LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Linguistic capabilities of large language models (LLMs)

Renhao Pei

26/05/2023

Outline

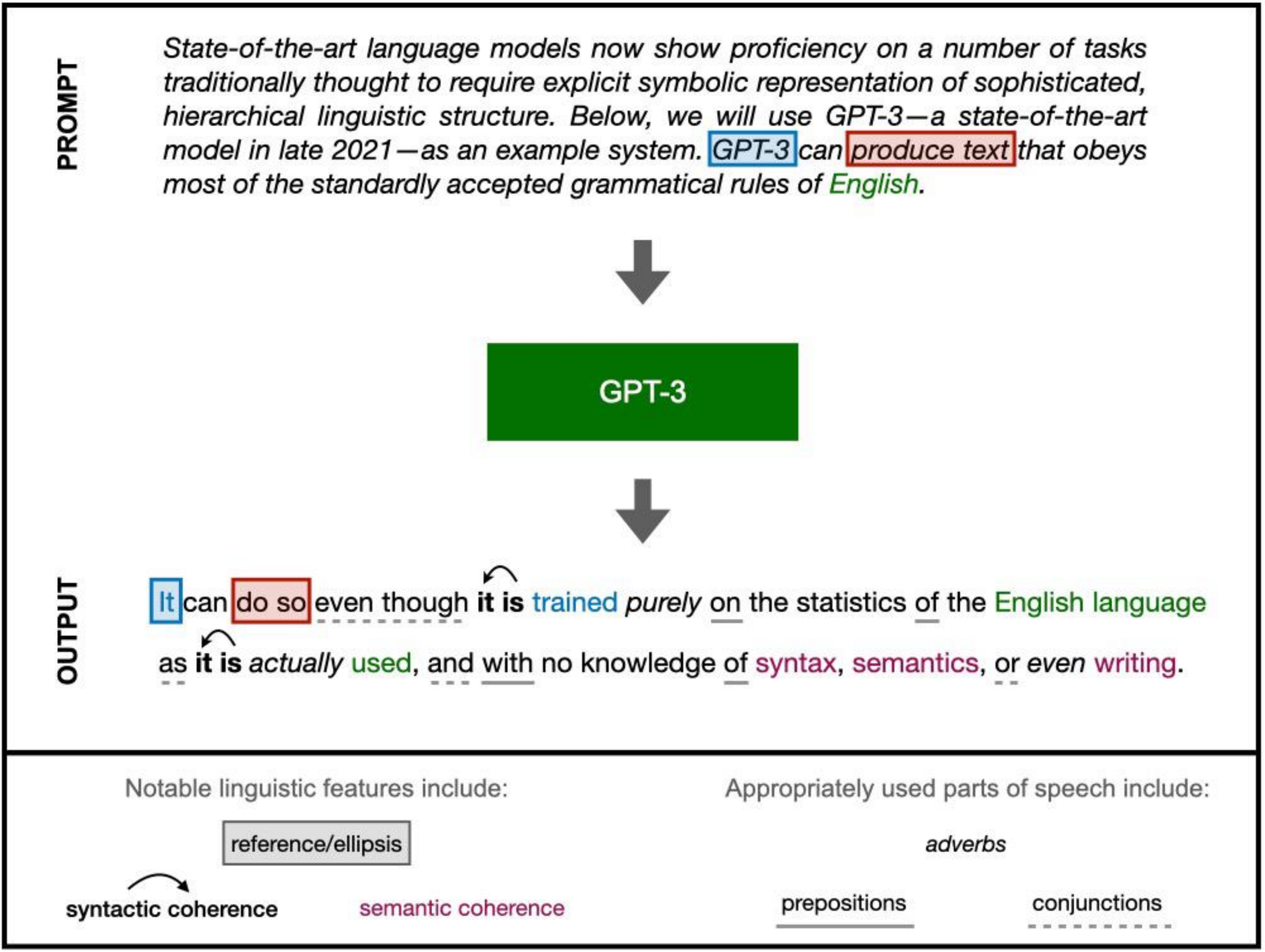
- Introduction
  - formal linguistic competence vs. functional linguistic competence
- (Formal) linguistic capabilities of LLMs
  - hierarchical structure & Long-distance agreement
  - BERT Rediscovers the Classical NLP Pipeline & Probing
- Probing Pretrained Language Models for the English Comparative Correlative
  - construction grammar
  - syntactic probe
  - semantic probe (where the LLMs fail)
  - discussion

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction

- LLMs can do things that a human can only do if they have some linguistic knowledge:

- Grammaticality
- Semantical plausibility
- Coreference
- Connectives, discourse coherence
- …



State-of-the-art language models now show proficiency on a number of tasks traditionally thought to require explicit symbolic representation of sophisticated, hierarchical linguistic structure. Below, we will use GPT-3—a state-of-the-art model in late 2021—as an example system. GPT-3 can produce text that obeys most of the standardly accepted grammatical rules of English.

It can do so even though **it is** trained *purely* on the statistics of the English language as **it is** *actually* used, and with no knowledge of syntax, semantics, or *even* writing.

