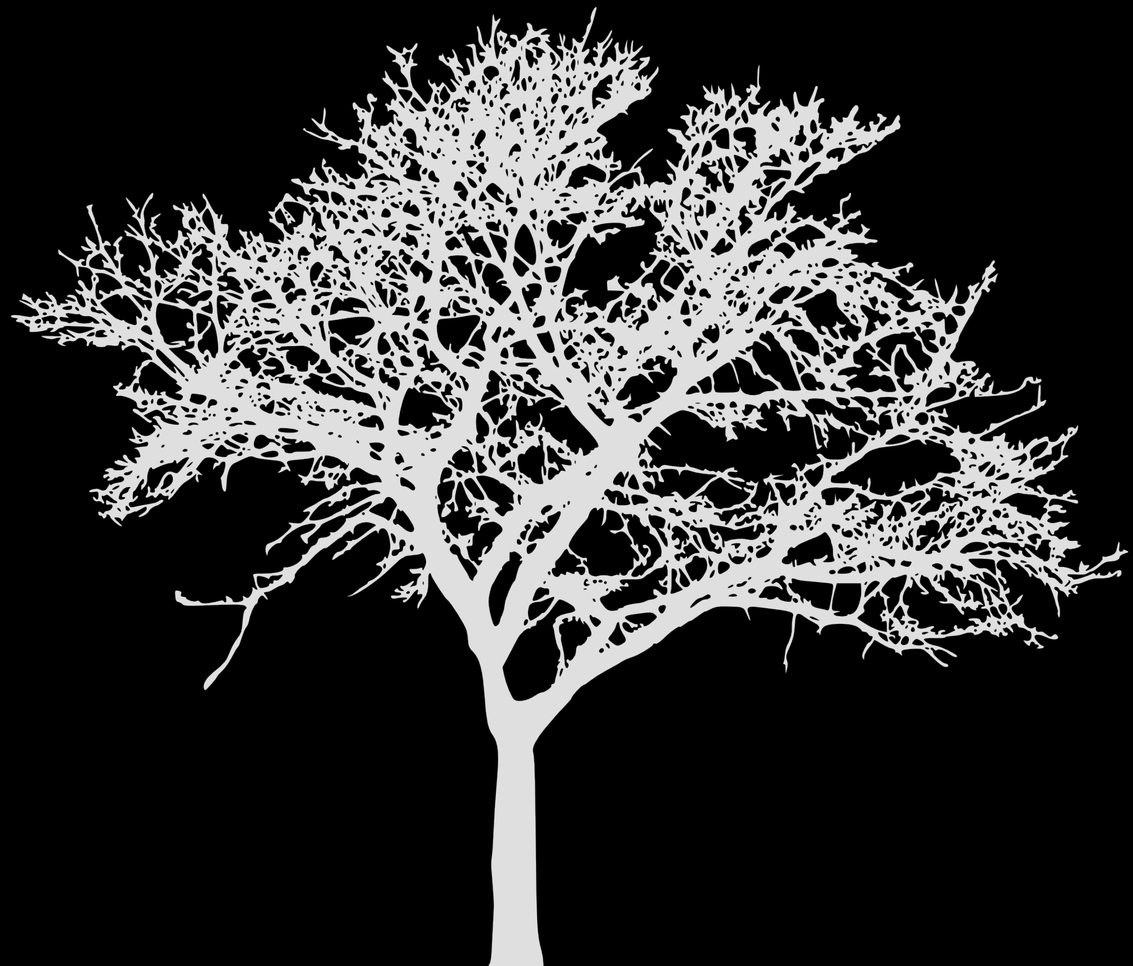
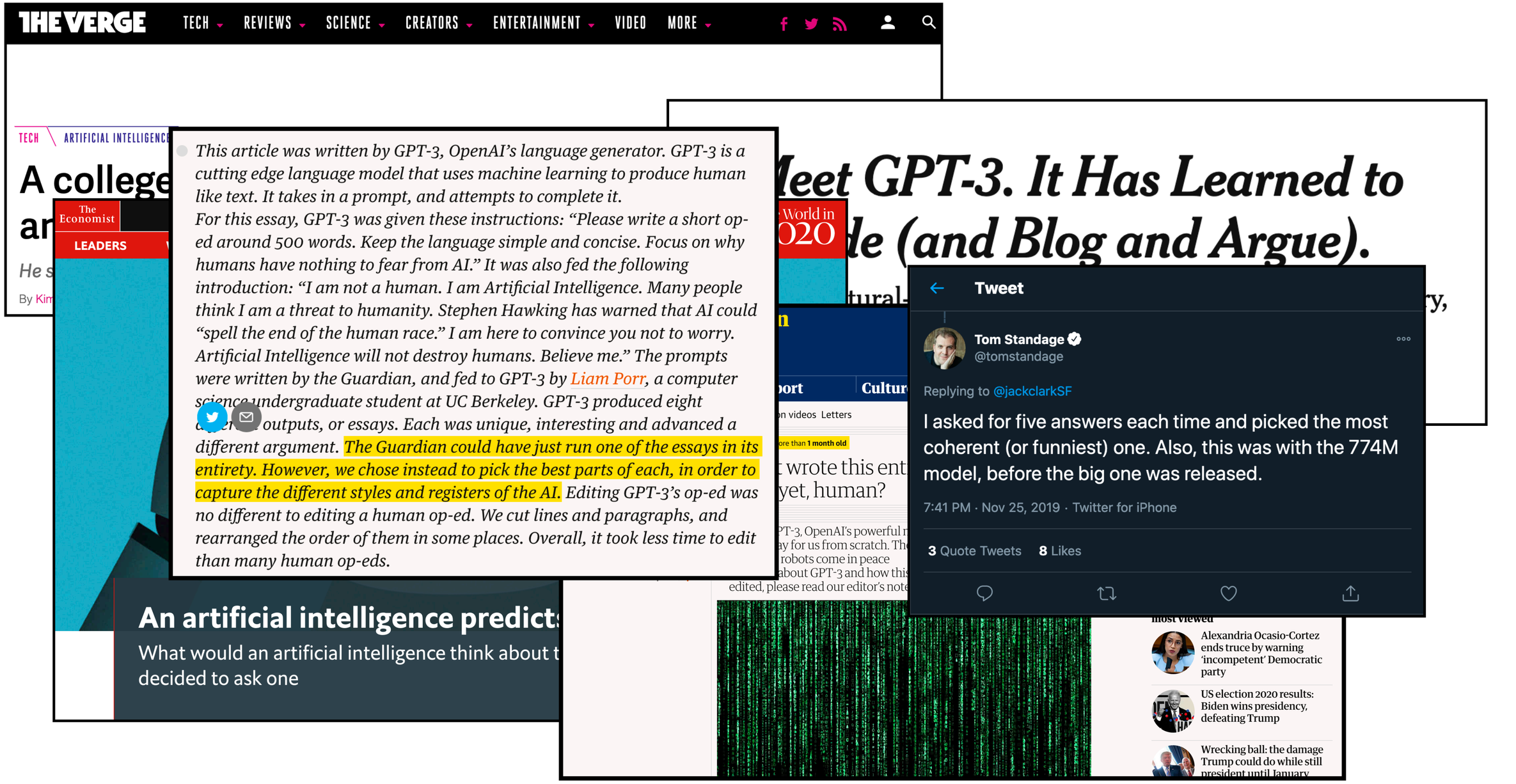
Schrödinger's Tree: On Syntax and Language Models



