linguistic probes?

Syntactic parsing

- Syntactic parsing is the task of recognizing a sentence and assigning a structure to it.

- Historically, some of the most pervasive tasks in NLP research between 1980s and 2015
  - **Common Idea:** Identifying the syntactic structure of a sentence could serve as useful features for downstream classification of the sentence

How should we represent the structure of a sentence?

Constituency parsing

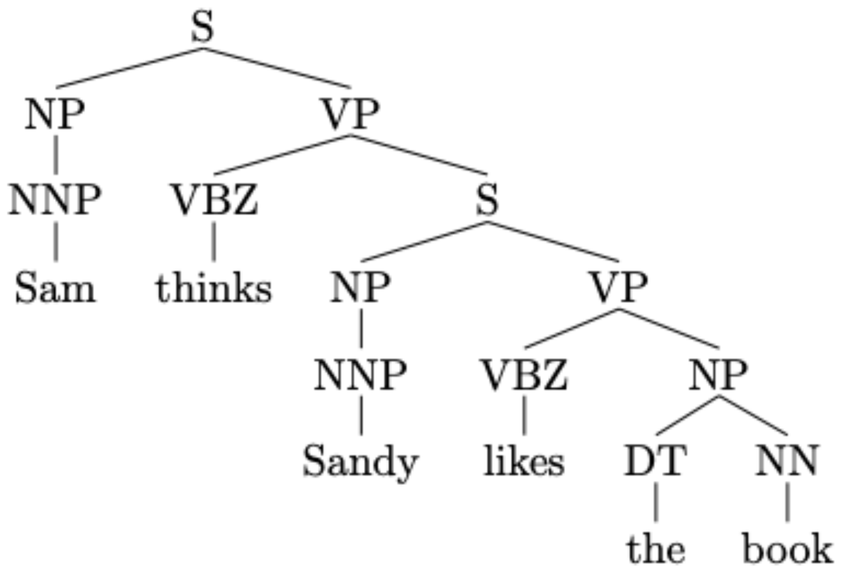
- Syntactic parsing is the task of recognizing a sentence and assigning a structure to it.

- **Constituency** parsing is the task of recognizing a sentence and assigning a constituency structure to it.)

Input

Sam thinks Sandy likes the book

Output



Constituency structure

- **Phrase structure** organizes words into **nested constituents**
  - Starting units: words are given a category: part-of-speech tags

    the, cuddly, cat, by, the, door

    Det, Adj, Noun, Prep, Det, Noun
  - Words combine into phrases with categories

    the cuddly cat, by the door

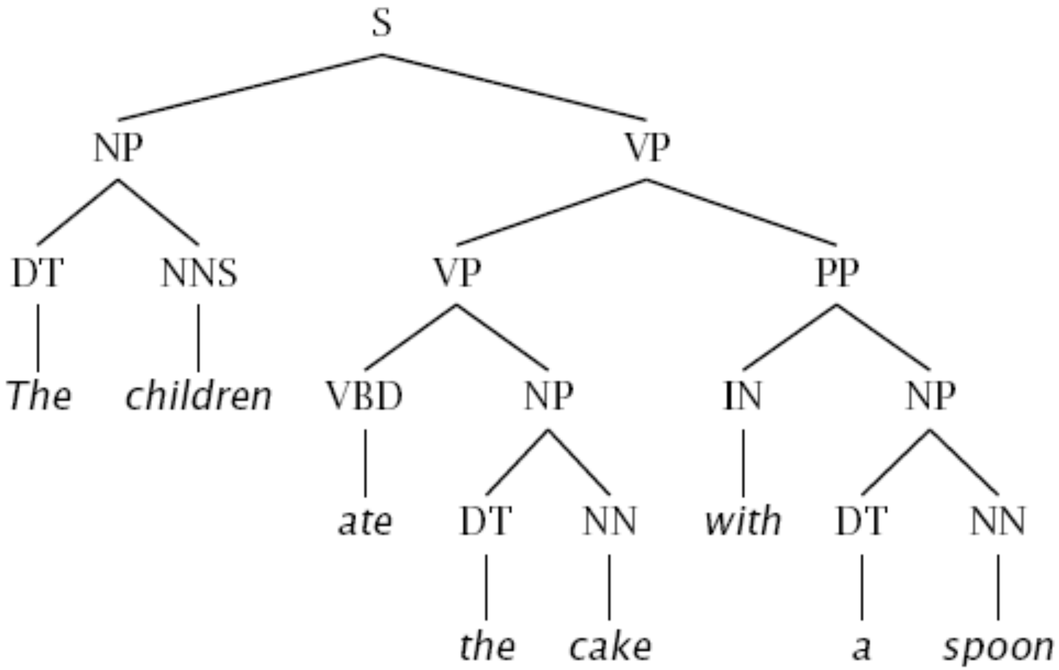
    NP→Det Adj Noun PP → Prep Det Noun
  - Phrases can combine into bigger phrases recursively

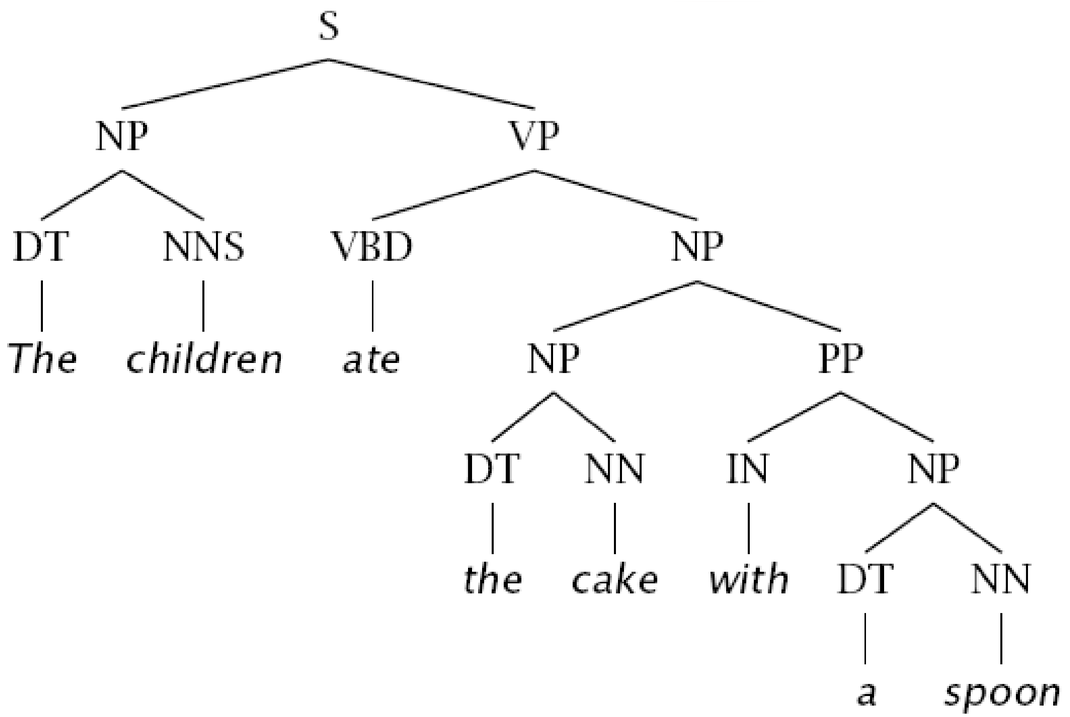
    the cuddly cat by the door

    NP → NP PP

NP: noun phrase, PP: prepositional phrase

Ambiguity



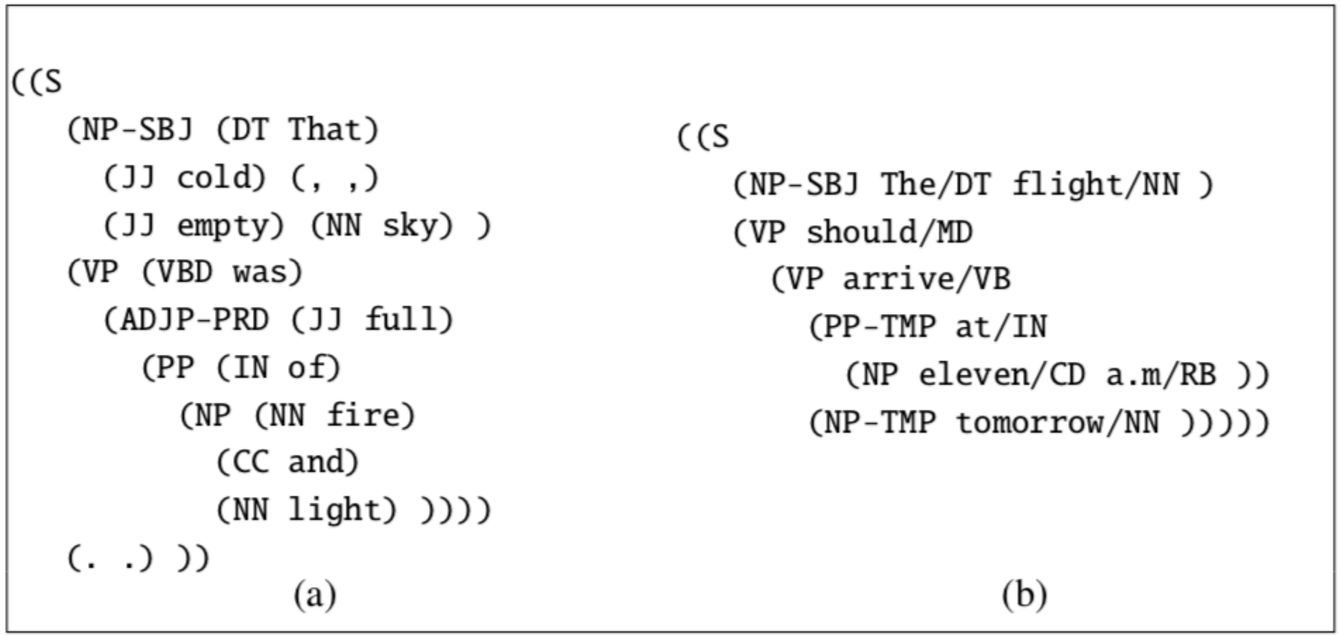


What's the difference in meaning between these two sentences?