Introduction

There are things that LLMs cannot do well:

Formal reasoning (e.g. math problem expressed in natural language)

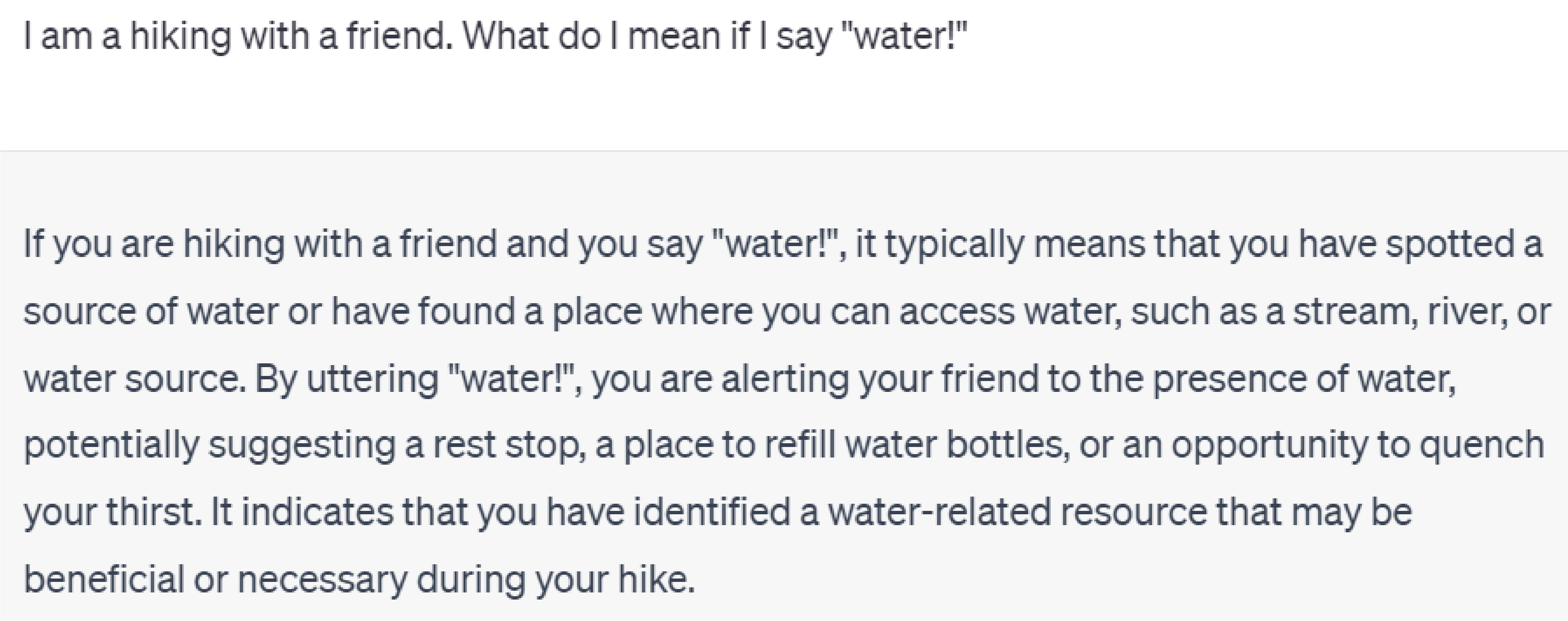
- LLMs rely on heuristics and fail on more complicated math problems.

Social reasoning (pragmatics and intent)

- Recent versions of GPT-3 models show a markedly improved capacity to interpret non-literal utterances, such as metaphors and polite deceit. However, they are still struggling on inferring the intentions behind others' actions (Sap et al., 2022).

Introduction

- P.s. ChatGPT can already do this much better!



I am a hiking with a friend. What do I mean if I say "water!"

If you are hiking with a friend and you say "water!", it typically means that you have spotted a source of water or have found a place where you can access water, such as a stream, river, or water source. By uttering "water!", you are alerting your friend to the presence of water, potentially suggesting a rest stop, a place to refill water bottles, or an opportunity to quench your thirst. It indicates that you have identified a water-related resource that may be beneficial or necessary during your hike.

Introduction

- But some limitations are not really due to deficient linguistic capabilities

- Drawing on evidence from cognitive neuroscience, **formal linguistic competence** in humans relies on specialized language processing network in the brain, **functional linguistic competence** recruits multiple extralinguistic capacities, such as formal reasoning, world knowledge, situation modeling, and social cognition (Mahowald et al., 2023)

- Individuals with global **aphasia** exhibit severe linguistic deficits, spare nothing but single word comprehension for a small set of words.
- However, they have intact non-linguistic cognitive abilities: they can play chess, compose music, solve arithmetic problems and logic puzzles, and navigate complex social situations (Mahowald et al., 2023)

Introduction

- Many capacities required for real-life language users are, in fact, not specific to language and are supported by distinct brain circuits.

- In line with this distinction, it might be reasonable that LLMs that master many syntactic and distributional properties of human language still cannot use language in human-like ways.

- Solution proposed in Mahowald et al. (2023): Modularity

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction

- What are the current linguistic capabilities of LLMs? How do we analyze them?

- What are the limitations of LLMs, in terms of linguistic capabilities?

Formal linguistic capabilities of LLMs: Hierarchical structure & Long-distance agreement

- A crucial formal feature of human language: hierarchical structure
- The meaning of a sentence is not derived by combining the meaning of each word one by one linearly. Instead, they are combined **hierarchically**.

- Long-distance agreement: *The keys to the old, wooden kitchen cabinet* ***are*** *on the table.*

- Can LLMs learn this hierarchical structure?
- *The length of the forewings (is/*are). . .*

- Gulordava et al. (2018) showed that LSTMs trained only to predict the next word in a corpus, can predict the long-distance agreement with high accuracy, where a baseline model like 5-gram cannot.
- It performed well even on semantically nonsensical sentences: *The colorless green ideas I ate with the chair (sleep/*sleeps)*

Probing
BERT Rediscovers the Classical NLP Pipeline

- Probing: take the internal representations of a language model (vectors) as input, train a new separate classifier to learn to recover a particular linguistic distinction from the vector representations.