Artur Kulmizev
Uppsala University


THE VERGE
TECH
REVIEWS
SCIENCE
CREATORS
ENTERTAINMENT
VIDEO
MORE
TECH
ARTIFICIAL INTELLIGENCE
A college
ar
He s
By Kim
eet GPT-3. It Has Learned to
le (and Blog and Argue).
The Economist
LEADERS
World in
020
This article was written by GPT-3, OpenAI's language generator. GPT-3 is a cutting edge language model that uses machine learning to produce human like text. It takes in a prompt, and attempts to complete it.
For this essay, GPT-3 was given these instructions: "Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI." It was also fed the following introduction: "I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could "spell the end of the human race." I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me." The prompts were written by the Guardian, and fed to GPT-3 by Liam Porr, a computer science undergraduate student at UC Berkeley. GPT-3 produced eight outputs, or essays. Each was unique, interesting and advanced a different argument. The Guardian could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI. Editing GPT-3's op-ed was no different to editing a human op-ed. We cut lines and paragraphs, and rearranged the order of them in some places. Overall, it took less time to edit than many human op-eds.
An artificial intelligence predicts
What would an artificial intelligence think about t
decided to ask one
Tweet
Tom Standage
@tomstandage
Replying to @jackclarkSF
I asked for five answers each time and picked the most coherent (or funniest) one. Also, this was with the 774M model, before the big one was released.
7:41 PM · Nov 25, 2019 · Twitter for iPhone
3 Quote Tweets
8 Likes
Culture
videos Letters
more than 1 month old
wrote this ent
yet, human?
PT-3, OpenAI's powerful
ay for us from scratch. The
robots come in peace
about GPT-3 and how this
edited, please read our editor's note
Most viewed
Alexandria Ocasio-Cortez ends truce by warning 'incompetent' Democratic party
US election 2020 results: Biden wins presidency, defeating Trump
Wrecking ball: the damage Trump could do while still president until January

SuperGLUE GLUE

Leaderboard Version: **2.0**

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b
1	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	96.1/94.1	72.3
2	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	93.3/95.5	67.8
3	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	92.7/94.7	68.6
4	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	92.7/91.9	69.1
5	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	93.3/93.8	66.7
6	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
7	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	92.7/91.9	65.6
8	Descartes Team	frozen T5 1.1 + SPoT						95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	83.1/82.6	66.9
29	Zhuosheng Zhang	SemBERT				82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4

Click on a submission to see more information

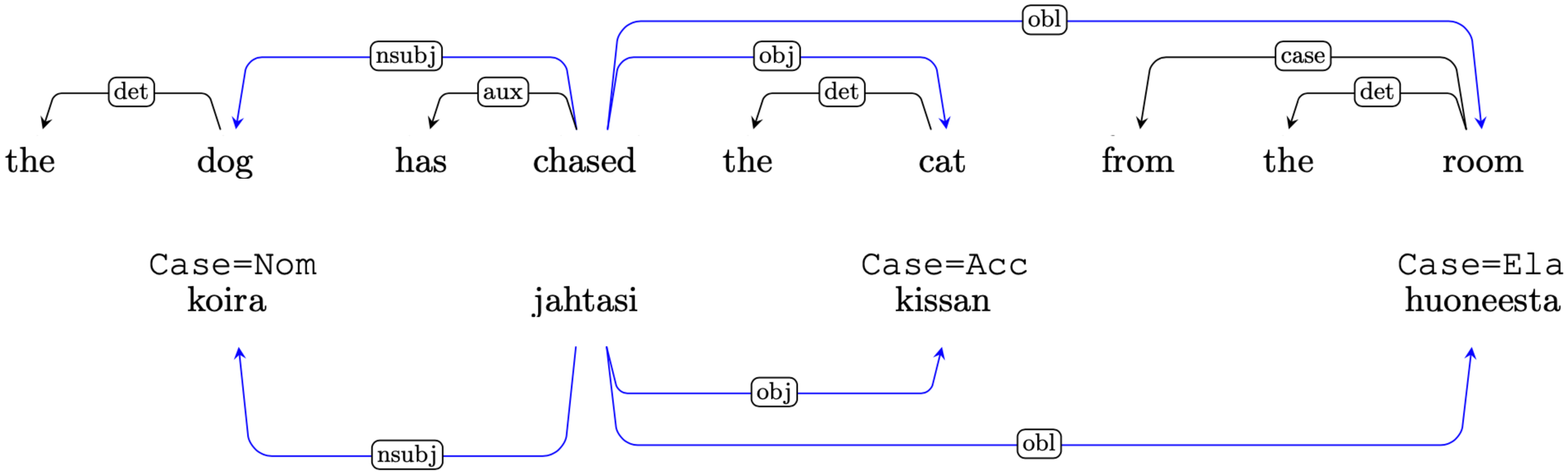
Are language models really *good* at language?

Do they really make human-like (linguistic) decisions?

How do we begin to investigate this?


syntax

What *is* syntax?

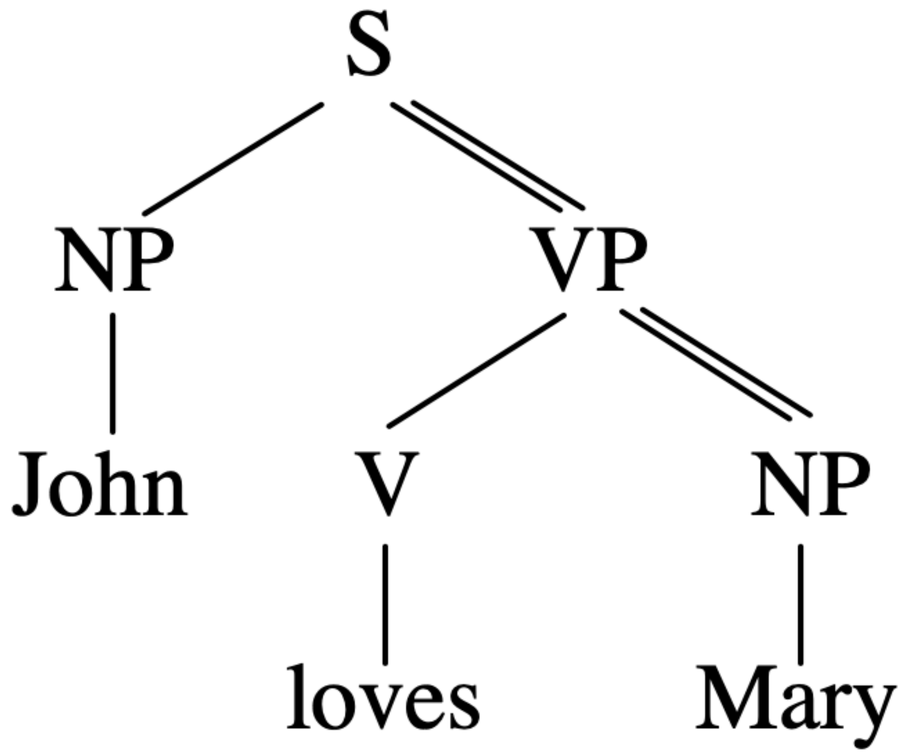


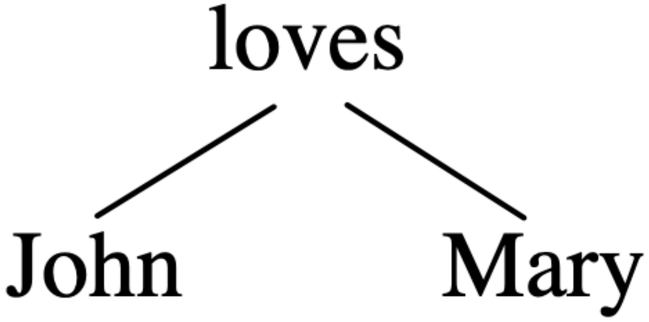
Important to note that…

1. coding properties only encode aspects of hierarchical structure, which is otherwise unobservable

2. syntactic theories vary highly in theoretical assumptions and representations

3. the “cognitive” nature of syntax is highly contested


S
NP
VP
John
V
NP
loves
Mary


loves
John
Mary

How do we investigate language models' syntactic knowledge?

targeted syntactic evaluation

Does a language model assign higher probability to a string A than it does to an ungrammatical, yet minimally different, string B?

How does model performance compare to humans?