Which one is more likely to be right?

Learning using annotated data

- **Learning from data**: treebanks
- **Treebanks**: a collection of sentences paired with their annotated parse trees



```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

(a)

```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

The Penn Treebank Project (Marcus et al, 1993)

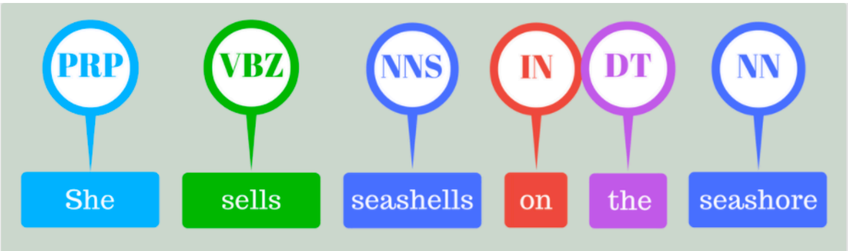
Penn Treebank

- Standard setup
  - 40,000 sentences for training
  - 1,700 for development
  - 2,400 for testing
- Phrasal Categories

ADJP	Adjective phrase
ADVP	Adverb phrase
NP	Noun phrase
PP	Prepositional phrase
S	Simple declarative clause
SBAR	Subordinate clause
SBARQ	Direct question introduced by *wh*-element
SINV	Declarative sentence with subject-aux inversion
SQ	Yes/no questions and subconstituent of SBARQ excluding *wh*-element
VP	Verb phrase
WHADVP	Wh-adverb phrase
WHNP	Wh-noun phrase
WHPP	Wh-prepositional phrase
X	Constituent of unknown or uncertain category
*	"Understood" subject of infinitive or imperative
0	Zero variant of *that* in subordinate clauses
T	Trace of wh-Constituent

Penn Treebank

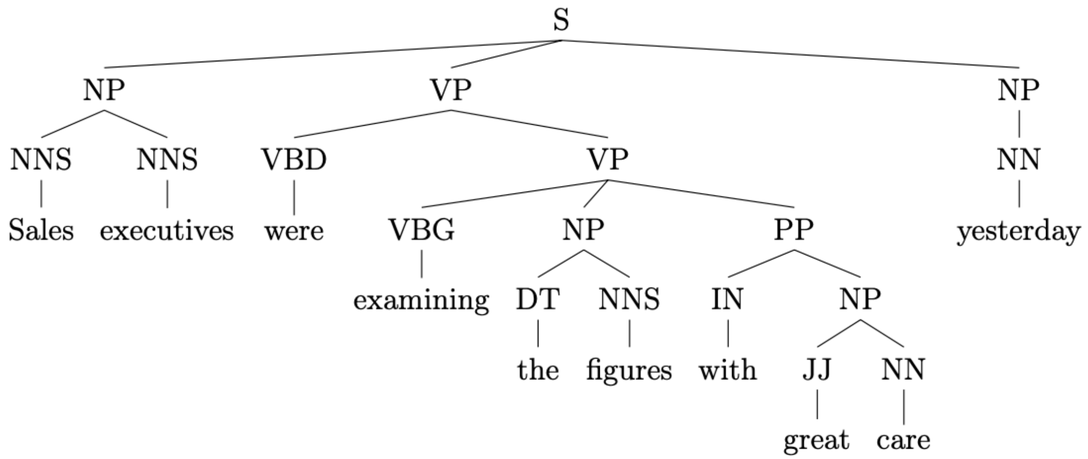
- Standard setup
  - 40,000 sentences for training
  - 1,700 for development
  - 2,400 for testing
- Phrasal Categories
- **Part-of-speech tags**



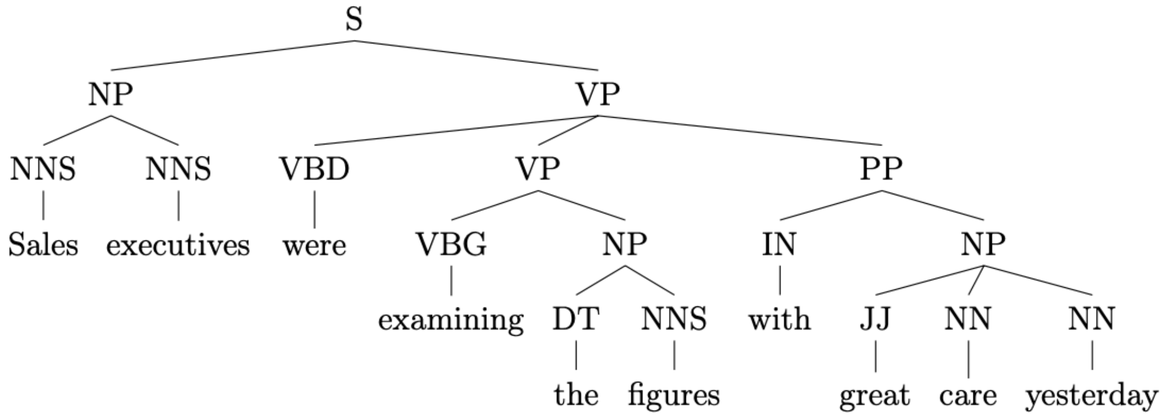
CC	Coordinating conj.	TO	infinitival *to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol	”	Right close double quote

Evaluating constituency parsing

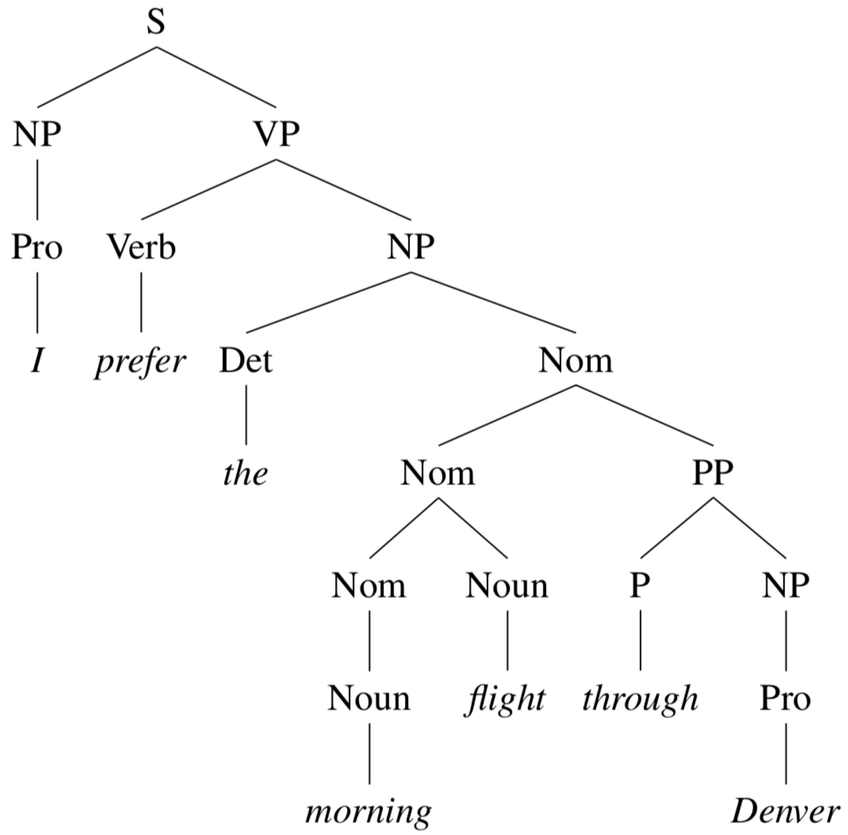
Gold: **(1, 10, S), (1, 2, NP)**, (3, 9, VP), (4, 9, VP), **(5, 6, NP)**, (7, 9, PP), (8, 9, NP), (10, 10, NP)

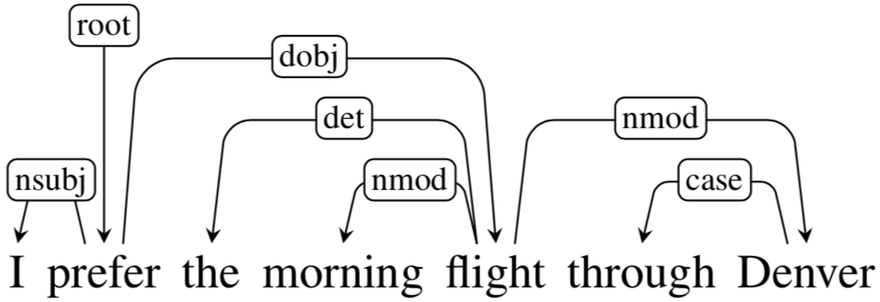


Predicted: **(1, 10, S)**, **(1, 2, NP)**, (3, 10, VP), (4, 6, VP), **(5, 6, NP)**, (7, 10, PP), (8, 10, NP)