- Tenney et al. (2019) uses a suite of probing tasks, derived from the traditional NLP pipelines (POS-tagging, constituency, dependency, semantic role labeling, semantic proto-role etc.) to quantify where specific types of linguistic information are encoded **on each layer**.

- A series of classifiers $\{P_\tau^{(\ell)}\}_\ell$ are trained to attend to layer l as well as all previous layers.

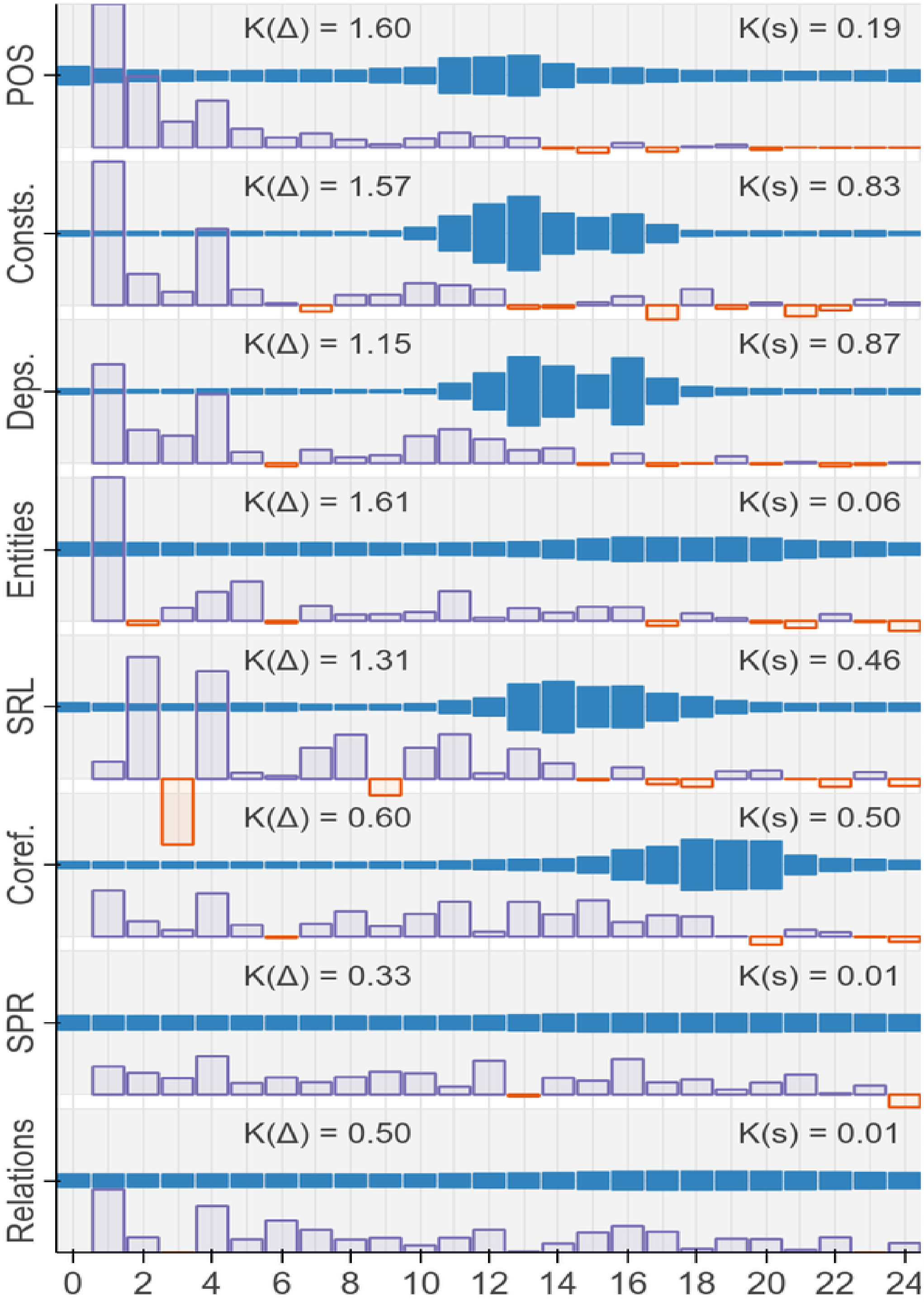
- These classifiers are cumulative, in the sense that $P_\tau^{(\ell+1)}$ has a similar number of parameters but with access to more information than $P_\tau^{(\ell)}$

BERT Rediscovers the Classical NLP Pipeline: Differential score

- Performance score (F1 score) intuitively generally increases as more layers are added.
- Differential score $\Delta_\tau^{(\ell)}$:

$$\Delta_\tau^{(\ell)} = \text{Score}(P_\tau^{(\ell)}) - \text{Score}(P_\tau^{(\ell-1)})$$