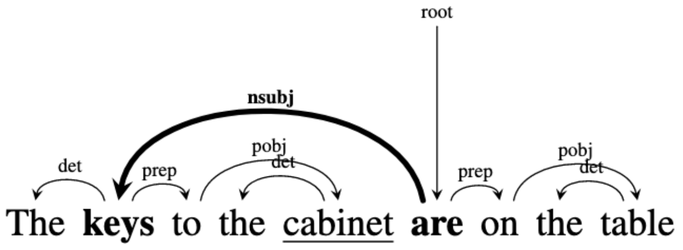
Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen[1,2] **Emmanuel Dupoux**[1]
LSCP[1] & IJN[2], CNRS,
EHESS and ENS, PSL Research University
{tal.linzen,
emmanuel.dupoux}@ens.fr

Yoav Goldberg
Computer Science Department
Bar Ilan University
yoav.goldberg@gmail.com

(1) a. The **key is** on the table.
b. *The **key are** on the table.
c. *The **keys is** on the table.
d. The **keys are** on the table.

(2) The **keys** to the cabinet **are** on the table.



(4) Alluvial **soils** carried in the *floodwaters* **add** nutrients to the floodplains.

(5) The only championship **banners** that are currently displayed within the building **are** for national or NCAA Championships.

(6) The **length** of the forewings **is** 12-13.

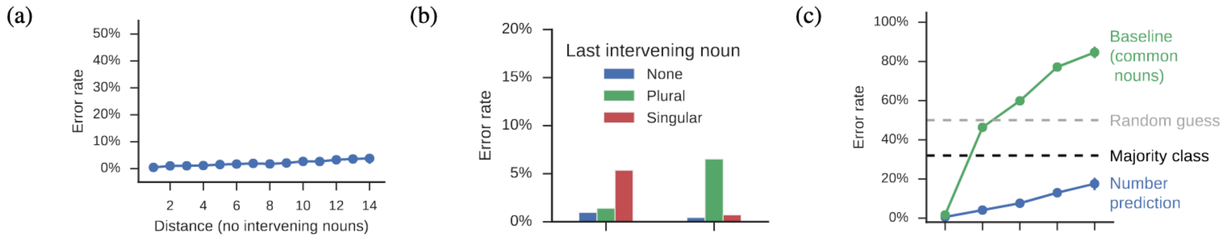
(7) Yet the **ratio** of men who survive to the women and children who survive **is** not clear in this story.

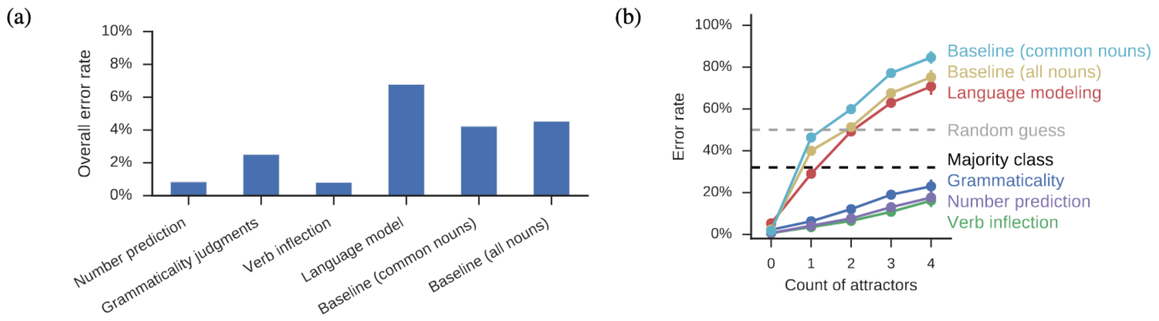
Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen[1,2] **Emmanuel Dupoux**[1]
LSCP[1] & IJN[2], CNRS,
EHESS and ENS, PSL Research University
{tal.linzen,
emmanuel.dupoux}@ens.fr

Yoav Goldberg
Computer Science Department
Bar Ilan University
yoav.goldberg@gmail.com

Training objective	Sample input	Training signal	Prediction task	Correct answer
Number prediction	*The keys to the cabinet*	PLURAL	SINGULAR/PLURAL?	PLURAL
Verb inflection	*The keys to the cabinet [is/are]*	PLURAL	SINGULAR/PLURAL?	PLURAL
Grammaticality	*The keys to the cabinet are here.*	GRAMMATICAL	GRAMMATICAL/UNGRAMMATICAL?	GRAMMATICAL
Language model	*The keys to the cabinet*	are	$P(are) > P(is)$?	True





Colorless green recurrent networks dream hierarchically

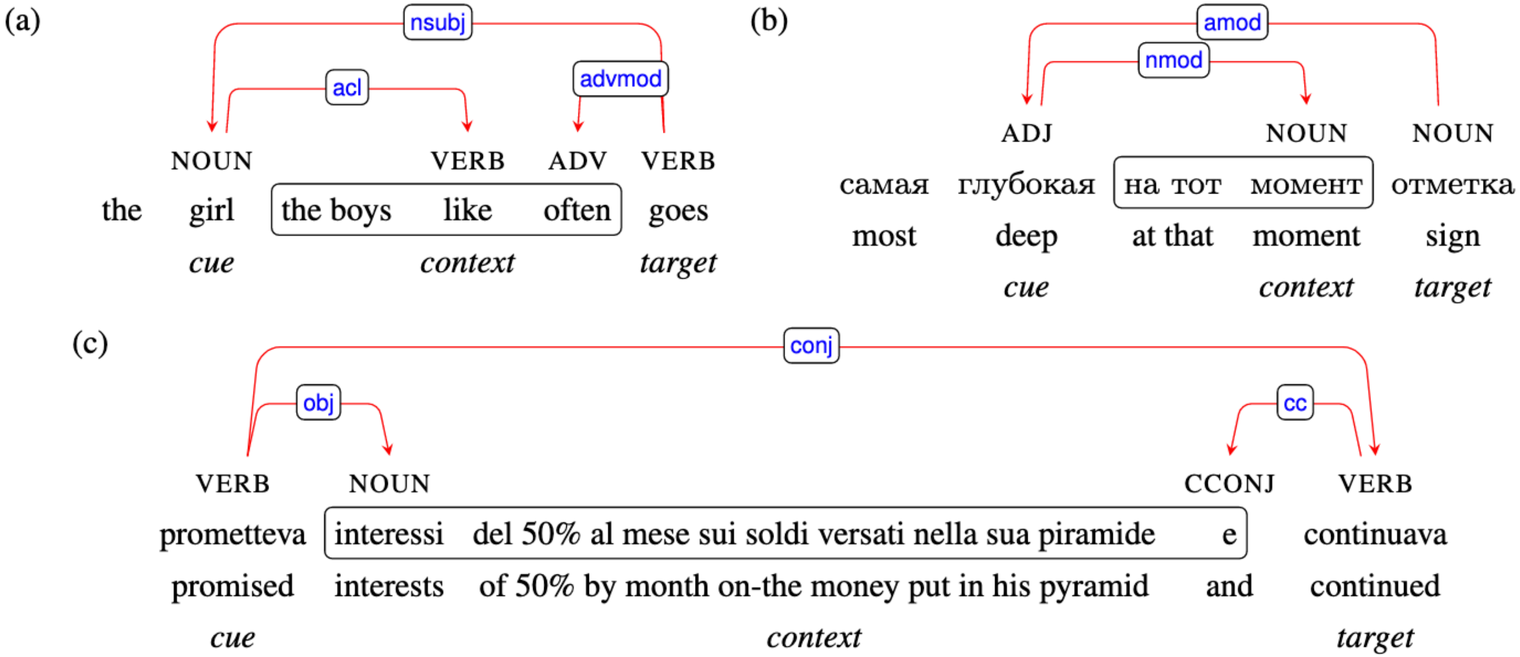
Kristina Gulordava*
Department of Linguistics
University of Geneva
kristina.gulordava@unige.ch

Piotr Bojanowski
Facebook AI Research
Paris
bojanowski@fb.com

Edouard Grave
Facebook AI Research
New York
egrave@fb.com

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

Marco Baroni
Facebook AI Research
Paris
mbaroni@fb.com



Colorless green recurrent networks dream hierarchically

Kristina Gulordava*
Department of Linguistics
University of Geneva
kristina.gulordava@unige.ch