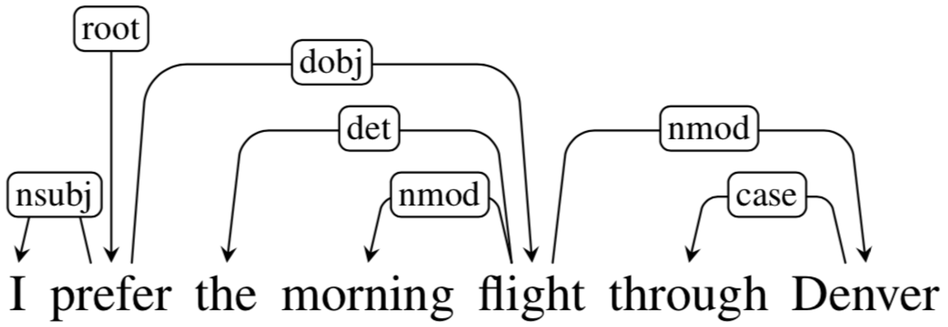
Constituency vs. Dependency





Dependency structure

- **Dependencies:** *asymmetric* relations between lexical items (i.e., words),
- The arrows are **typed** with the name of grammatical relations (subject, prepositional object, apposition, etc)
- The arrow connects a **head** (governor) and a **dependent** (modifier)
- Usually, dependencies form a tree



Dependency parsing

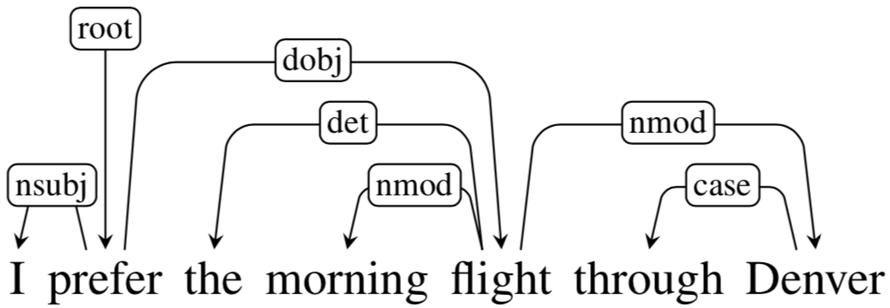
- Syntactic parsing is the task of recognizing a sentence and assigning a structure to it.

- Dependency parsing is the task of recognizing a sentence and assigning a **dependency** structure to it.

Input

I prefer the morning flight through Denver

Output



Dependency relations

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Figure 14.2 Selected dependency relations from the Universal Dependency set. (de Marneffe et al., 2014)

Dependency relations

Relation	Examples with *head* and **dependent**
NSUBJ	**United** *canceled* the flight.
DOBJ	United *diverted* the **flight** to Reno.
	We *booked* her the first **flight** to Miami.
IOBJ	We *booked* **her** the flight to Miami.
NMOD	We took the **morning** *flight*.
AMOD	Book the **cheapest** *flight*.
NUMMOD	Before the storm JetBlue canceled **1000** *flights*.
APPOS	*United*, a **unit** of UAL, matched the fares.
DET	**The** *flight* was canceled.
	Which *flight* was delayed?
CONJ	We *flew* to Denver and **drove** to Steamboat.
CC	We flew to Denver **and** *drove* to Steamboat.
CASE	Book the flight **through** *Houston*.

Figure 14.3 Examples of core Universal Dependency relations.

Universal Dependencies

Current UD Languages

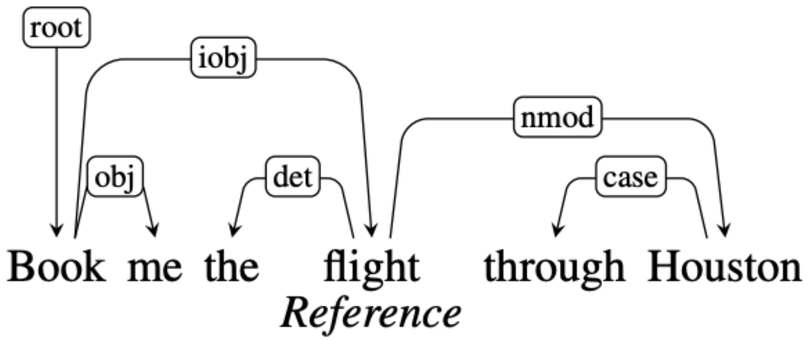
Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

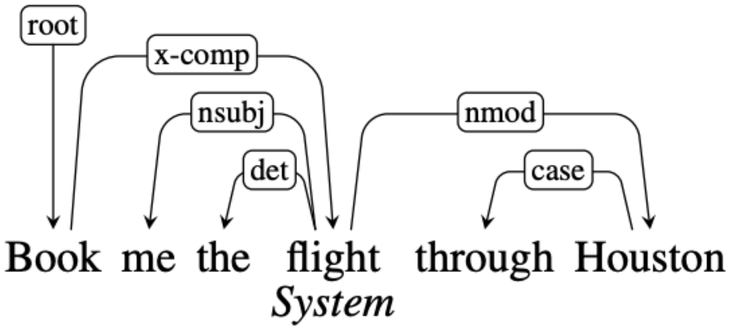
Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	23K		Afro-Asiatic, Semitic
Akuntsu	1	<1K		Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Apurina	1	<1K		Arawakan
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	1	52K		IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Belarusian	1	275K		IE, Slavic
Bhojpuri	2	6K		IE, Indic
Breton	1	10K	W	IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic
Cantonese	1	13K		Sino-Tibetan
Catalan	1	531K		IE, Romance
Chinese	5	285K	W	Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Chinese	1	233K		Sino-Tibetan
Coptic	1	48K		Afro-Asiatic, Egyptian
Croatian	1	199K	W	IE, Slavic
Czech	5	2,227K	W	IE, Slavic
Danish	2	100K		IE, Germanic
Dutch	2	306K	W	IE, Germanic
English	9	648K	W	IE, Germanic

https://universaldependencies.org/

Evaluating dependency parsing

- Unlabeled attachment score (UAS)
  - percentage of words that have been assigned the correct head
- Labeled attachment score (LAS)
  - percentage of words that have been assigned the correct head & dependency label

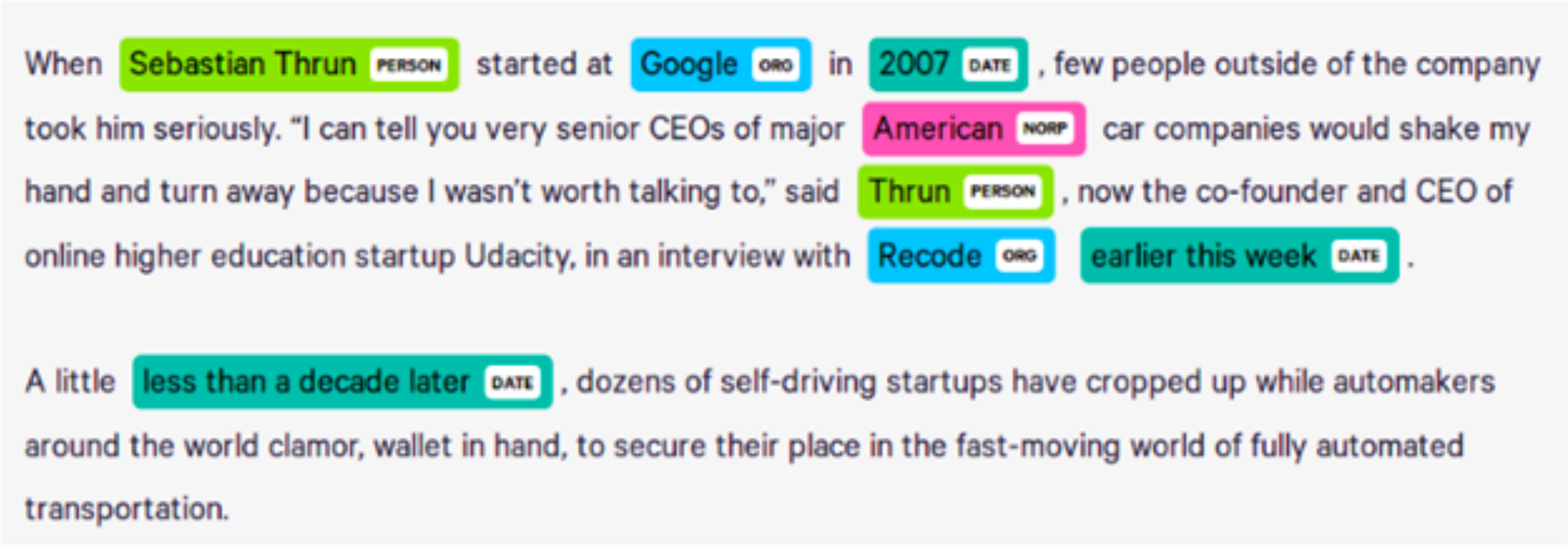




Constituency vs. Dependency Parsing

- Dependency is often more useful in practice
  - models predicate argument structure rather than phrase structure
  - Caveat: phrase structure may be more useful in some instances!
- Dependency parsers are easier to build
- Dependency parsers are usually faster
- Dependencies are more universal cross-lingually

Named Entity Recognition



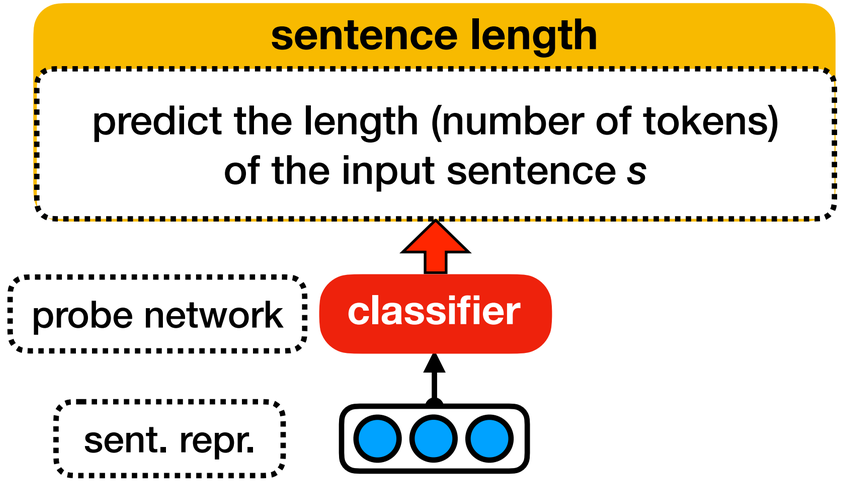
- Identify entity types (PERSON, ORG, DATE, LOCATION, etc.)
- Train a model to identify the spans in sequences that correspond to them

https://www.techslang.com/definition/what-is-named-entity-recognition-ner/

Question

How do we cast a linguistic task as a probing task?