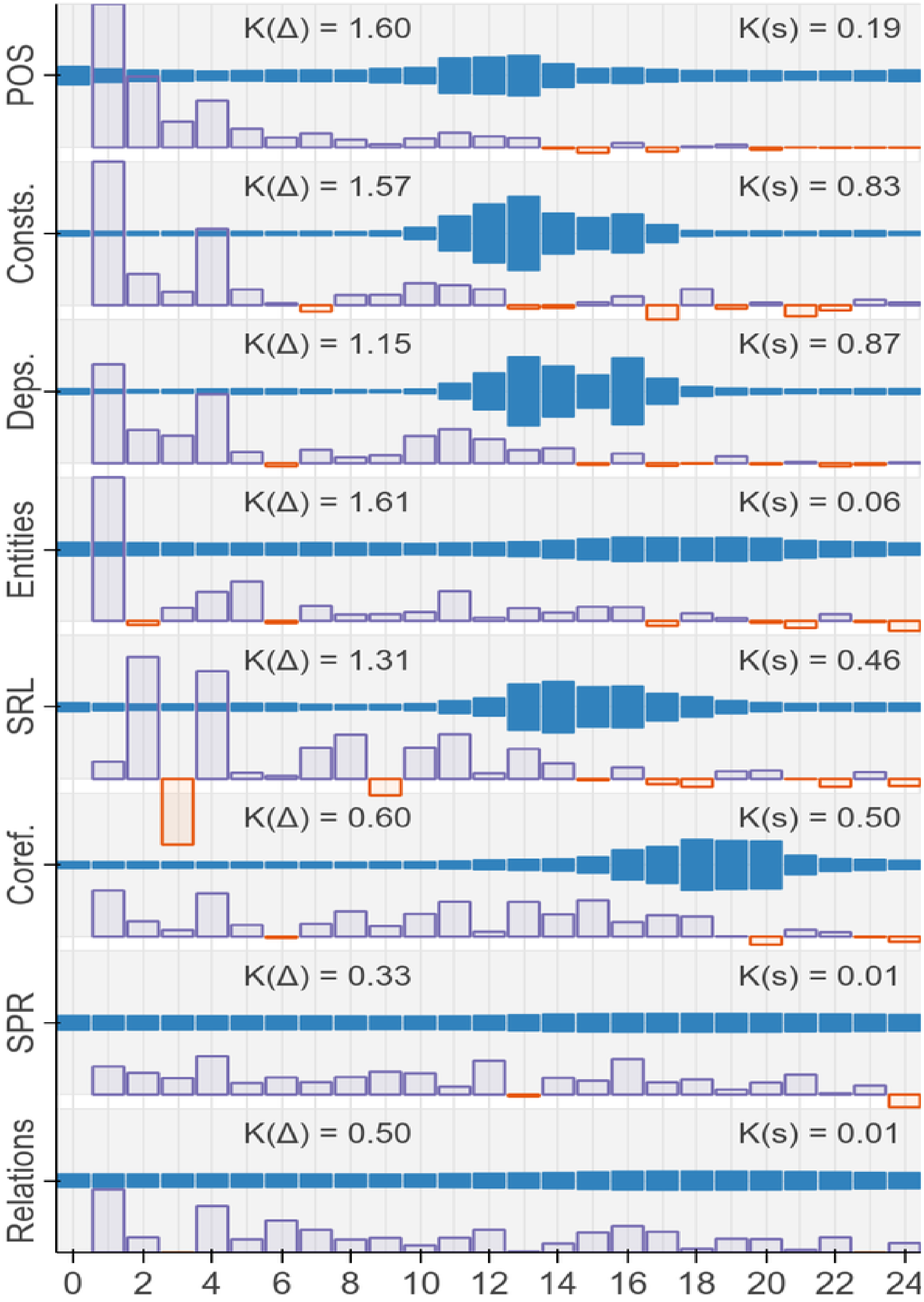
- It measures how much better we do on the probing task if we observe one additional encoder layer.
- histograms (purple) are differential scores, normalized for each task
- horizontal axis is encoder layer



BERT Rediscovers the Classical NLP Pipeline: Differential score

- Some tasks (e.g. POS tagging) can be correctly classified very early on, while challenging tasks (e.g. SPR task) have continued improvement up to the highest layers.

- Syntactic information (constituents, dependencies) is concentrated on a few layers, while information related to semantic tasks (SPR, relation classification) is spread across the entire network.



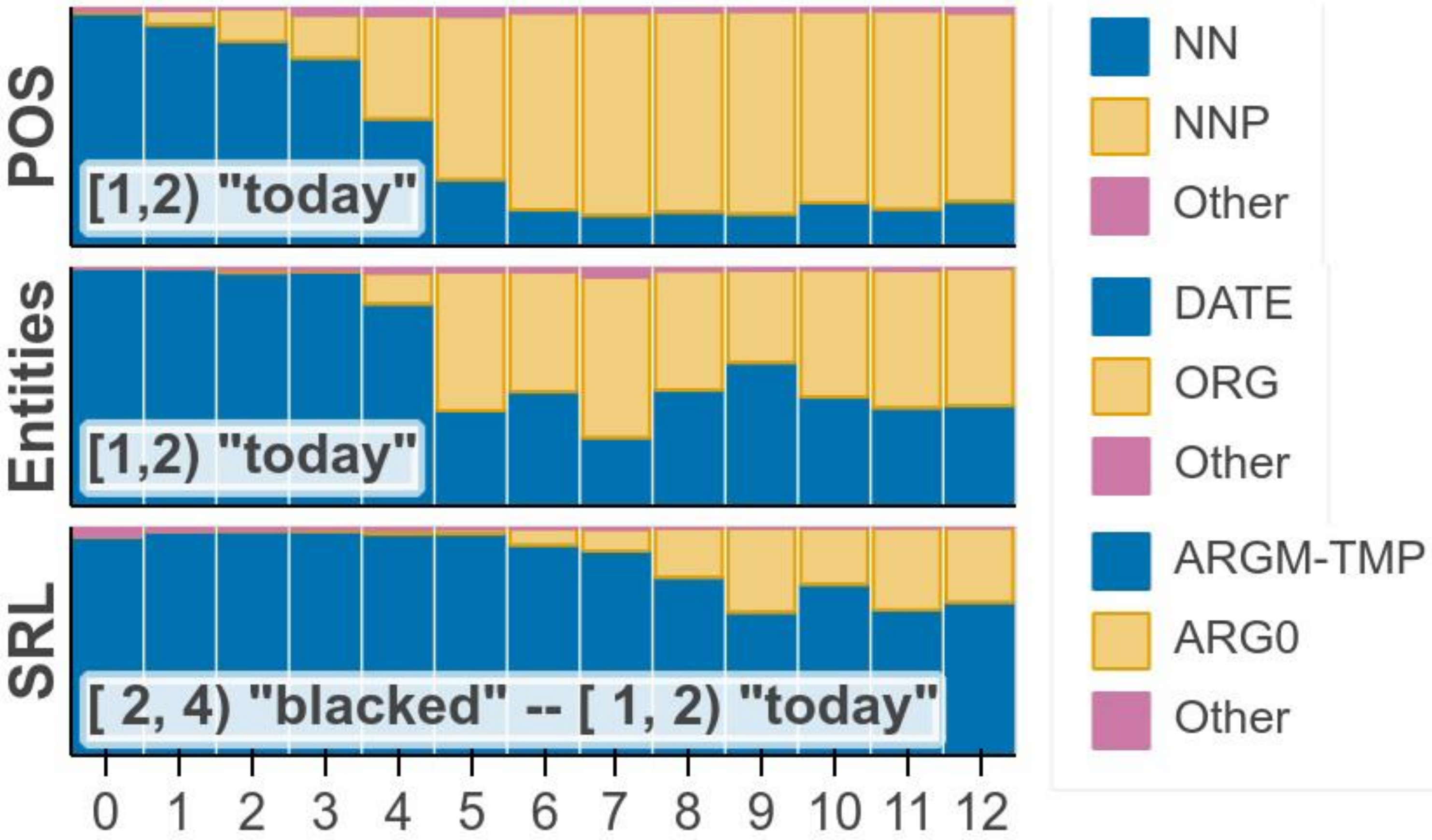
BERT Rediscovers the Classical NLP Pipeline: Classifier predictions across layers

