LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTER FOR INFORMATION AND LANGUAGE PROCESSING

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

Joint work with Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze

Leonie Weissweiler

06.05.2023

Construction Grammar

It matters which theory of Syntax we use in NLP

- Overgeneralisation: Universal Dependencies → Dependency Grammar → Syntax
- Assessment of progress of the field: „Have Language Models acquired Syntax?“
- Making recommendations from the Linguistics niche to the broader community:
  - ‚Are we climbing the wrong hill?‘
  - ‚Are language models learning language the right way?‘
  - ‚Are language models learning the same way that humans do?‘

How is Construction Grammar different?

- No strong line between lexicon and Syntax → Patterns (called Constructions) are stored in the brain the same way words are
- Focus on surface form: no deep structure, no underlying transformations
- Basic unit of analysis: pairing of form and meaning (construction)

Construction Name	Construction Template	Examples
Word		Banana
Word (partially filled)	pre-N, V-ing	Pretransition, Working
Idiom (filled)		Give the devil his due
Idiom (partially filled)	Jog <someone's> memory	She jogged his memory
Idiom (minimally filled)	The X-er the Y-er	The more I think about it, the less I know
Ditransitive construction (unfilled)	Subj V Obj1 Obj2	He baked her a muffin
Passive (unfilled)	Subj aux VPpp (PP by)	The armadillo was hit by a car

Table 1: Standard examples of constructions at various levels, adapted from Goldberg (2013)

Probing for Construction Grammar: Key Questions

- If this is how humans process language, do language models, too?
- To what extent do language models acquire constructions?
- If they can identify the construction, do they learn what it means?

Probing for the English Comparative Correlative (CC)

Comparative phrase

Sentences being compared, with long-distance filler-gap dependency and/or ellipsis

Fixed ‚the‘ at the start of each half

The funnier the example,

the more citations the paper will have.

→ If the example is funnier, the paper will have more citations.

→ As the funniness of the example increases, so will the citations of the paper.

How can we probe whether LMs understand this construction?

→ Split into two questions

Can PLMs learn the **syntactic** features of the construction?

Can PLMs learn the **semantic** features of the construction?

Syntactic Features: Probing Setup

Question: Can the model distinguish CC sentences from non-CC sentences?

→ Find minimal pairs of sentences that differ only in this one feature: do they include the CC?

→ Difficulty: finding very similar-looking sentences, that are still grammatically acceptable, and don‘t give any exploitable clues to the probing model

Minimal Pairs

First idea: **Minimal Pairs from corpora**

☑ She thinks the more water she drinks the better her skin looks.

✗ The way the older guys help out the younger guys is fantastic.

Easy vocabulary workarounds for the probing classifier, like occurrences of ‘the’

→ Complementary: **Minimal Pairs generated by a CFG**

☑ The flatter the fourteen lions push , the deeper and smaller the sixteen deer burn under the roof.

✗ The flatter fourteen push the lions, the deeper and smaller sixteen burn the deer under the roof.