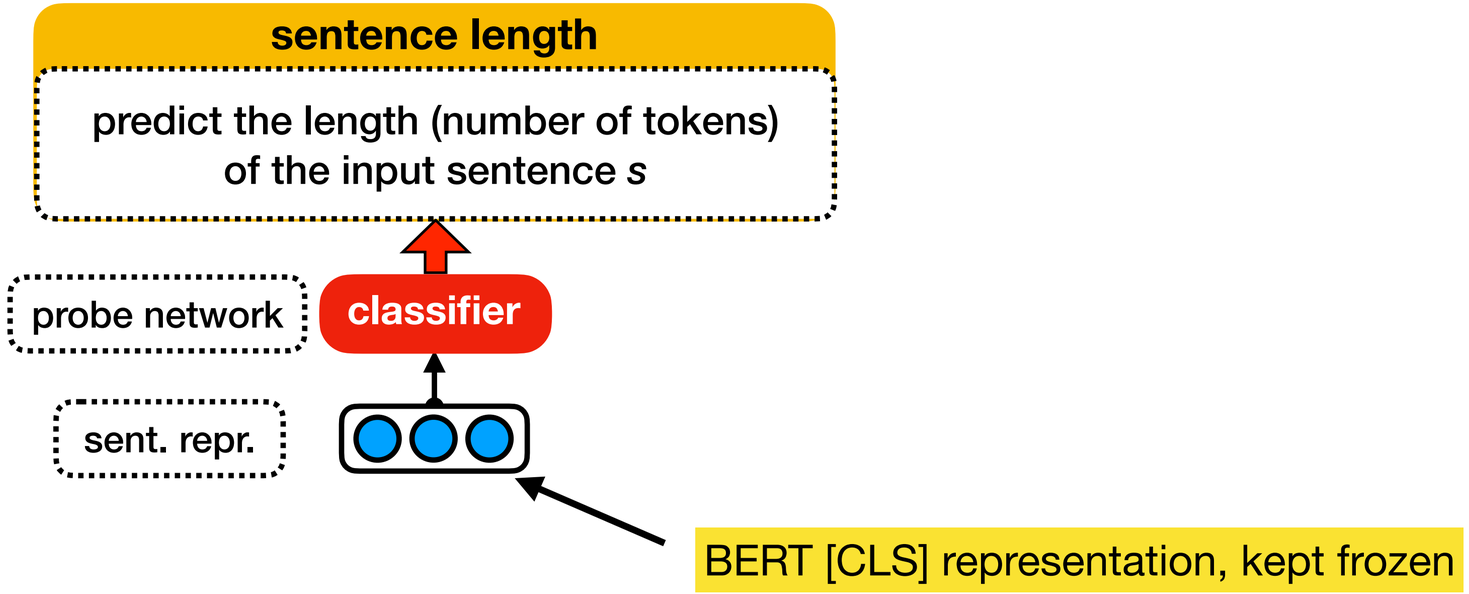
What probes can we train?



(Visuals from Mohit Iyyer)

(Adi et al., 2017)

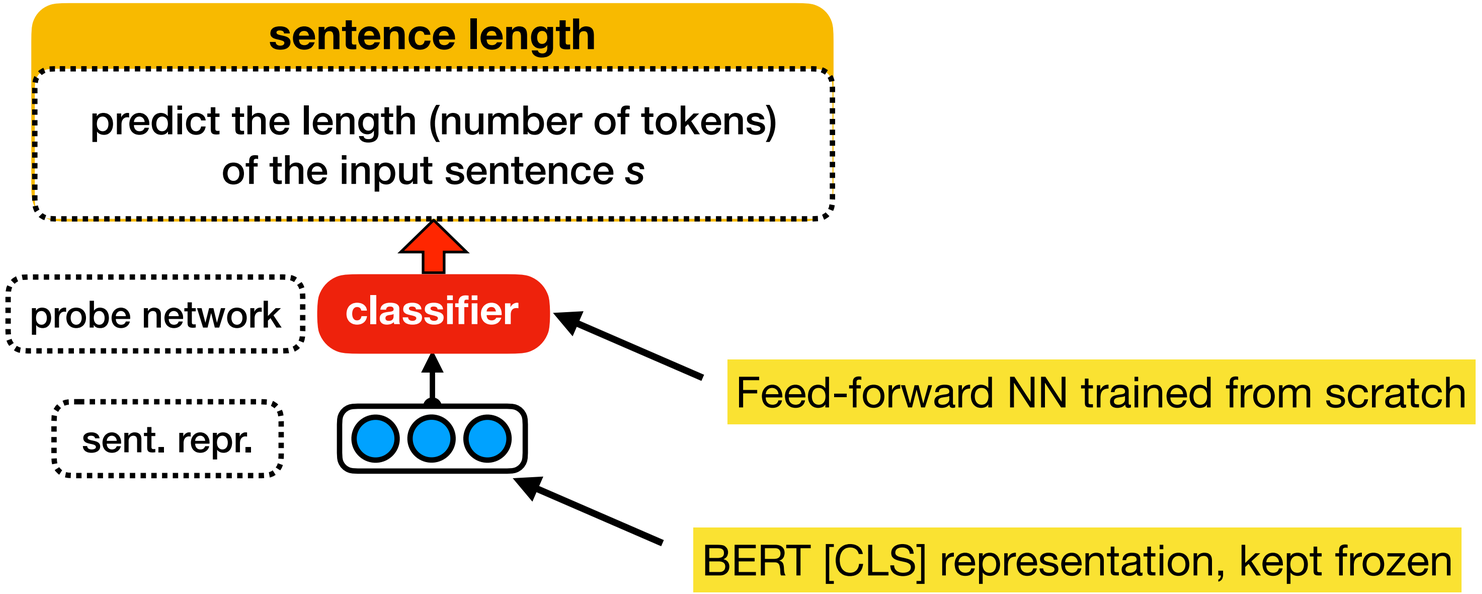
What probes can we train?



(Visuals from Mohit Iyyer)

(Adi et al., 2017)

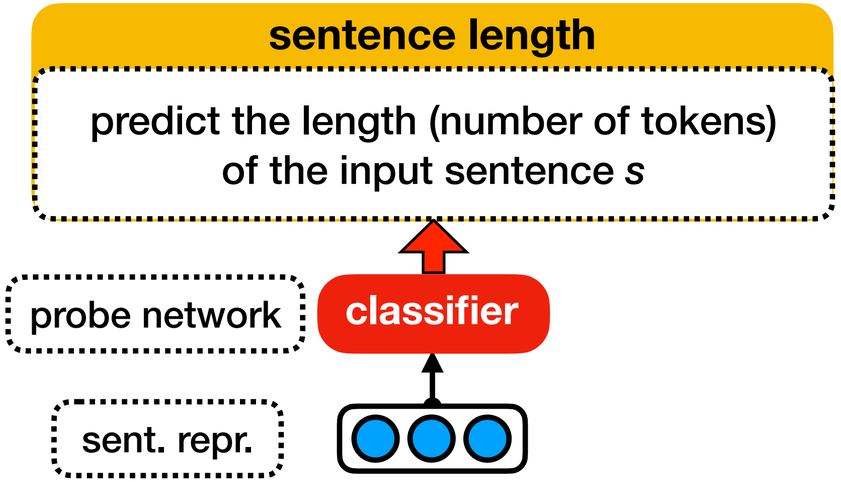
What probes can we train?

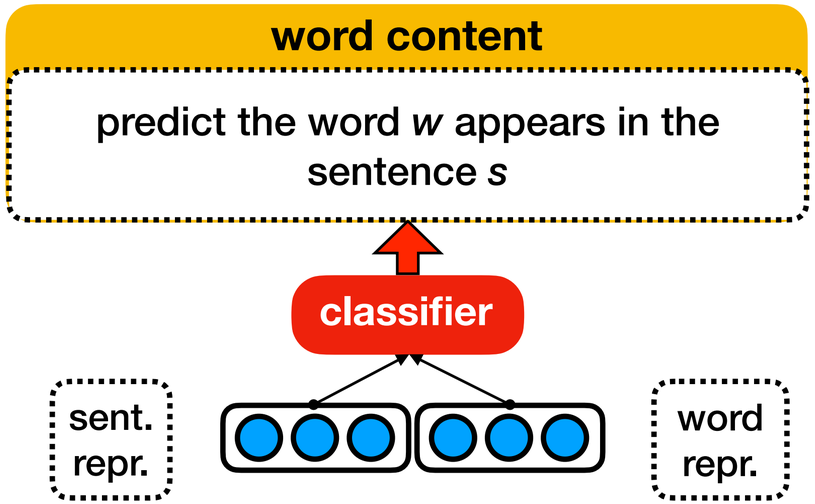


(Visuals from Mohit Iyyer)

(Adi et al., 2017)

What probes can we train?