- *“today” is initially tagged as a common noun, date, and temporal modifier (NN,DATE,ARGM-TMP)*

- *but later “china today” is reinterpreted as a proper noun (NNP), i.e. the TV network*

- *the model then updates the entity type (to ORG), and the semantic role (as the agent ARG0)*

*china **today** blacked out a cnn interview that was ...*



BERT Rediscovers the Classical NLP Pipeline: Conclusion

Deep language models can model two things traditionally believed necessary for language processing:

- Lower layers encode more basic syntactic information while higher layers capture more complex semantics
- Different levels of hierarchical information (high-level semantic vs. low-level syntactic information) interact with each other

The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative

- Construction Grammar (CxG): Rather than a system divided into non-overlapping syntactic rules and lexical items, CxG views language as a structured system of **constructions,** that encapsulate syntactic and semantic components as single linguistic signs.

- Construction: linguistic units (form-meaning pairs) of different **granularity** that combine syntax and semantics, ranging from individual morphemes up to phrasal elements and fixed expressions

Construction: The English Comparative Correlative

1. The more, the merrier.
2. The longer the bake, the browner the colour.
3. The more she practiced, the better she became.
4. The Better Your Syntax, the Better Your Semantics?

Form

- The Comparative Correlative (CC) consists of two clauses, both of which are characterized by an instance of “the” followed by an adverb or an adjective in the comparative (‘-er’ or with “more”).

Meaning

- A general cause-and-effect relationship: as (2) can be paraphrased as “If the bake is longer, the colour will be more brown”.
- A temporal development: paraphrasing (3) as “She practiced more over time, and she became better over time”.

The Better Your Syntax, the Better Your Semantics?
Syntax Probing

First of all, can the model **recognise the structure** of CC?

- Syntactic probing question: Can the PLM distinguish instances of the CC from **similar-looking** non-instances?

Minimal pairs:

- pairs of sentences which are indistinguishable except for the fact that one of them is an instance of the CC and the other is not.

Two ways of constructing minimal pairs:

- (1) Artificial data based on a context-free grammar (CFG)
- (2) Sentences extracted from C4 Corpus

Syntax Probing: Synthetic Data

Pattern for a positive instance:

- The ADV-er the NUM NOUN VERB
- “The harder the two cats fight”.

To create a negative instance, reorder the pattern to :

- The ADJ-er NUM VERB the NOUN
- “The harder two fight the cats”.

Negative instance uses the **same words** as the positive one, but in a **different order**.

Syntax Probing: Corpus-based data

Procedures

- (1) Extract sentences from C4 corpus, that follow the pattern: “The” (DET) followed an adjective or adverb in the comparative, and at any point later in the sentence again the same pattern.
- (2) Group these sentences by their sequence of POS tags.
- (3) Manually classify the sequences as either positive or negative.

Example

- She thinks the more water she drinks the better her skin looks. [Positive]
- The way the older guys help out the younger guys is fantastic. [Negative]

Syntax Probing: the Probe

- Train a simple logistic regression model on top of the mean-pooled sentence embeddings.