Piotr Bojanowski
Facebook AI Research
Paris
bojanowski@fb.com

Edouard Grave
Facebook AI Research
New York
egrave@fb.com

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

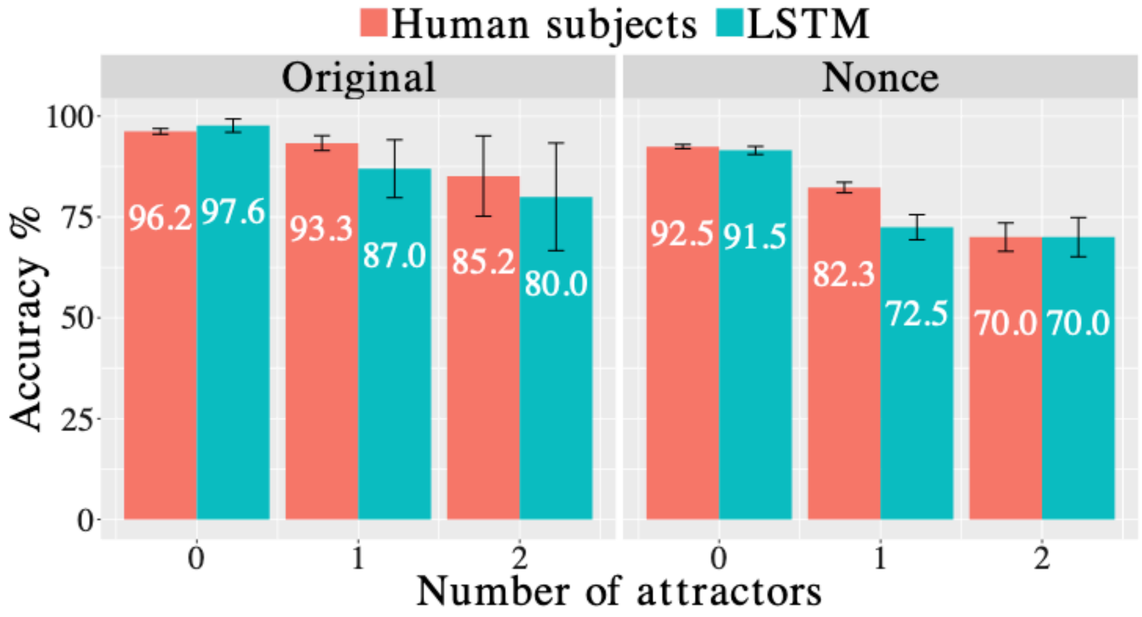
Marco Baroni
Facebook AI Research
Paris
mbaroni@fb.com

	IT	EN	HE	RU
#constructions	8	2	18	21
#original	119	41	373	442
Unigram				
Original	54.6	65.9	67.8	60.2
Nonce	54.1	42.5	63.1	54.0
5-gram KN				
Original	63.9	63.4	72.1	73.5
Nonce	52.8	43.4	61.7	56.8
Perplexity	147.8	168.9	122.0	166.6
5-gram LSTM				
Original	81.8 ±3.2	70.2 ±5.8	90.9 ±1.2	91.5 ±0.4
Nonce	78.0 ±1.3	58.2 ±2.1	77.5 ±0.8	85.7 ±0.7
Perplexity	62.6 ±0.2	71.6 ±0.3	59.9 ±0.2	61.1 ±0.4
LSTM				
Original	92.1 ±1.6	81.0 ±2.0	94.7 ±0.4	96.1 ±0.7
Nonce	85.5 ±0.7	74.1 ±1.6	80.8 ±0.8	88.8 ±0.9
Perplexity	45.2 ±0.3	52.1 ±0.3	42.5 ±0.2	48.9 ±0.6

		N V V	V NP conj V
Italian	Original	$93.3_{\pm 4.1}$	$83.3_{\pm 10.4}$
	Nonce	$92.5_{\pm 2.1}$	$78.5_{\pm 1.7}$
English	Original	$89.6_{\pm 3.6}$	$67.5_{\pm 5.2}$
	Nonce	$68.7_{\pm 0.9}$	$82.5_{\pm 4.8}$
Hebrew	Original	$86.7_{\pm 9.3}$	$83.3_{\pm 5.9}$
	Nonce	$65.7_{\pm 4.1}$	$83.1_{\pm 2.8}$
Russian	Original	-	$95.2_{\pm 1.9}$
	Nonce	-	$86.7_{\pm 1.6}$

Table 2: LSTM accuracy in the constructions N V V (subject-verb agreement with an intervening embedded clause) and V NP conj V (agreement between conjoined verbs separated by a complement of the first verb).

(1) a. It presents the case for marriage equality and states...
b. It stays the shuttle for honesty insurance and finds...



Do Language Models Understand *Anything*?

On the Ability of LSTMs to Understand Negative Polarity Items

Jaap Jumelet **Dieuwke Hupkes**

What do RNN Language Models Learn about Filler–Gap Dependencies?

Ethan Wilcox[1], **Roger Levy**[2], **Takashi Morita**[3,4], and **Richard Futrell**[5]

Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State

Richard Futrell[1], **Ethan Wilcox**[2], **Takashi Morita**[3,4], **Peng Qian**[5], **Miguel Ballesteros**[6], and **Roger Levy**[5]

Do RNNs learn human-like abstract word order preferences?

Richard Futrell[1] and **Roger P. Levy**[2]

Targeted Syntactic Evaluation of Language Models

Rebecca Marvin **Tal Linzen**

	RNN	Multitask	*n*-gram	Humans	# sents
SUBJECT-VERB AGREEMENT:					
Simple	0.94	1.00	0.79	0.96	280
In a sentential complement	0.99	0.93	0.79	0.93	3360
Short VP coordination	0.90	0.90	0.51	0.94	1680
Long VP coordination	0.61	0.81	0.50	0.82	800
Across a prepositional phrase	0.57	0.69	0.50	0.85	44800
Across a subject relative clause	0.56	0.74	0.50	0.88	22400
Across an object relative clause	0.50	0.57	0.50	0.85	44800
Across an object relative (no *that*)	0.52	0.52	0.50	0.82	44800
In an object relative clause	0.84	0.89	0.50	0.78	44800
In an object relative (no *that*)	0.71	0.81	0.50	0.79	44800
REFLEXIVE ANAPHORA:					
Simple	0.83	0.86	0.50	0.96	560
In a sentential complement	0.86	0.83	0.50	0.91	6720
Across a relative clause	0.55	0.56	0.50	0.87	44800
NEGATIVE POLARITY ITEMS:					
Simple	0.40	0.48	0.06	0.98	792
Across a relative clause	0.41	0.73	0.60	0.81	31680

BLiMP: The Benchmark of Linguistic Minimal Pairs for English

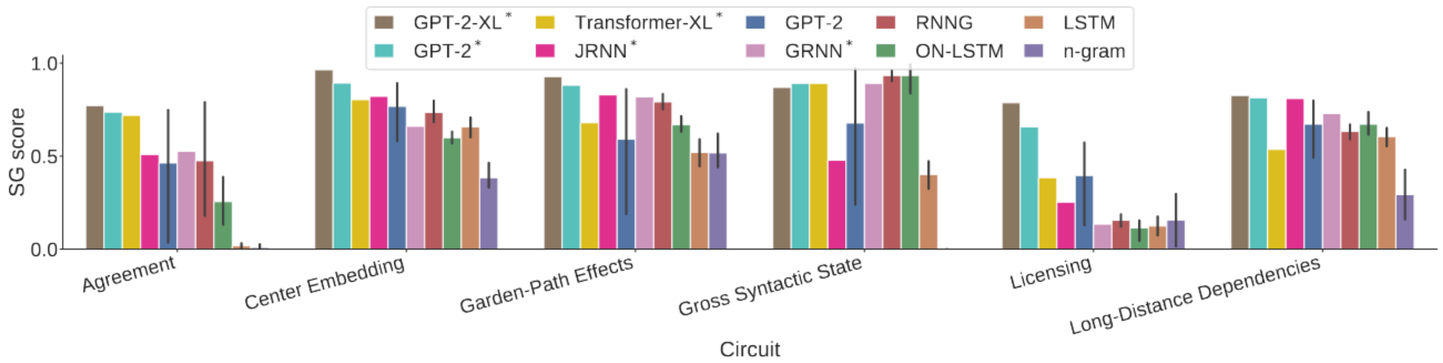
Alex Warstadt[1], Alicia Parrish[1], Haokun Liu[2], Anhad Mohananey[2], Wei Peng[2], Sheng-FuWang[1], Samuel R. Bowman[1,2,3]

Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	61.2	47.9	71.9	64.4	68.5	70.0	36.9	60.2	79.5	57.2	45.5	53.5	60.3
LSTM	69.8	91.7	73.2	73.5	67.0	85.4	67.6	73.9	89.1	46.6	51.7	64.5	80.1
TXL	69.6	94.1	69.5	74.7	71.5	83.0	77.2	66.6	78.2	48.4	55.2	69.3	76.0
GPT-2	81.5	99.6	78.3	80.1	80.5	93.3	86.6	81.3	84.1	70.6	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Table 3: Percentage accuracy of four baseline models and raw human performance on BLiMP using a forced-choice task. A random guessing baseline would achieve an accuracy of 50%.

A Systematic Assessment of Syntactic Generalization in Neural Language Models

Jennifer Hu[1], Jon Gauthier[1], Peng Qian[1], Ethan Wilcox[2], and Roger P. Levy[1]



How do we investigate language models' syntactic knowledge?

targeted syntactic evaluation

Does a language model assign higher probability to a string A than it does to an ungrammatical, yet minimally different, string B?

How does model performance compare to humans?

hidden state probing

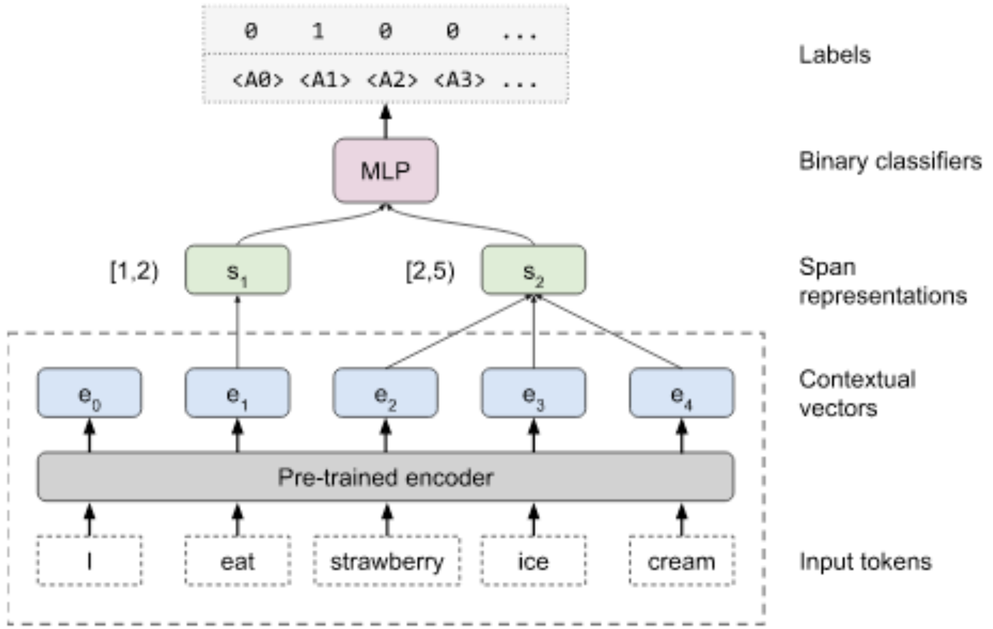
Can we decode linguistic information (e.g. POS, DEP) from the hidden states of a model with simple classifiers?

Which layers are responsible for encoding different types of linguistic information? And at what point during training does a model encode it?

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] **Patrick Xia,**[2] **Berlin Chen,**[3] **Alex Wang,**[4] **Adam Poliak,**[2]
R. Thomas McCoy,[2] **Najoung Kim,**[2] **Benjamin Van Durme,**[2] **Samuel R. Bowman,**[4]
Dipanjan Das,[1] **and Ellie Pavlick**[1,5]

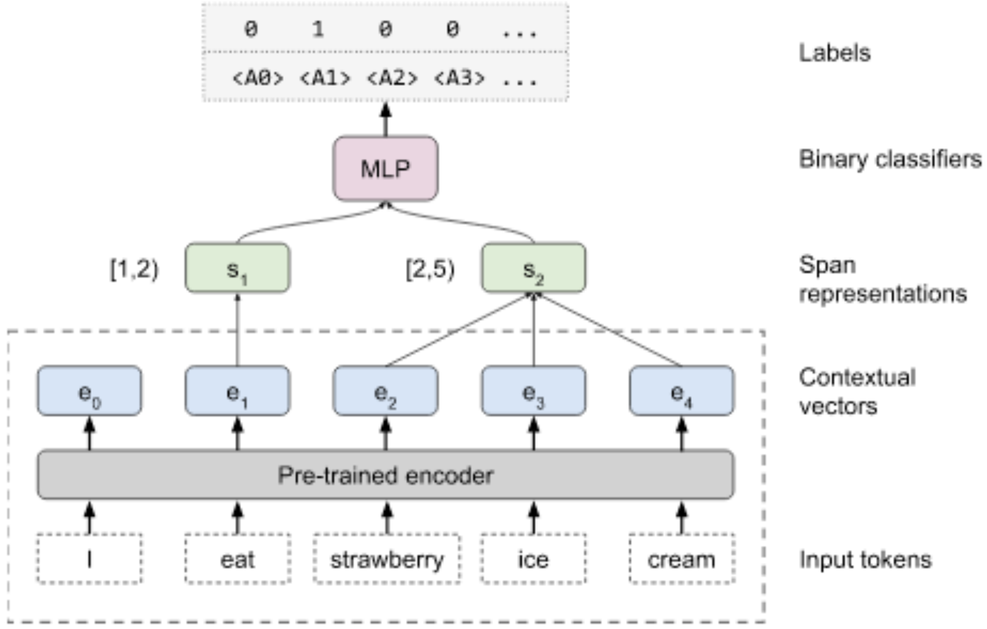
[1]Google AI Language, [2]Johns Hopkins University, [3]Swarthmore College,
[4]New York University, [5]Brown University



WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] **Patrick Xia,**[2] **Berlin Chen,**[3] **Alex Wang,**[4] **Adam Poliak,**[2]
R. Thomas McCoy,[2] **Najoung Kim,**[2] **Benjamin Van Durme,**[2] **Samuel R. Bowman,**[4]
Dipanjan Das,[1] **and Ellie Pavlick**[1,5]

[1]Google AI Language, [2]Johns Hopkins University, [3]Swarthmore College,
[4]New York University, [5]Brown University