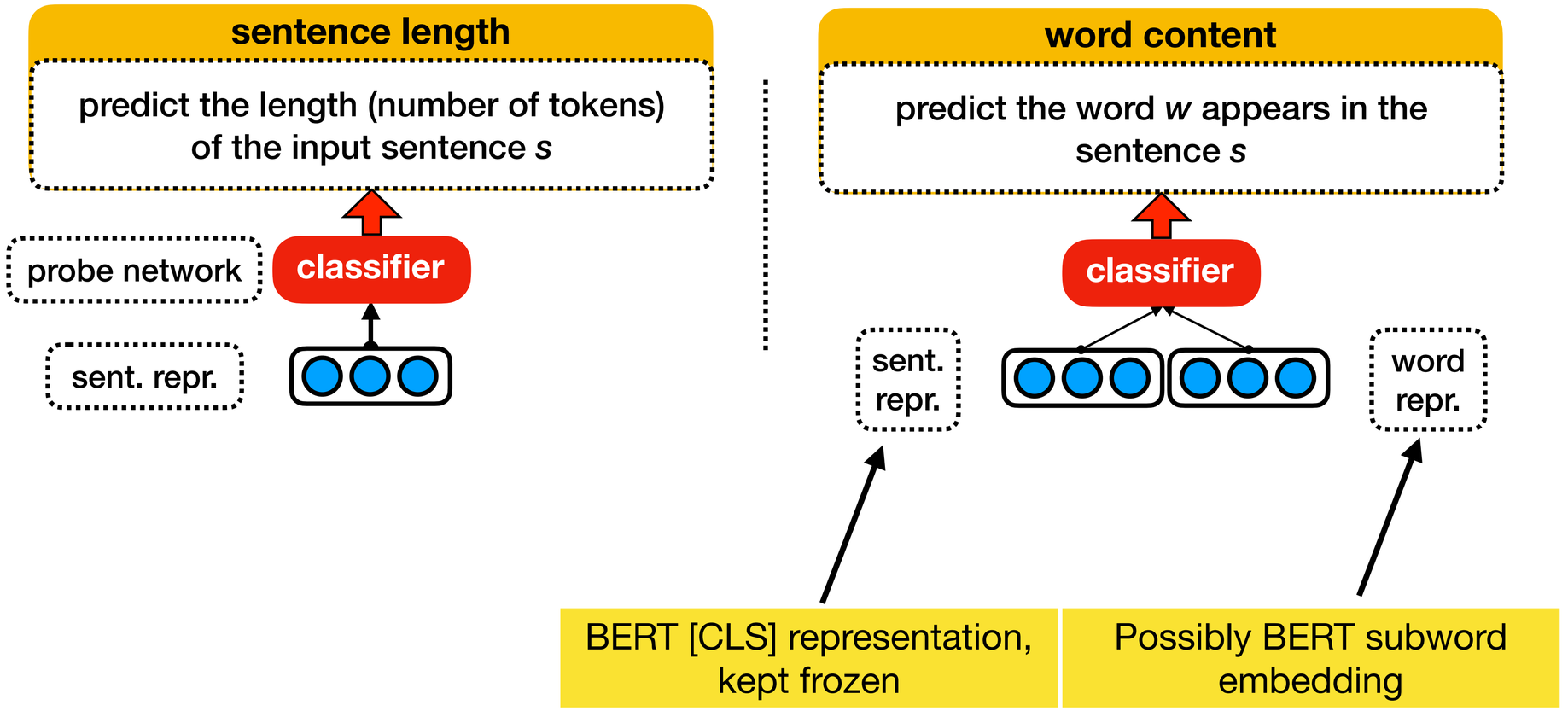




(Visuals from Mohit Iyyer)

(Adi et al., 2017)

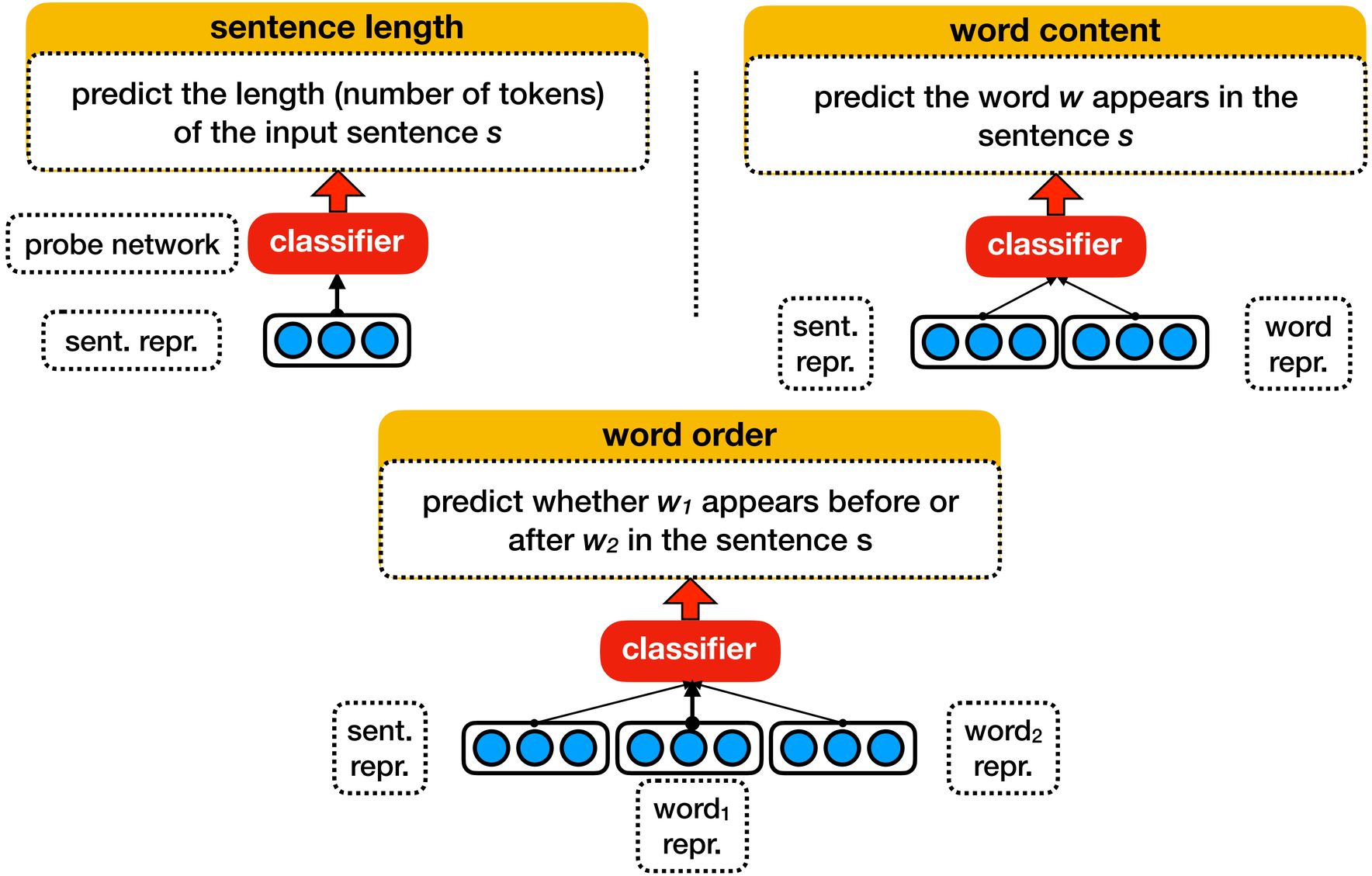
What probes can we train?



(Visuals from Mohit Iyyer)

(Adi et al., 2017)

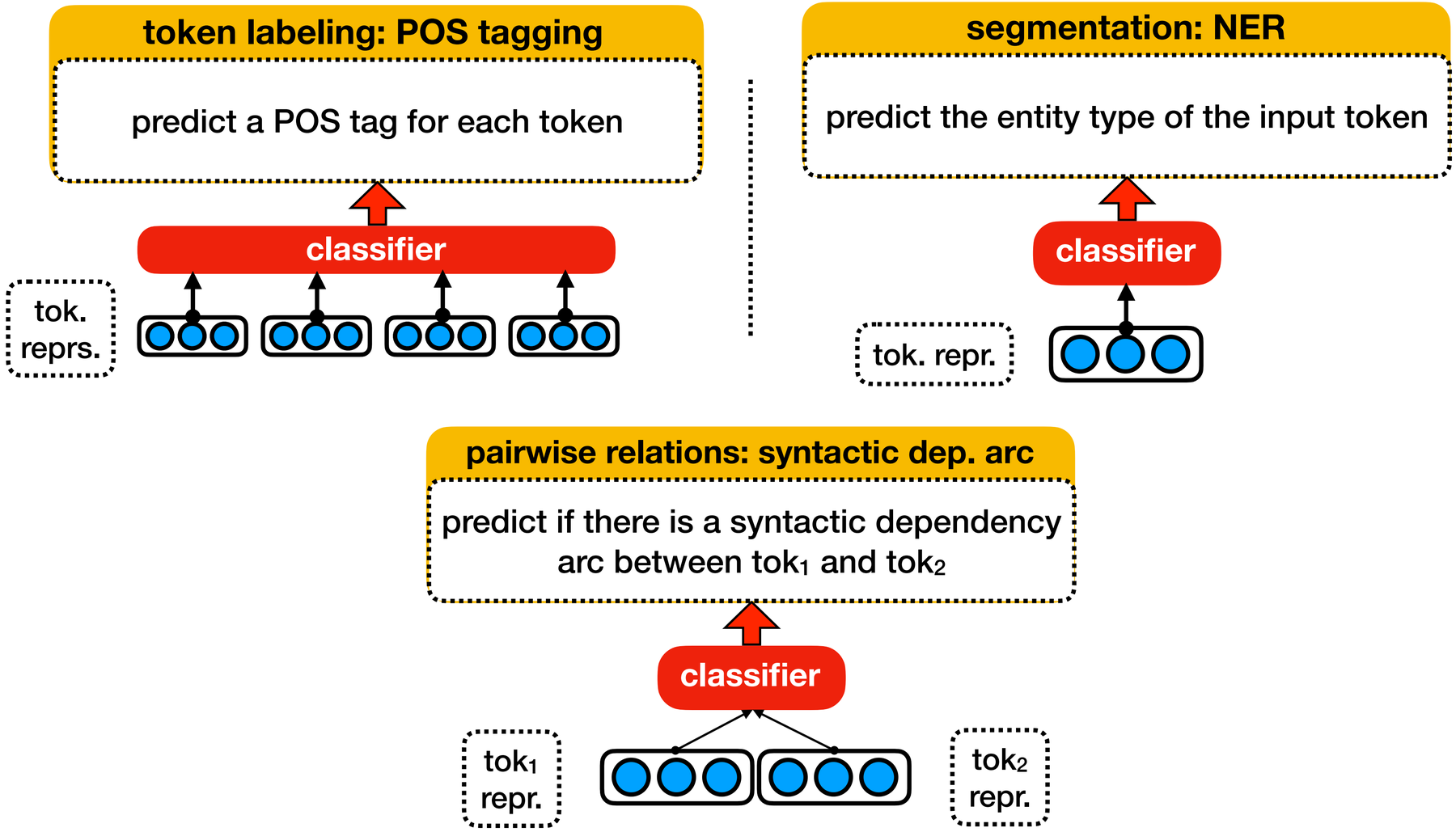
What probes can we train?