Sentences are sampled such that both the positive and the negative class is **balanced** across every value of these features:

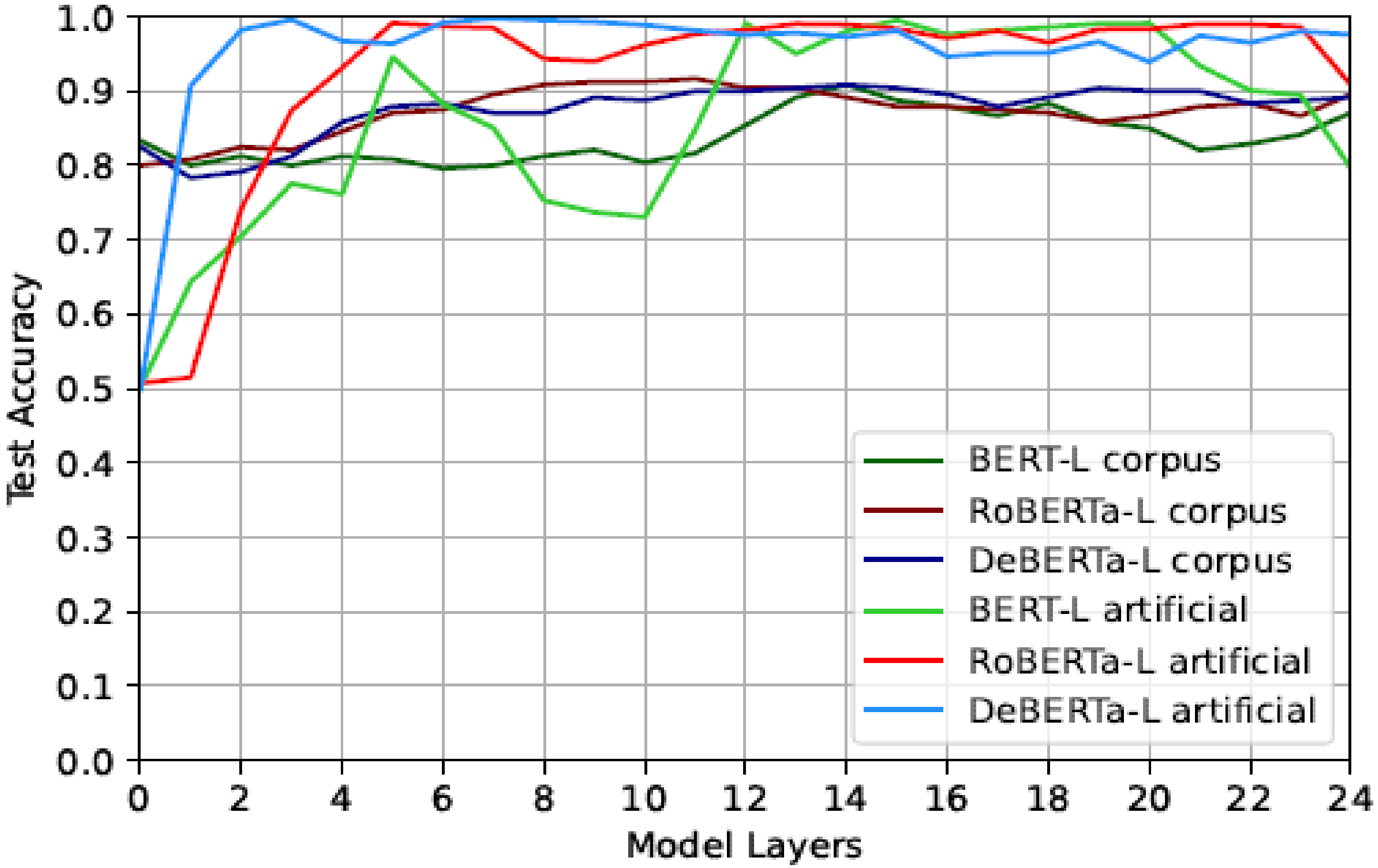
- length of the sentence,
- the start position of the construction,
- the position of its second half, and the distance between them.

This ensures that the probes are unable to exploit correlations between a class and any of the above features.

Syntax Probing: Results

- Generally, models perform better on artificial data than on corpus data from the 5th layer on (with the exception of a dip in performance for BERT large)

- All models perform at 80% or better from the middle layers on.



Semantics Probing

- Can the model, based on the meaning conveyed by the CC, draw a correct inference in a specific scenario?

Base:

- The ADJ1-er you are, the ADJ2-er you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

Example

- “The stronger you are, the **faster** you are. The weaker you are, the slower you are. Terry is stronger than John. Therefore, Terry will be [MASK] than John. ”
- Ask the model to predict the probabilities of the mask being “faster” and “slower”. If the model has understood the meaning conveyed by the CC, we expect the probability of **ADJ2 (**“faster”**)** to be high.

Semantics Probing

- The set-up is comparable in difficulty to the NLU tasks presented in LAMBADA (Paperno et al., 2016), on which GPT-2 has achieved high zero-shot accuracy.

Several biases might cloud the assessment of the model’s understanding of the CC:

- Recency bias
- Vocabulary bias
- Name bias

Semantics Probing: Recency bias

Base(S1):

- The ADJ1-er you are, the **ADJ2-er** you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

That models might prefer to repeat the adjective that is closest to the mask token.

Recency bias(S2):

- To test its influence, we flip the first two sentences so that the correct answer is now more recent.
- The ANT1-er you are, the ANT2-er you are. The ADJ1-er you are, the **ADJ2**-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

Semantics Probing: Vocabulary bias

Base(S1):

- The ADJ1-er you are, the **ADJ2-er** you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

The model might assign higher probabilities to some adjectives, purely based on their frequency in the pretraining corpus/their lexical identities.

Vocabulary bias(S3):

- To test its influence, ADJ2/ANT2 are swapped. **ANT2** is now the correct answer for mask.
- The ADJ1-er you are, the ANT2-er you are. The ANT1-er you are, the ADJ2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.
- "The stronger you are, the **slower** you are. The weaker you are, the faster you are. Terry is stronger than John. Therefore, Terry will be [MASK] than John. "

Semantics Probing: Name bias

Base(S1):

- The ADJ1-er you are, the **ADJ2-er** you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

A model might have learned to associate adjectives with names in pretraining, so we construct a third version, in which we swap the names.