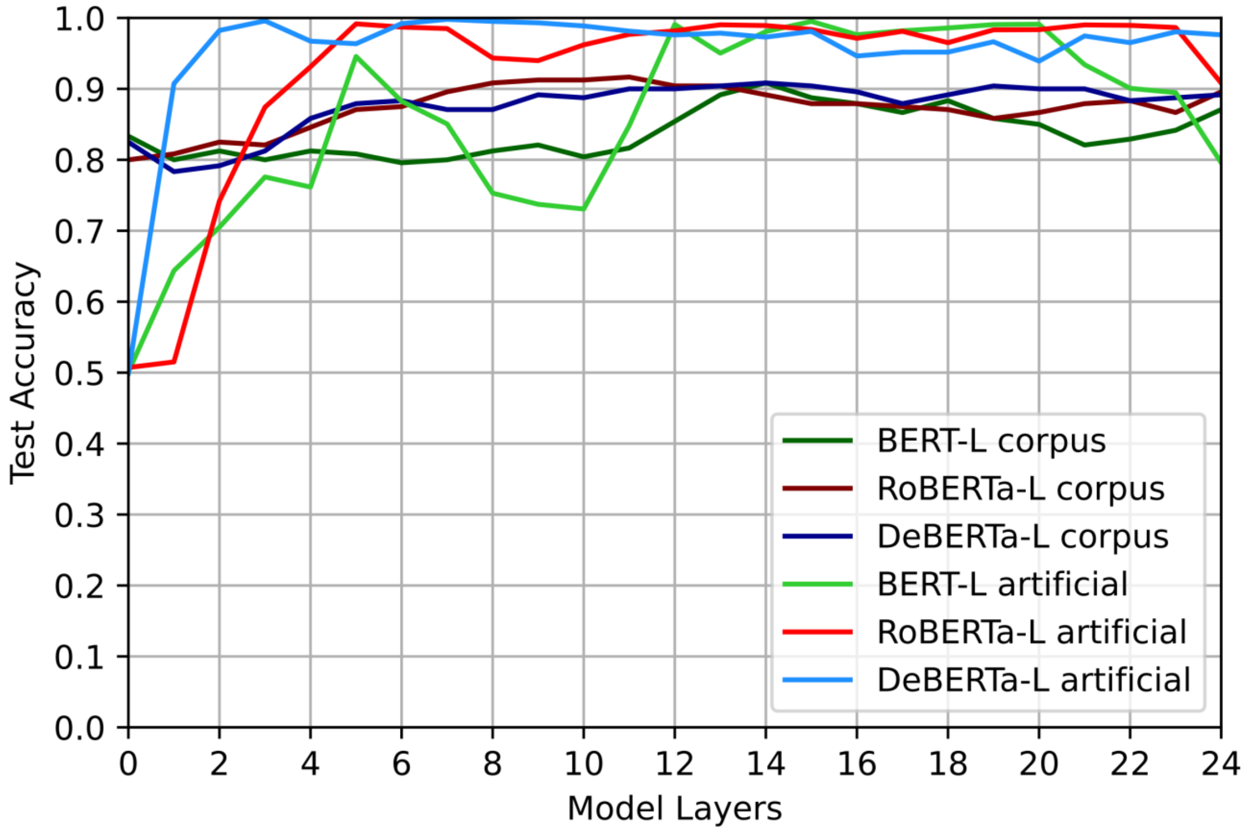
Syntax Probing Results

- Models: BERT, RoBERTa, DeBERTa (large)
- One-layer perceptron as probing classifier on top of every layer's contextual embeddings

→ Artificial sentences are at 50% accuracy on embedding layer, corpus sentences at 80%

→ 90% or better accuracy for all models

→ The *form* of the CC seems to be recognised



Probing accuracies for each model, layer, and data type

Semantic Features: Probing Setup

Question: Can PLMs understand the meaning of the CC?

→ Can they use information given to them in a CC in a NLU task?

The stronger you are, the faster you are. Terry is stronger than John. Therefore, Terry will be [MASK] than John.

→ Can the model correctly predict faster?

Problem: the wrong answer should be included in the context

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.

→ p(faster) > p(slower)?

Bias

Bias: the model could always predict the adjective closest to the [MASK].

→ *recency bias*

Test: swap first two sentences

S2: The weaker you are, the slower you are. The stronger you are, the faster you are. Terry is stronger than John. Therefore, Terry will be [MASK] than John.

Bias

Bias: the model could always predict the more frequent adjective.

→ *vocabulary bias*

Test: swap sentence halves so that the correct answer changes

S3: 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.

Bias

Bias: the model could associate some names strongly with some adjectives

→ *name bias*

Test: swap names

S4: The weaker you are, the slower you are. The stronger you are, the faster you are. John is stronger than Terry. Therefore, John will be [MASK] than Terry.

First Results

S2: test for recency bias

S3: test for vocabulary bias

S4: test for name bias

	Accuracy		Decision Flip		
	S1	S2	S2	S3	S4
BERT_{B}	37.65	64.64	26.98	75.69	02.70
BERT_{L}	36.85	67.21	30.44	73.31	02.32
$\text{RoBERTa}_{\text{B}}$	61.60	52.84	09.91	76.18	02.76
$\text{RoBERTa}_{\text{L}}$	55.71	68.00	14.33	79.47	04.33
$\text{DeBERTa}_{\text{B}}$	49.72	49.80	00.91	99.66	01.07
$\text{DeBERTa}_{\text{L}}$	50.88	51.40	07.04	94.83	02.23
$\text{DeBERTa}_{\text{XL}}$	47.73	49.33	05.46	89.28	02.51
$\text{DeBERTa}_{\text{XXL}}$	47.34	48.72	03.59	82.09	01.13

→ Accuracy is consistently better when the correct answer is closer to the MASK

→ Changing the correct answer by swapping sentence halves very strongly influences the answer

→ No recoverable significant performance from any of the models

One last chance: Calibration

Idea: if we can measure the ‚default' probabilities for each answer before we give the model any information, we can *calibrate* the actual answer by dividing by the default

C1: leave out CC sentence

→ Terry is stronger than John. Therefore, Terry will be [MASK] than John.