(Visuals from Mohit Iyyer)

(Adi et al., 2017)

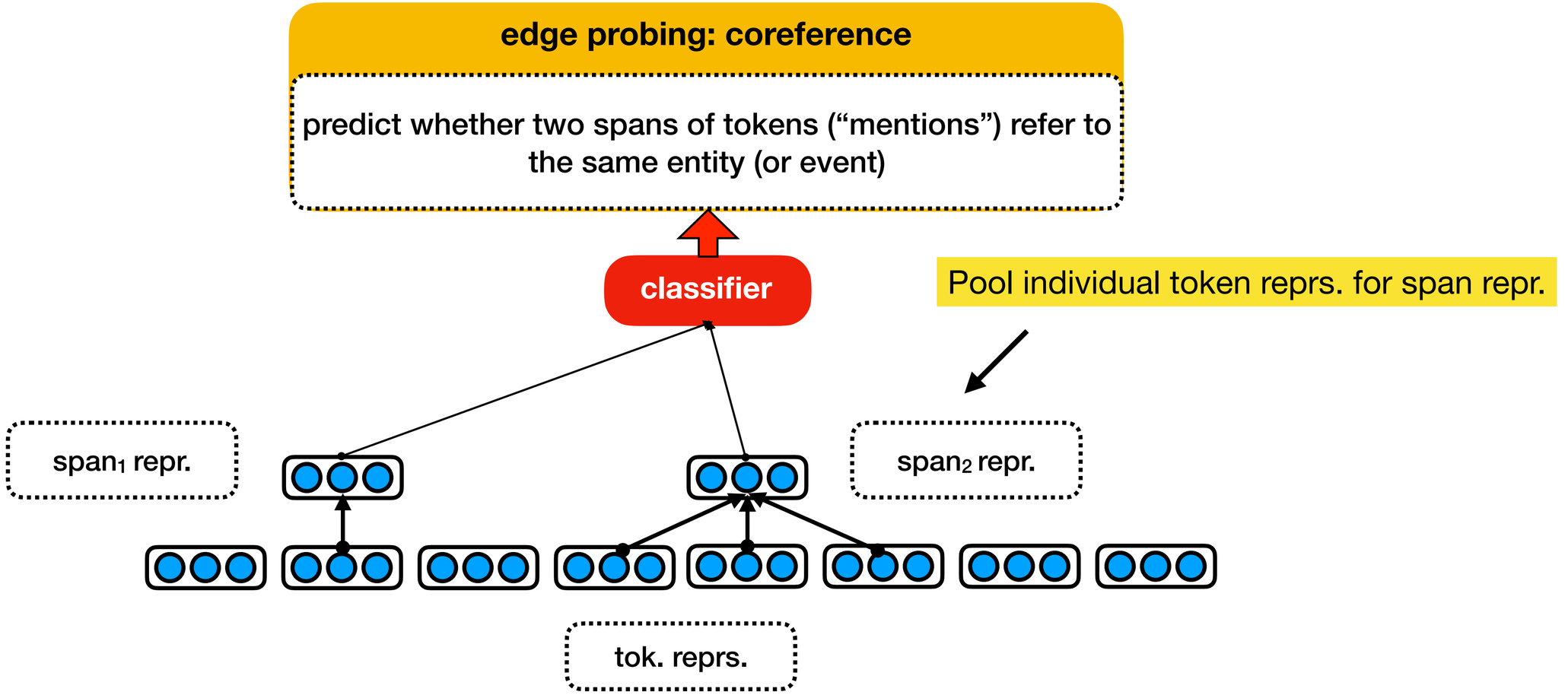
What probes can we train?



(Visuals from Mohit Iyyer)

(Liu et al., 2017)

What probes can we train?



(Visuals from Mohit Iyyer)

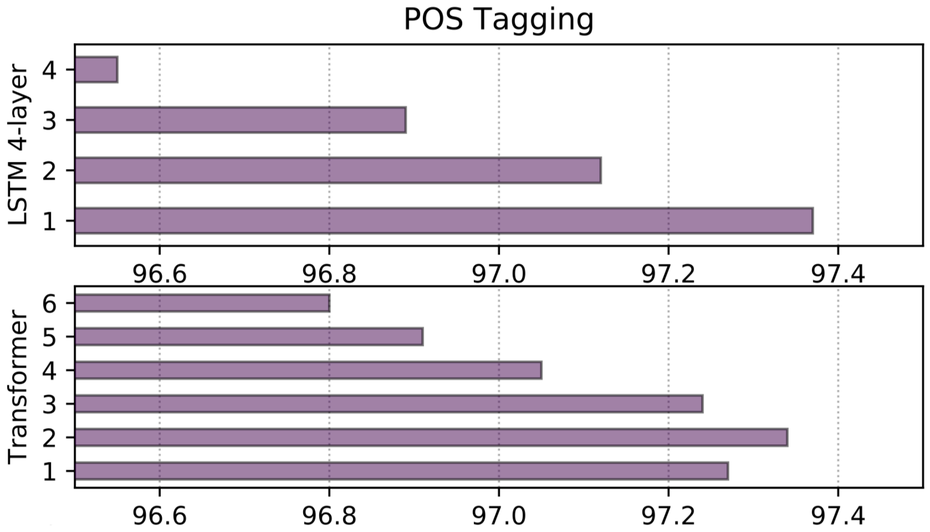
(Tenney et al., 2017)

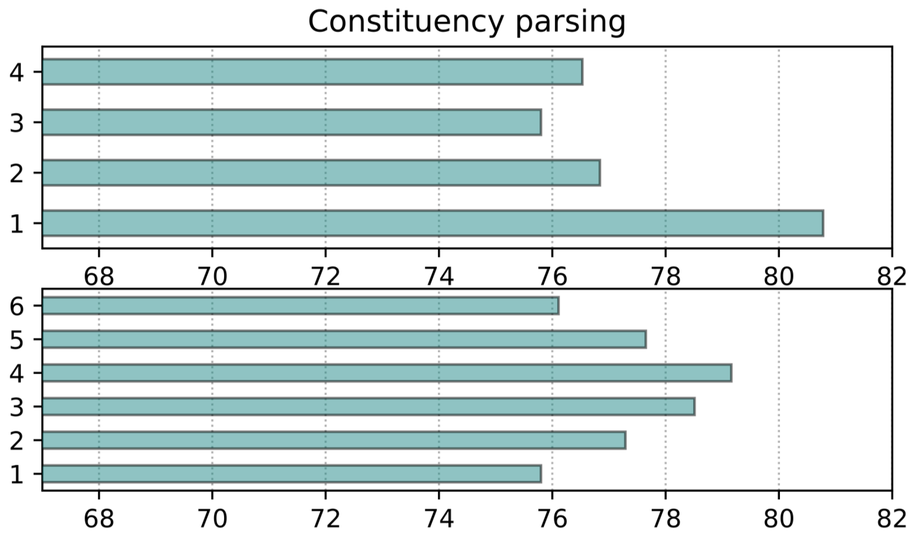
Question

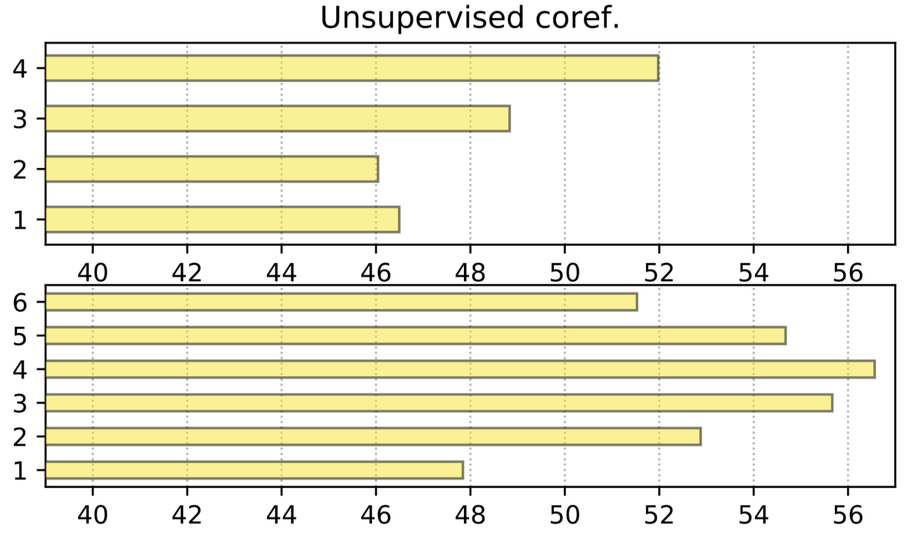
Why is any of this relevant to interpretability?