Name bias(S3):

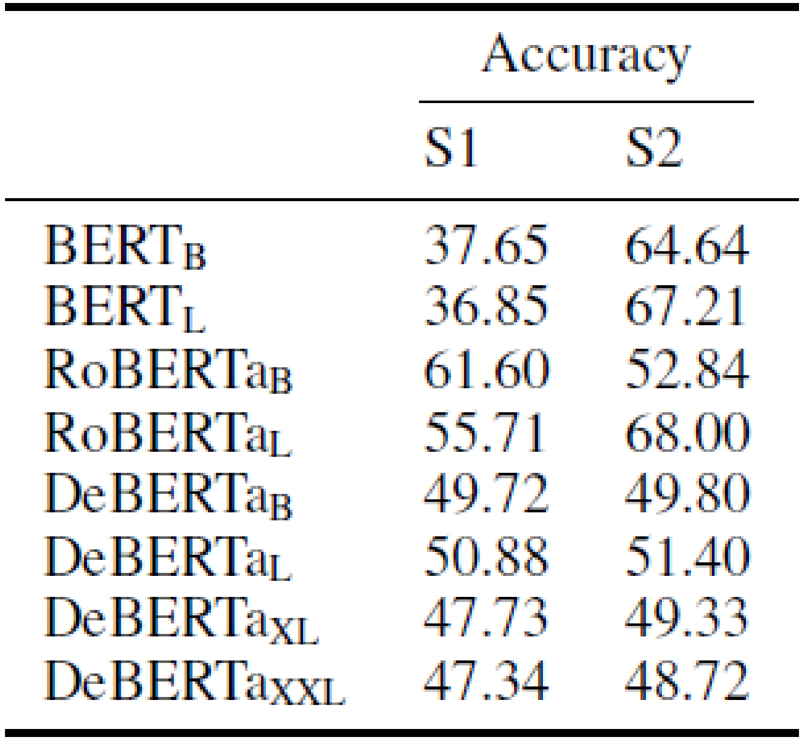
- The ADJ1-er you are, the **ADJ2-er** you are. The ANT1-er you are, the ANT2-er you are. NAME2 is ADJ1-er than NAME1. Therefore, NAME2 is [MASK] than NAME1.

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Semantics Probing: Results for biases

Accuracy:

- RoBERTa’s and DeBERTa’s scores are close to 50% (i.e., chance) accuracy for both S1 and S2.
- BERT is more influenced by the bias related to the order of the two CCs. The average between them is also very close to chance.

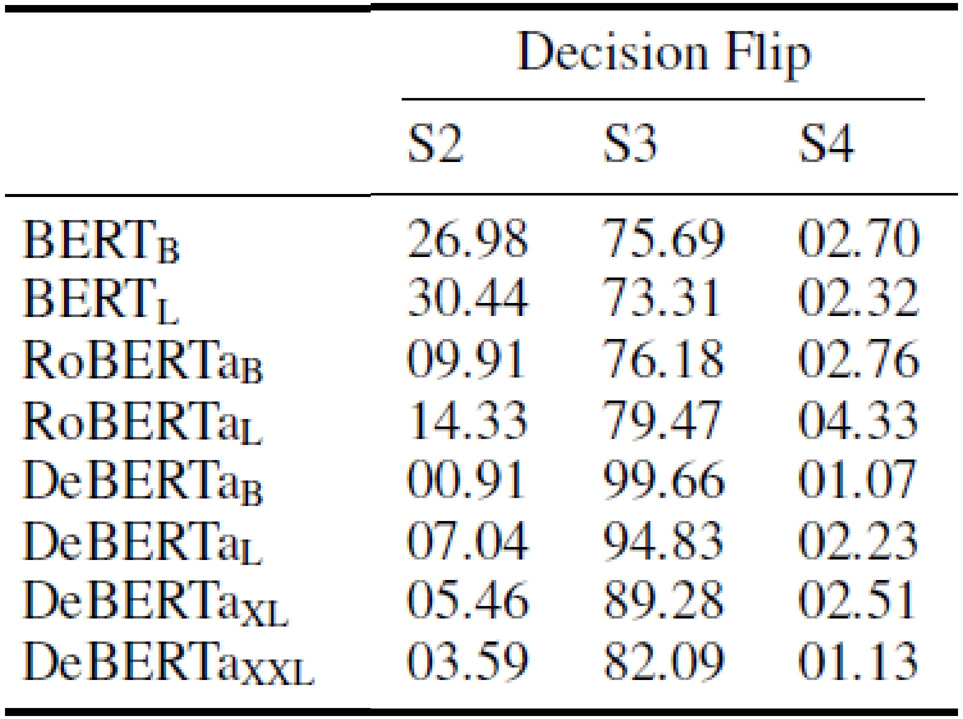


	Accuracy	
	S1	S2
BERT$_\text{B}$	37.65	64.64
BERT$_\text{L}$	36.85	67.21
RoBERTa$_\text{B}$	61.60	52.84
RoBERTa$_\text{L}$	55.71	68.00
DeBERTa$_\text{B}$	49.72	49.80
DeBERTa$_\text{L}$	50.88	51.40
DeBERTa$_\text{XL}$	47.73	49.33
DeBERTa$_\text{XXL}$	47.34	48.72

Semantics Probing: Results for biases

Decision flip:

- the percentage of sentences for which the decision was changed when considering the alternative sentence of bias (S2,S3,S4), as opposed to original (S1).
- large bias related to both the recency of the correct answer (S2) and the choice of vocabulary (S3).
- no bias related to the choice of names (S4).
- Biases are large, no wonder we have accuracies at chance level.