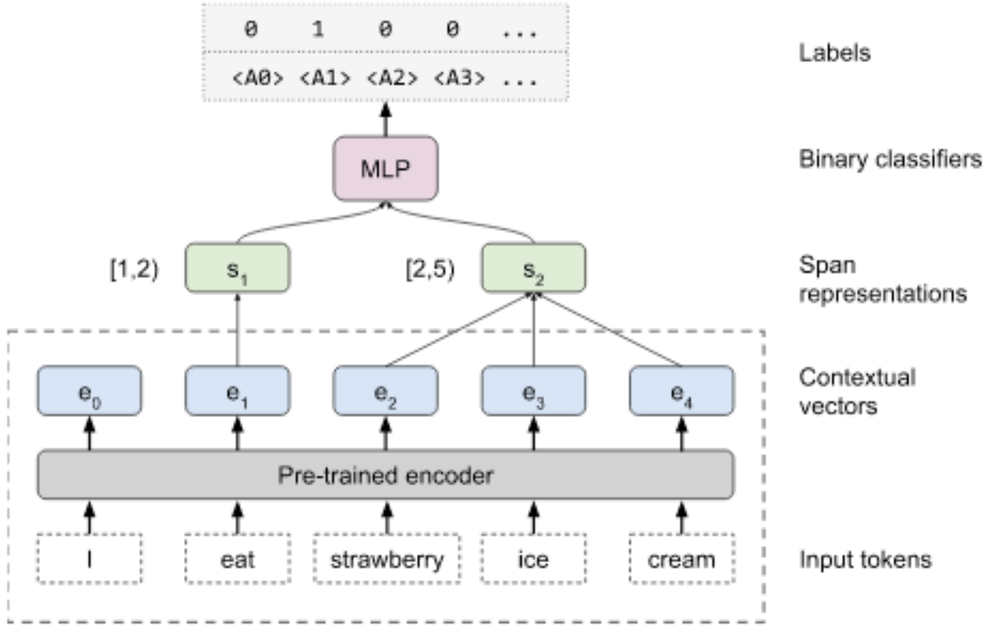
POS	The important thing about Disney is that it is a global $[\text{brand}]_1$. → NN (Noun)
Constit.	The important thing about Disney is that it $[\text{is a global brand}]_1$. → VP (Verb Phrase)
Depend.	$[\text{Atmosphere}]_1$ is always $[\text{fun}]_2$ → nsubj (nominal subject)
Entities	The important thing about $[\text{Disney}]_1$ is that it is a global brand. → Organization
SRL	$[\text{The important thing about Disney}]_2$ $[\text{is}]_1$ that it is a global brand. → Arg1 (Agent)
SPR	$[\text{It}]_1$ $[\text{endorsed}]_2$ the White House strategy. . . → {awareness, existed_after, . . . }
Coref.^{O}	The important thing about $[\text{Disney}]_1$ is that $[\text{it}]_2$ is a global brand. → True
Coref.^{W}	$[\text{Characters}]_2$ entertain audiences because $[\text{they}]_1$ want people to be happy. → True Characters entertain $[\text{audiences}]_2$ because $[\text{they}]_1$ want people to be happy. → False
Rel.	The $[\text{burst}]_1$ has been caused by water hammer $[\text{pressure}]_2$. → Cause-Effect(e_2, e_1)

Table 1: Example sentence, spans, and target label for each task. O = OntoNotes, W = Winograd.

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] **Patrick Xia,**[2] **Berlin Chen,**[3] **Alex Wang,**[4] **Adam Poliak,**[2]
R. Thomas McCoy,[2] **Najoung Kim,**[2] **Benjamin Van Durme,**[2] **Samuel R. Bowman,**[4]
Dipanjan Das,[1] **and Ellie Pavlick**[1,5]

[1]Google AI Language, [2]Johns Hopkins University, [3]Swarthmore College,
[4]New York University, [5]Brown University



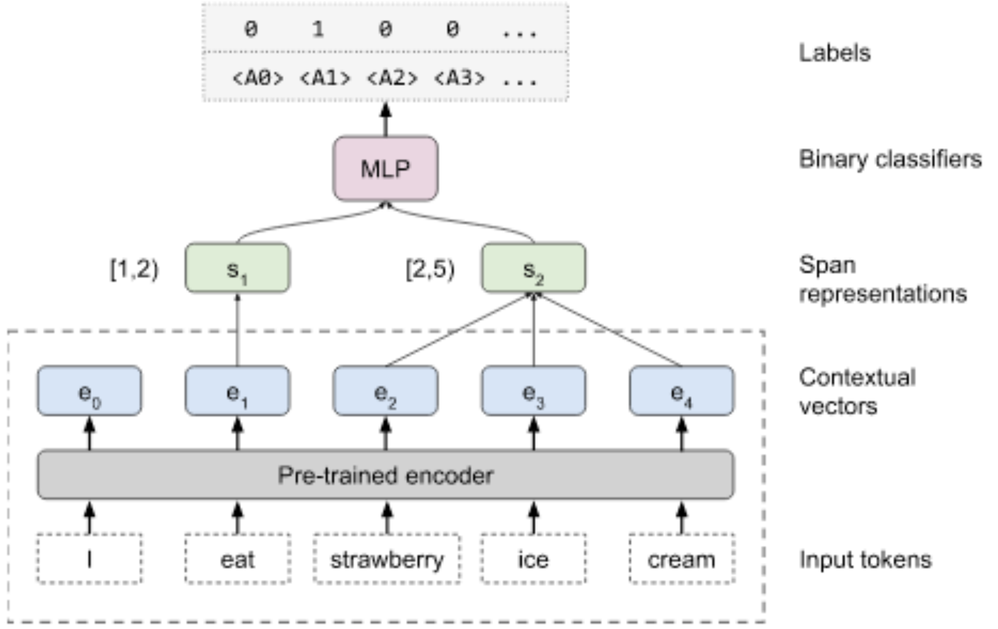
	CoVe			ELMo			GPT		
	Lex.	**Full**	**Abs. Δ**	**Lex.**	**Full**	**Abs. Δ**	**Lex.**	**cat**	**mix**
Part-of-Speech	85.7	94.0	8.4	90.4	**96.7**	6.3	88.2	94.9	95.0
Constituents	56.1	81.6	25.4	69.1	**84.6**	15.4	65.1	81.3	**84.6**
Dependencies	75.0	83.6	8.6	80.4	**93.9**	13.6	77.7	92.1	**94.1**
Entities	88.4	90.3	1.9	92.0	**95.6**	3.5	88.6	92.9	92.5
SRL (all)	59.7	80.4	20.7	74.1	**90.1**	16.0	67.7	86.0	89.7
Core roles	*56.2*	*81.0*	*24.7*	*73.6*	***92.6***	*19.0*	*65.1*	*88.0*	*92.0*
Non-core roles	*67.7*	*78.8*	*11.1*	*75.4*	***84.1***	*8.8*	*73.9*	*81.3*	***84.1***
OntoNotes coref.	72.9	79.2	6.3	75.3	84.0	8.7	71.8	83.6	**86.3**
SPR1	73.7	77.1	3.4	80.1	**84.8**	4.7	79.2	83.5	83.1
SPR2	76.6	80.2	3.6	82.1	83.1	1.0	82.2	**83.8**	**83.5**
Winograd coref.	52.1	**54.3**	2.2	**54.3**	**53.5**	-0.8	51.7	52.6	**53.8**
Rel. (SemEval)	51.0	60.6	9.6	55.7	77.8	22.1	58.2	**81.3**	**81.0**
Macro Average	69.1	78.1	9.0	75.4	**84.4**	9.1	73.0	83.2	**84.4**

	BERT-base				BERT-large				
	F1 Score			Abs. Δ	F1 Score			Abs. Δ	
	Lex.	**cat**	**mix**	**ELMo**	**Lex.**	**cat**	**mix**	**(base)**	**ELMo**
Part-of-Speech	88.4	**97.0**	96.7	0.0	88.1	96.5	**96.9**	0.2	0.2
Constituents	68.4	83.7	86.7	2.1	69.0	80.1	**87.0**	0.4	2.5
Dependencies	80.1	93.0	95.1	1.1	80.2	91.5	**95.4**	0.3	1.4
Entities	90.9	96.1	96.2	0.6	91.8	96.2	**96.5**	0.3	0.9
SRL (all)	75.4	89.4	91.3	1.2	76.5	88.2	**92.3**	1.0	2.2
Core roles	*74.9*	*91.4*	*93.6*	*1.0*	*76.3*	*89.9*	***94.6***	*1.0*	*2.0*
Non-core roles	*76.4*	*84.7*	*85.9*	*1.8*	*76.9*	*84.1*	***86.9***	*1.0*	*2.8*
OntoNotes coref.	74.9	88.7	90.2	6.3	75.7	89.6	**91.4**	1.2	7.4
SPR1	79.2	84.7	**86.1**	1.3	79.6	85.1	**85.8**	-0.3	1.0
SPR2	81.7	83.0	**83.8**	0.7	81.6	83.2	**84.1**	0.3	1.0
Winograd coref.	54.3	53.6	54.9	1.4	53.0	53.8	**61.4**	6.5	7.8
Rel. (SemEval)	57.4	78.3	82.0	4.2	56.2	77.6	**82.4**	0.5	4.6
Macro Average	75.1	84.8	86.3	1.9	75.2	84.2	**87.3**	1.0	2.9

Table 2: Comparison of representation models and their respective lexical baselines. Numbers reported are micro-averaged F1 score on respective test sets. **Lex.** denotes the lexical baseline (§ 4) for each model, and bold denotes the best performance on each task. Lines in *italics* are subsets of the targets from a parent task; these are omitted in the macro average. SRL numbers consider core and non-core roles, but ignore references and continuations. Winograd (DPR) results are the average of five runs each using a random sample (without replacement) of 80% of the training data. 95% confidence intervals (normal approximation) are approximately ±3 (±6 with BERT-large) for Winograd, ±1 for SPR1 and SPR2, and ±0.5 or smaller for all other tasks.