Because we can explore *how* the model encodes this linguistic information!

Which layers encode which linguistic properties?







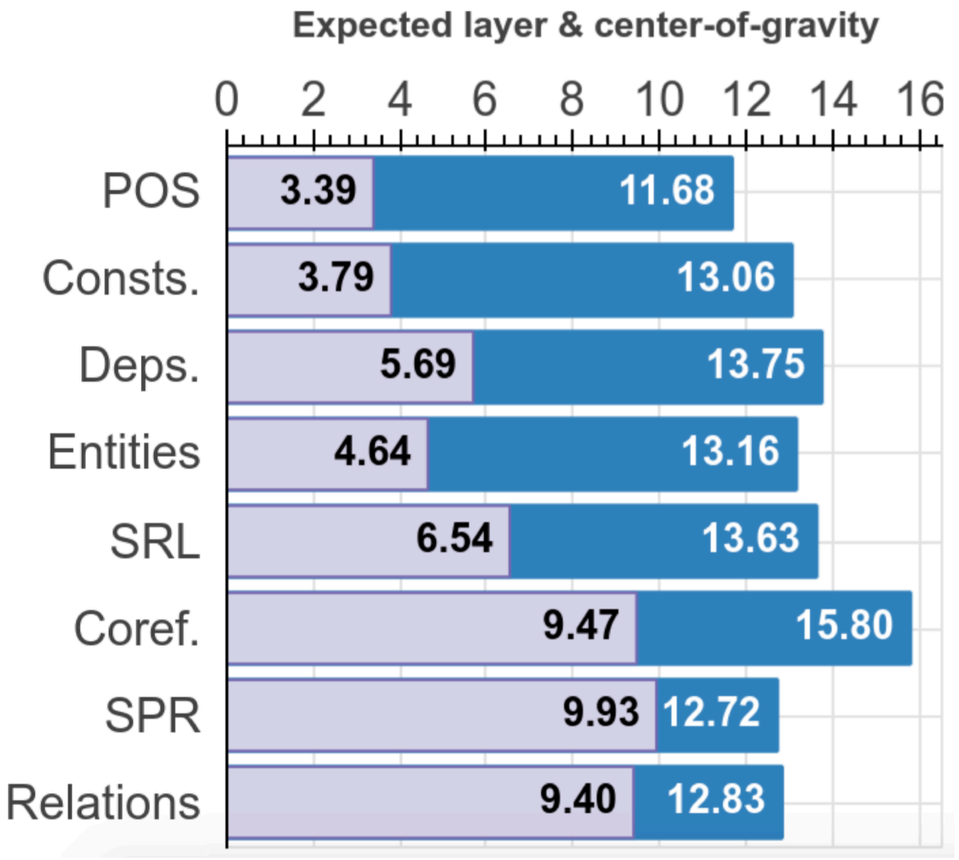
lowest layers focus on local syntax

upper layers focus more semantic content

(Peters et al., 2018)

Recovery of traditional NLP pipeline

- 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)

Question

What might be some shortcomings of linguistic probes?

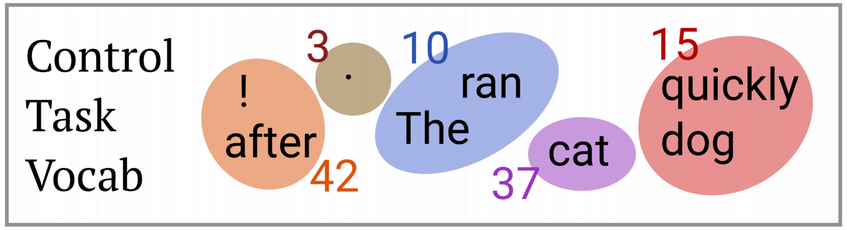
Probe Complexity

- How accessible should linguistic information in LM representations be?
  - **Simplicity**: use a linear classifier
  - **Complex**: as long as *any* model, no matter how powerful, can make the prediction correctly, the linguistic information is represented (Pimentel et al., 2020)

Selectivity

- How accessible should linguistic information in LM representations be?
  - **Simplicity**: use a linear classifier
  - **Complex**: as long as *any* model, no matter how powerful, can make the prediction correctly, the linguistic information is represented (Pimentel et al., 2020)

- **Selectivity**: how well can the probe perform on randomly labeled linguistic examples?
  - Identify a control label for each word in the vocabulary
  - Train a probe to predict these control labels

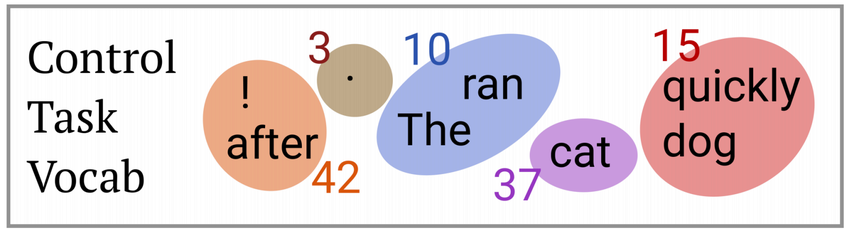


Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	10	37	10	15	3
Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	10	15	10	42	42

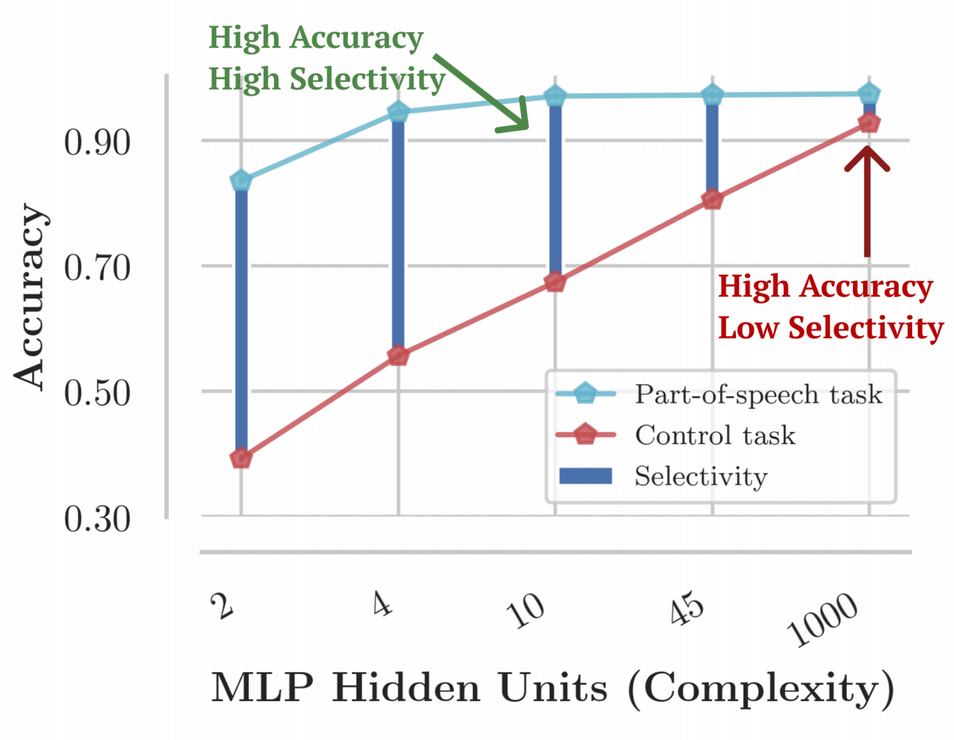
(Hewitt et al., 2019)

Selectivity

- **Selectivity**: how well can the probe perform on randomly labeled linguistic examples?
  - Identify a control label for each word in the vocabulary
  - Train a probe to predict these control labels