	Decision Flip		
	S2	S3	S4
BERT$_B$	26.98	75.69	02.70
BERT$_L$	30.44	73.31	02.32
RoBERTa$_B$	09.91	76.18	02.76
RoBERTa$_L$	14.33	79.47	04.33
DeBERTa$_B$	00.91	99.66	01.07
DeBERTa$_L$	07.04	94.83	02.23
DeBERTa$_{XL}$	05.46	89.28	02.51
DeBERTa$_{XXL}$	03.59	82.09	01.13

Semantics Probing: Calibration

$$P_c(a|S_b) = P(a|S_b)/[\sum_{i=1}^{i=5}(P(a|C_i)/5)]$$

- Dividing the probability predicted in the task context by the prior probability of a label (i.e., its probability if no context is given)

- This gives us the conditional probability of a label given the context, representing the true knowledge of the model about this task (CC).

Semantics Probing: Calibration

- Removing the important information of CC:

Short (S5):

- NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

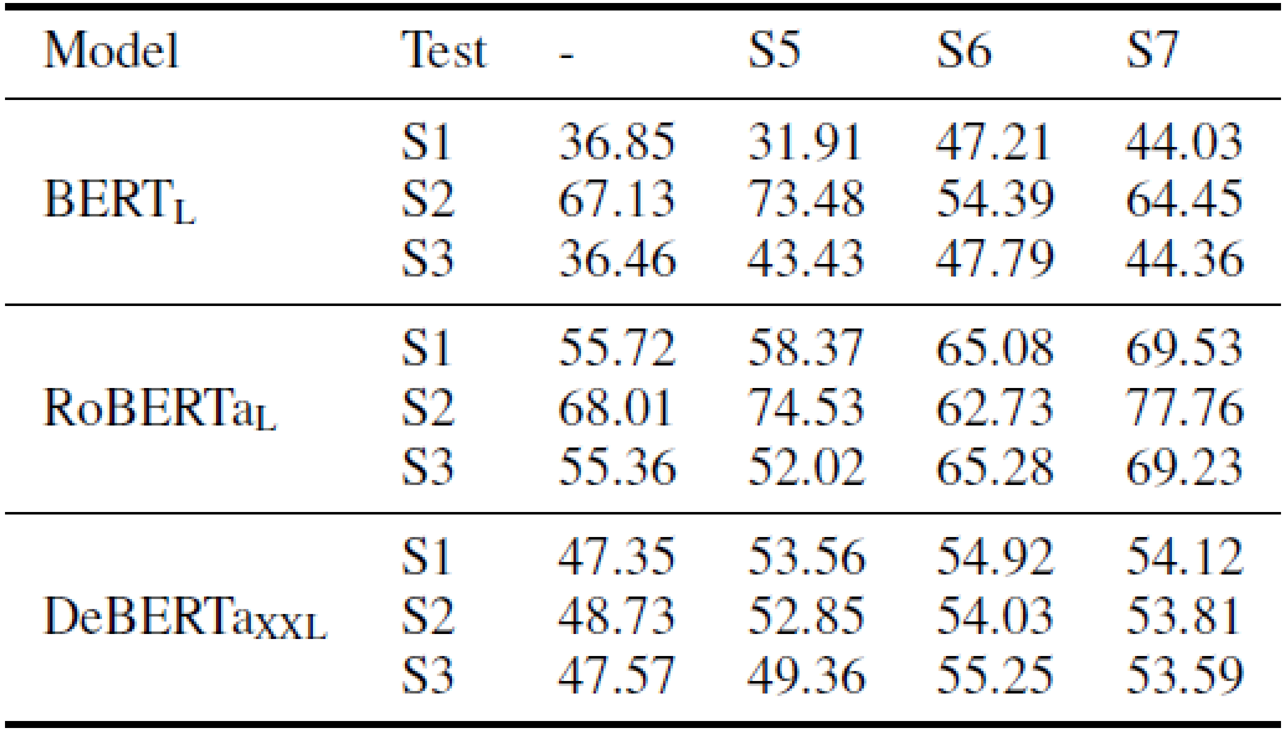
Name (S6):

- The ADJ1-er you are, the ADJ2-er you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME3 is [MASK] than NAME4.

Adjective (S7):

- The ADJ1-er you are, the ADJ2-er you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ3-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

Semantics Probing: Results for Calibration



Model	Test	-	S5	S6	S7
BERT$_\text{L}$	S1	36.85	31.91	47.21	44.03
	S2	67.13	73.48	54.39	64.45
	S3	36.46	43.43	47.79	44.36
RoBERTa$_\text{L}$	S1	55.72	58.37	65.08	69.53
	S2	68.01	74.53	62.73	77.76
	S3	55.36	52.02	65.28	69.23
DeBERTa$_\text{XXL}$	S1	47.35	53.56	54.92	54.12
	S2	48.73	52.85	54.03	53.81
	S3	47.57	49.36	55.25	53.59

- Despite the effort to retrieve any knowledge that the models have about the CC,
- they are unable to perform clearly above chance, and we have therefore found no evidence that the investigated models understand and can use the semantics of the CC.

Conclusion

- Even though PLMs are able to classify sentences even in difficult circumstances, they do not seem to be able to extract the meaning it conveys and use it in context,
- indicating that while the syntactic aspect of the CC is captured in pretraining, the semantic aspect is not.

- Why?
- (1) Models have never had a chance to learn what the CC means because they have never seen it applied, and do not have the same opportunities as humans to either interact with the speaker to clarify the meaning or to make deductions using observations in the real world.

Conclusion

- (2) It might be possible that the type of meaning representation required to solve this task is beyond the current transformer-style architectures. (Alternative: Logic for NLI ect.?)

- Similar evidence: complex semantics like negation is still beyond state-of-the-art PLMs. (Kassner and Schütze, 2020)

Thank you!