C2: add two unrelated names

→ The stronger you are, the faster you are. The weaker you are, the slower you are. Terry is stronger than John. Therefore, Eric will be [MASK] than Michael.

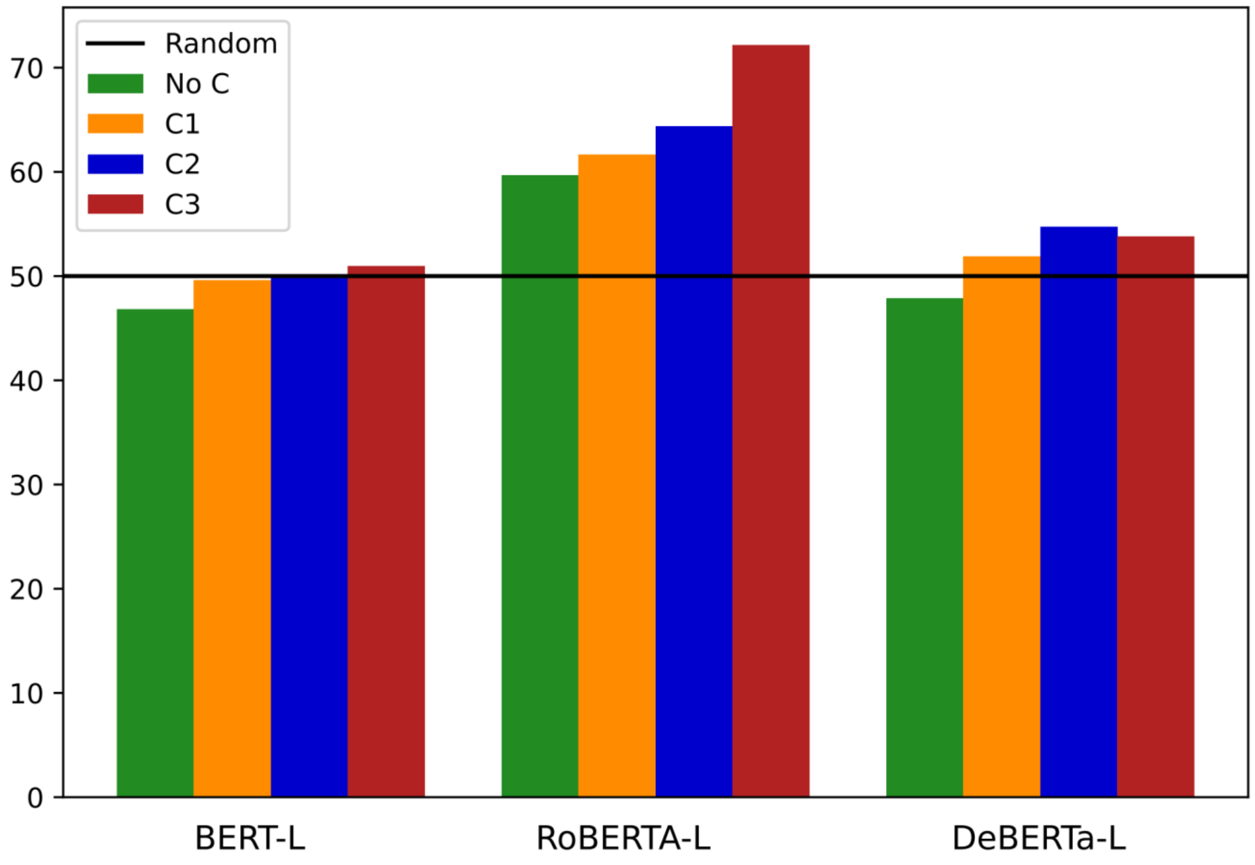
C3: add a third adjective

→ The weaker you are, the slower you are. The stronger you are, the faster you are. Terry is funnier than John. Therefore, Terry will be [MASK] than John.

Calibrated Results

- All calibration methods were somewhat helpful, especially for RoBERTa
- BERT and DeBERTA perform at chance level
- RoBERTa gets up to 70% accuracy

→ We can not conclude that PLMs understand the CC



Calibrated and averaged accuracies for each model

Takeaways

We saw that…

- The English Comparative Correlative is an interesting construction with many complex features
- PLMs can reliably distinguish CC sentences from non-CC sentences
- PLMS struggle to understand and use CC meaning in our setup

Thank you for listening!

Leonie Weissweiler
weissweiler@cis.lmu.de · www.cis.lmu.de/~weissweiler · @laweissweiler