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

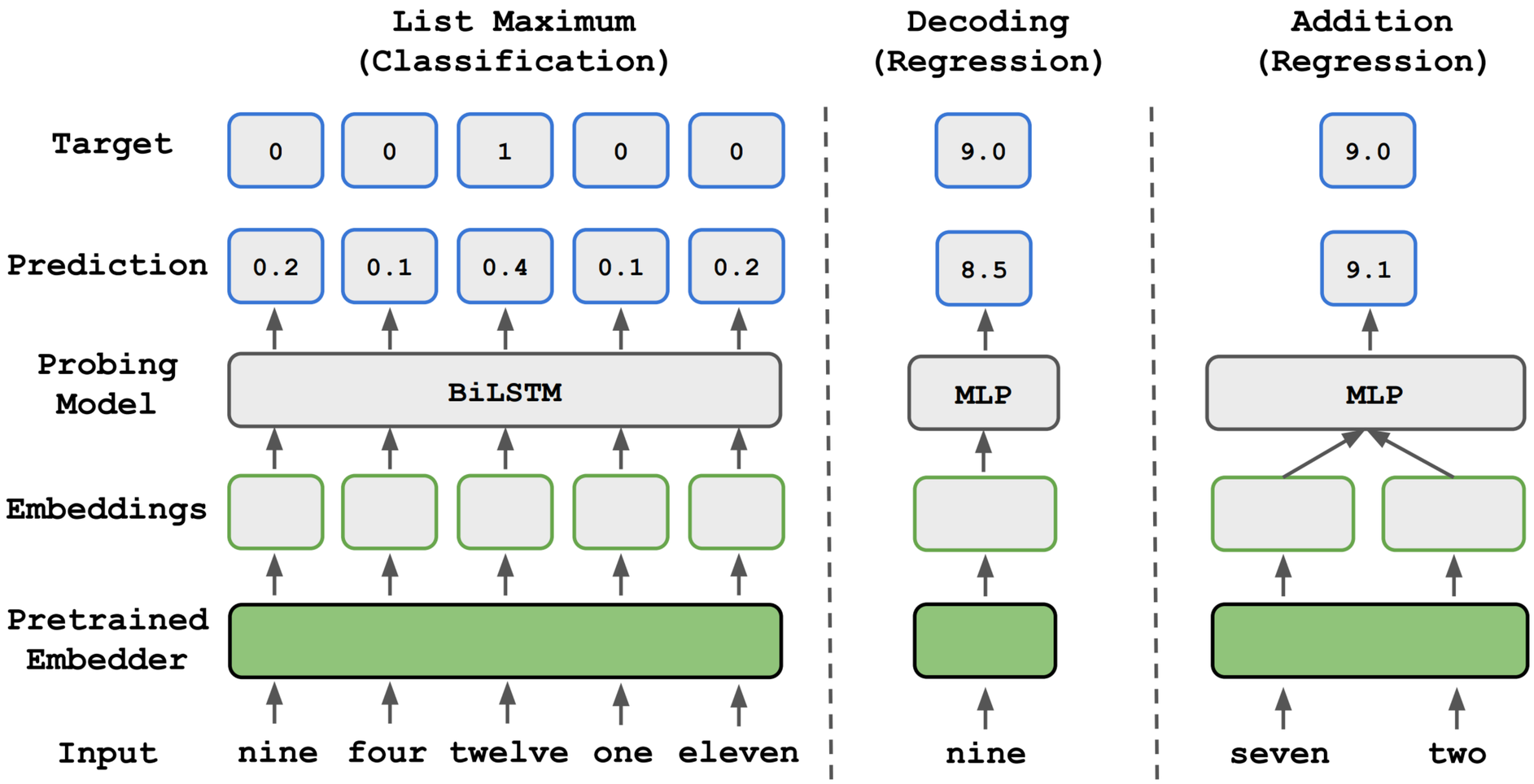


(Hewitt et al., 2019)

Question

Can we probe for something other than linguistic properties?

Numeracy



(Wallace et al., 2019)

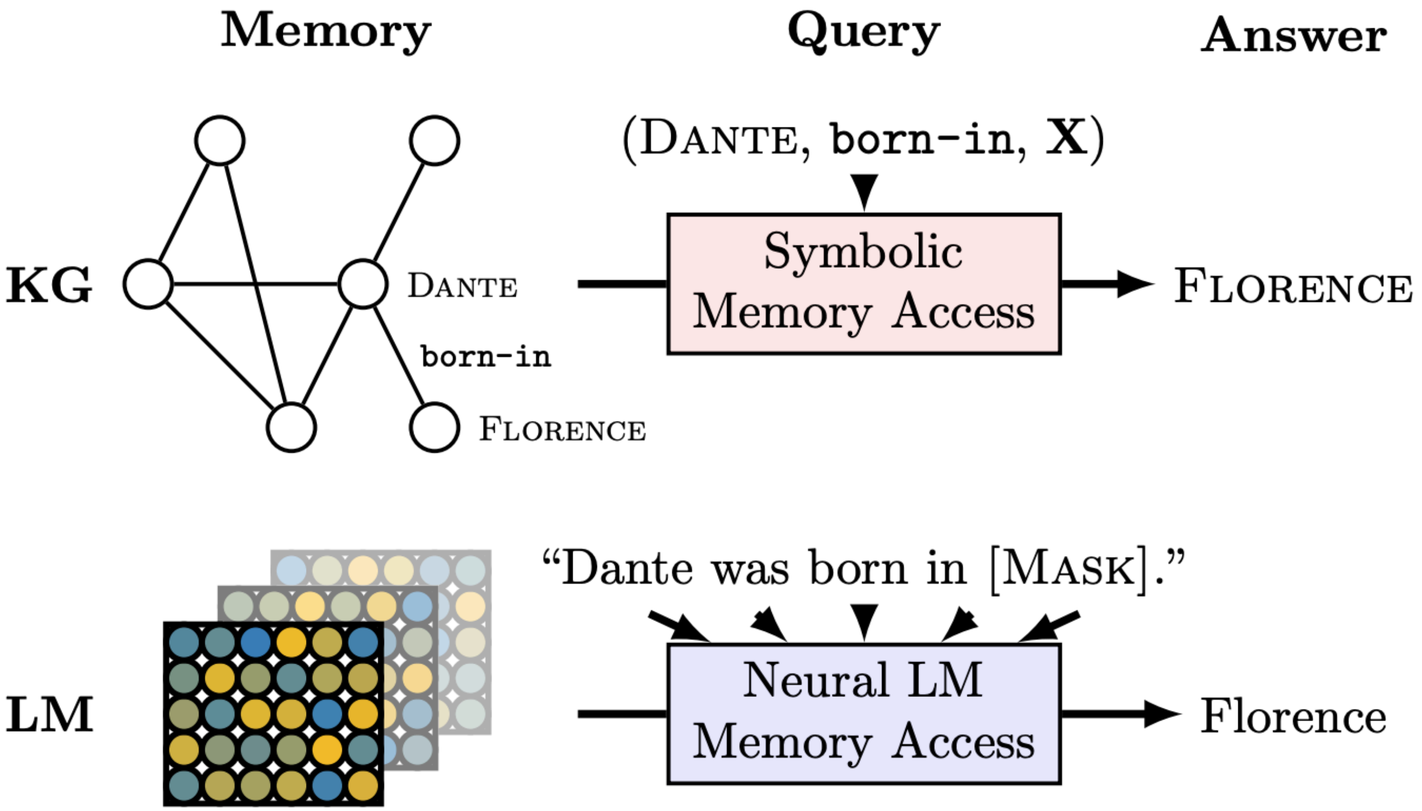
Numeracy

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

- Decent performance if you only test within the range of numbers for which you trained the probe
  - **no extrapolation beyond this range**
- **Why?** subword tokenisation is a bad method to encode digits
  - two numbers which are similar in value can have very different sub-word divisions

(Wallace et al., 2019)

Knowledge



(Petroni et al., 2019)

Knowledge

- Define templates for relations in a knowledge base, e.g., "[S] was born in [O]" for `place_of_birth`
- Mask the object in the template and use the language model to directly predict the object

	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]

(Petroni et al., 2019)

Knowledge

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)

Recap

- Probes are classification models
  - Freeze language model and uses its output (or layer-wise representations) as input features to train a classifier
  - **Probing task doesn't care the encoder architecture, only output representation vectors**
  - **Probes can be designed to test for syntax, semantics, numeracy, knowledge, commonsense , etc.**
- Need to be careful about how much capacity your probe is providing — will this information be accessible?
- Because of their simplicity, it is easier to control for biases in probing tasks than in downstream tasks — probing tasks become proxies for model reliability
  - **Correlation ≠ causality — good probe performance doesn't mean you'll succeed on downstream task!**

Question

How else could we interpret why a model makes a prediction?

Local explanations of the input

Local Explanations

- Treat model as black box
- Interpret model *behavior*, rather than representations
- **Simple:** Given an input x, what do I observe the output to be?
- **Counterfactual:** if the input were x’ instead of x, what would the output be?

Local Explanations

Sentence	**Prediction**
that movie was not great , in fact it was terrible !	-
that movie was not ____ , in fact it was terrible !	-
that movie was ____ great , in fact it was ____ !	+

Leave-one-out Method

- Delete words one by one and see how prediction probability changes

that movie was not great , in fact it was terrible ! – prob = 0.97

___ movie was not great , in fact it was terrible ! – prob = 0.97

that ____ was not great , in fact it was terrible ! – prob = 0.97

that movie ____not great, in fact it was terrible ! – prob = 0.97

that movie was ___ great, in fact it was terrible ! – prob = 0.8

that movie was not ____, in fact it was terrible ! – prob = 0.99

Leave-one-out Method

- Delete words one by one and see how prediction probability changes
- **Result**: highlights of the input based on effect of each word on prediction

that movie was not great , in fact it was terrible !

  - **not** contributed to predicting the negative class (removing it made it less negative),
  - **great** contributed to predicting the positive class (removing it made it more negative)

- **Problem:** If multiple features contribute to a prediction, removing them individually may hide the effect

More advanced: LIME

- Locally-interpretable, model-agnostic explanations
- Similar to leave-one-out, but delete groups of features at same time

Ribeiro et al. (2016)

Question

Doesn’t a lot of this interpretability work depend on human interpretation of the “interpretability” result?

Faithfulness vs. Plausibility

- Suppose our model is a bag-of-words model with the following:
  - the = -1, movie = -1, good = +3, bad =0
  - the movie was good prediction score=+1
  - the movie was bad prediction score=-2
- Suppose explanation returned by local explanation method is:
  - the movie was good
  - the movie was bad
- Is this a “correct" explanation?

Faithfulness vs. Plausibility

- *Plausible* explanation: intuitive to a human

the movie was good the movie was bad

  - Intuitive to a human, but not what the model is really doing (according to our model) !

- *Faithful* explanation: reflects the behavior of the model

the movie was good the movie was bad

  - We should prefer faithful explanations;
  - non-faithful explanations are actually deceiving us about what our models are doing!

Question

Could machines interpret their own operation?

Model Explanations

Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Expl a large wingspan, hooked yellow beal

Laysan Albatross

Description the water.
Class Defir ed yellow beak, black back and white b
Visual Expl a hooked yellow beak white neck and b

Hot off the presses:

GPT-4 can provide explanations of its own neurons?

- Are the explanati haking its decision?
- Model-generated explanations may not match probing of model behavior
- Explanations may be plausible, but unfaithful!

Hendricks et al. (2016)

Recap

- Model interpretability methods allow us to explore *how* our models function when they make predictions
- **Many different approaches**: probing, local explanation, model-based explanation, and many not covered here today!
- Interpretability methods can be misleading in what they show — always be aware of the assumptions baked into your interpretation
- Plausibility is not the same as faithfulness — is this actually how your model is making a prediction? Or do you just want to believe it is?