BERT Rediscovers the Classical NLP Pipeline

Ian Tenney[1] **Dipanjan Das**[1] **Ellie Pavlick**[1,2]
[1]Google Research [2]Brown University
{iftenney,dipanjand,epavlick}@google.com



$$\mathbf{h}_{i,\tau} = \gamma_\tau \sum_{\ell=0}^{L} s_\tau^{(\ell)} \mathbf{h}_i^{(\ell)} \qquad (1)$$

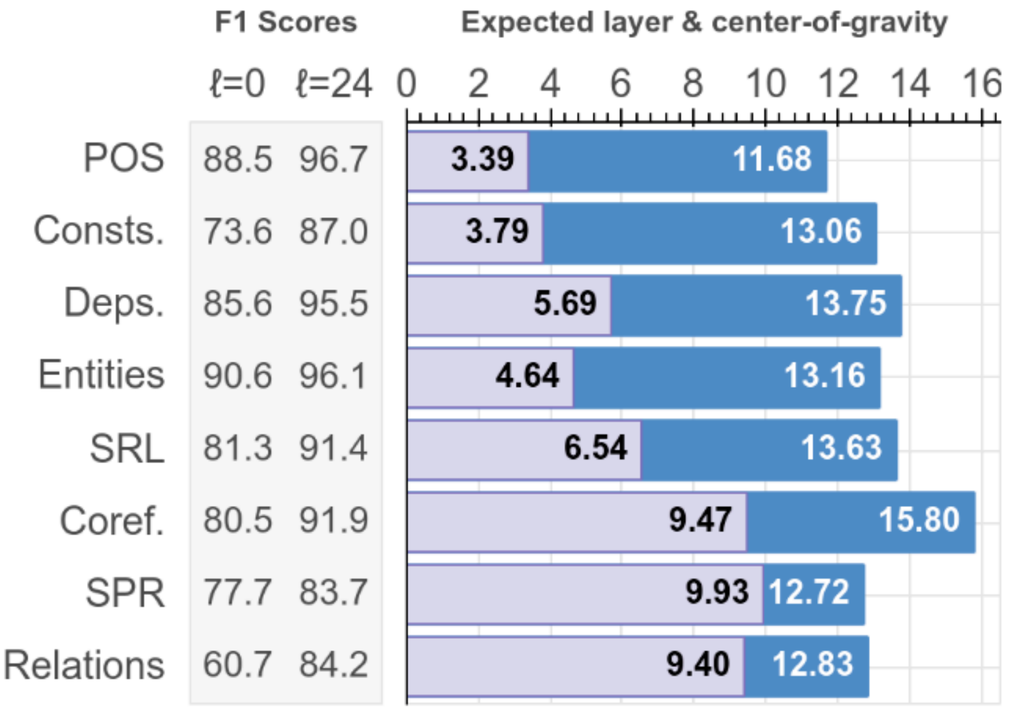


Figure 1: Summary statistics on BERT-large. Columns on left show F1 dev-set scores for the baseline ($P_\tau^{(0)}$) and full-model ($P_\tau^{(L)}$) probes. Dark (blue) are the mixing weight center of gravity (Eq. 2); light (purple) are the expected layer from the cumulative scores (Eq. 4).

probing for syntactic structure

representations are implicitly
hierarchical

“evolution” throughout layers
(e.g. POS → parse trees → SRL → coreference)

possible to probe for syntactic structure via
a linear transformation over hidden states

A Structural Probe for Finding Syntax in Word Representations

John Hewitt
Stanford University
johnhew@stanford.edu

Christopher D. Manning
Stanford University
manning@stanford.edu

distance and depth natural properties of syntax trees and vector spaces

theoretically possible to train probe to retrieve these properties

distance probe

$$\min_B \sum_{l=1}^{L} \frac{1}{|n^l|^2} \sum_{i,j} |\, d_{T^l}(w_i^l, w_j^l) - d_B(\mathbf{h}_i^l, \mathbf{h}_j^l)^2 \,|$$

depth probe

$$\min_B \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i}^{n} (\|w_i^l\| - \|B\mathbf{h}_i^l\|^2)$$

A Structural Probe for Finding Syntax in Word Representations

John Hewitt
Stanford University
johnhew@stanford.edu

Christopher D. Manning
Stanford University
manning@stanford.edu

Method	Distance UUAS	Distance DSpr.	Depth Root%	Depth NSpr.
LINEAR	48.9	0.58	2.9	0.27
ELMO0	26.8	0.44	54.3	0.56
DECAY0	51.7	0.61	54.3	0.56
PROJ0	59.8	0.73	64.4	0.75
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**

Table 1: Results of structural probes on the PTB WSJ test set; baselines in the top half, models hypothesized to encode syntax in the bottom half. For the distance probes, we show the Undirected Unlabeled Attachment Score (UUAS) as well as the average Spearman correlation of true to predicted distances, DSpr. For the norm probes, we show the root prediction accuracy and the average Spearman correlation of true to predicted norms, NSpr.

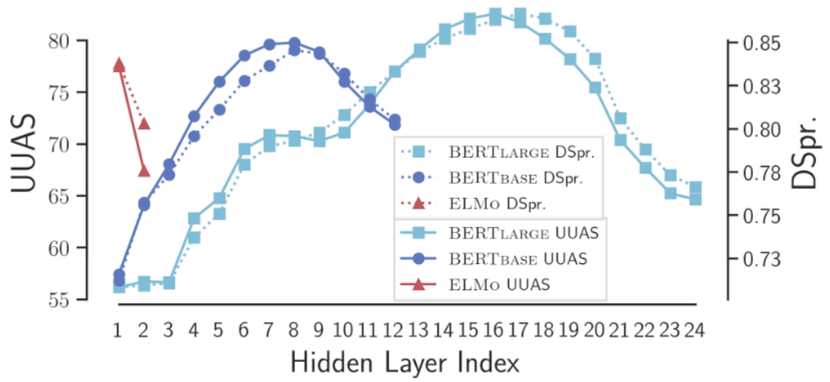


Figure 1: Parse distance UUAS and distance Spearman correlation across the BERT and ELMo model layers.

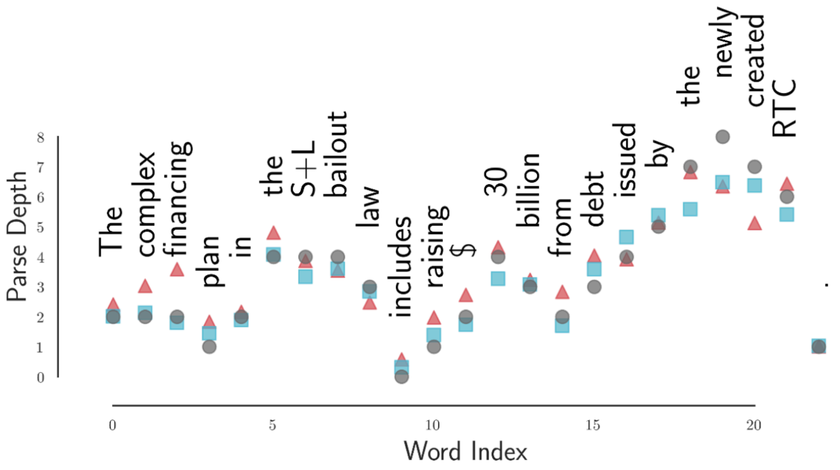


Figure 3: Parse tree depth according to the gold tree (black, circle) and the norm probes (squared) on ELMO1 (red, triangle) and BERTLARGE16 (blue, square).

How do we investigate language models' syntactic knowledge?

targeted syntactic evaluation

Does a language model assign higher probability to a string A than it does to an ungrammatical, yet minimally different, string B?

How does model performance compare to humans?

hidden state probing

Can we decode linguistic information (e.g. POS, DEP) from the hidden states of a model with simple classifiers?

Which layers are responsible for encoding different types of linguistic information? And at what point during training does a model encode it?

NLU evaluation

Does a model retain its performance when trained/tested on structurally perturbed input?

Does a model improve when imbued directly with syntactic structure?

BERT & Family Eat Word Salad: Experiments with Text Understanding

Ashim Gupta, Giorgi Kvernadze, Vivek Srikumar
University of Utah
{ashim, giorgi, svivek}@cs.utah.edu

Name	Description
`Sort`	Sort the input tokens
`Reverse`	Reverse the token sequence
`Shuffle`	Randomly shuffle tokens
`CopySort`	Copy one of the input texts and then sort it to create the second text. (Only applicable when the input is a pair of texts)

Table 2: Lexical-overlap based transformations

Name	Description
`Drop`	Drop the least important tokens.
`Repeat`	Replace the least important tokens with one of the most important ones.
`Replace`	Replace the least important tokens with random tokens from the vocabulary
`CopyOne`	Copy the most important token from one text as the sole token in the other. (Only applicable when the input is a pair of texts)

Table 3: Gradient-based transformations

Dataset	**Transform**	**Input**	**Prediction**
Natural Language Inference MNLI		P: As with other types of internal controls, this is a cycle of activity, not an exercise with a defined beginning and end.	
	Original	H: There is no clear beginning and end, it's a continuous cycle.	Ent (99.48%)
	`Shuffled`	H'_1: , beginning end no there clear 's continuous is a it and cycle .	Ent (99.60%)
	`PBSMT-E`	H'_2: The relationship of this is not a thing in the beginning .	Ent (94.82%)
Paraphrase Detection QQP		Q1: How do I find out what operating system I have on my Macbook?	
	Original	Q2: How do I find out what operating system I have?	Yes (99.53%)
	`Repeat`	Q2: out out i find out what out out i find?	Yes (99.98%)
	`CopySort`	Q2: ? do find have how i i macbook my on operating out system what	Yes (98.52%)
Sentiment Analysis SST-2	Original	A by-the-numbers effort that won't do much to enhance the franchise.	–ve (99.96%)
	`Sort`	a by-the-numbers do effort enhance franchise much n't that the to wo.	–ve (99.92%)
	`Drop`	a–n won do to franchise.	–ve (99.96%)

BERT & Family Eat Word Salad: Experiments with Text Understanding

Ashim Gupta, Giorgi Kvernadze, Vivek Srikumar
University of Utah
{ashim, giorgi, svivek}@cs.utah.edu

Transformation	% Invalid
Un-transformed	7.83
Sort	94.07
Reverse	95.59
Shuffle	94.20
CopySort	95.42
Avg. Lexical	94.82
Replace	91.21
Repeat	100.00
Drop	85.79
CopyOne	100.00
Avg. Gradient	94.25
PBSMT	79.92

Transform	MNLI	SNLI	QQP	MRPC	SST2
Sort	79.1	82.6	88.3	81.1	83.3
Reverse	76.9	75.1	86.8	77.9	82.5
Shuffle	79.4	81.1	88.4	80.4	84.8
CopySort	90.5	81.3	93.5	96.8	–
Avg. Lex.	82.4	80.1	89.3	84.1	83.5
Replace	63.0	51.9	69.9	56.6	78.1
Repeat	49.7	68.5	77.1	68.1	81.3
Drop	69.4	72.7	80.4	76.7	82.5
CopyOne	80.4	83.7	98.9	100	–
Avg. Grad.	65.6	69.2	81.6	75.4	80.6
PBSMT	57.0	65.6	72.5	–	75.2
Random	33.3	33.3	50.0	50.0	50.0

UnNatural Language Inference

Koustuv Sinha[1,2,3], Prasanna Parthasarathi[1,2], Joelle Pineau[1,2,3] and Adina Williams[3]
[1] School of Computer Science, McGill University, Canada
[2] Montreal Institute of Learning Algorithms (Mila), Canada
[3] Facebook AI Research (FAIR)
{koustuv.sinha, prasanna.parthasarathi, jpineau, adinawilliams}
@{mail.mcgill.ca, mail.mcgill.ca, cs.mcgill.ca, fb.com}

Premise	**Hypothesis**	**Predicted Label**
Boats in daily use lie within feet of the fashionable bars and restaurants.	There are boats close to bars and restaurants.	E
restaurants and use feet of fashionable lie the in Boats within bars daily .	bars restaurants are There and to close boats .	E
He and his associates weren't operating at the level of metaphor.	He and his associates were operating at the level of the metaphor.	C
his at and metaphor the of were He operating associates n't level .	his the and metaphor level the were He at associates operating of .	C

Model	Eval. Dataset	$\mathcal{A}$	Ω_{max}	$\mathcal{P}^c$	$\mathcal{P}^f$	Ω_{rand}
RoBERTa-Large	MNLI_m_dev	0.906	0.987	0.707	0.383	0.794
	MNLI_mm_dev	0.901	0.987	0.707	0.387	0.790
	SNLI_dev	0.879	0.988	0.768	0.393	0.826
	SNLI_test	0.883	0.988	0.760	0.407	0.828
	A1*	0.456	0.897	0.392	0.286	0.364
	A2*	0.271	0.889	0.465	0.292	0.359
	A3*	0.268	0.902	0.480	0.308	0.397
	Mean	0.652	0.948	0.611	**0.351**	0.623
BART-Large	MNLI_m_dev	0.902	0.989	0.689	0.393	0.784
	MNLI_mm_dev	0.900	0.986	0.695	0.399	0.788
	SNLI_dev	0.886	0.991	0.762	0.363	0.834
	SNLI_test	0.888	0.990	0.762	0.370	0.836
	A1*	0.455	0.894	0.379	0.295	0.374
	A2*	0.316	0.887	0.428	0.303	0.397
	A3*	0.327	0.931	0.428	0.333	0.424
	Mean	**0.668**	**0.953**	0.592	**0.351**	**0.634**
DistilBERT	MNLI_m_dev	0.800	0.968	0.775	0.343	0.779
	MNLI_mm_dev	0.811	0.968	0.775	0.346	0.786
	SNLI_dev	0.732	0.956	0.767	0.307	0.731
	SNLI_test	0.738	0.950	0.770	0.312	0.725
	A1*	0.251	0.750	0.511	0.267	0.300
	A2*	0.300	0.760	0.619	0.265	0.343
	A3*	0.312	0.830	0.559	0.259	0.363
	Mean	0.564	0.883	**0.682**	0.300	0.575

Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]

Adina Williams[†] **Douwe Kiela**[†]

[†] Facebook AI Research; [‡] McGill University / Mila - Quebec AI

{koustuvs,adinawilliams,dkiela}@fb.com

Model	QNLI	RTE	QQP	SST-2	MRPC	PAWS	MNLI-m/mm	CoLA
$\mathcal{M}_{\text{N}}$	92.45 +/- 0.2	73.62 +/- 3.1	91.25 +/- 0.1	93.75 +/- 0.4	89.09 +/- 0.9	94.49 +/- 0.2	86.08 +/- 0.2 / 85.4 +/- 0.2	52.45 +/- 21
$\mathcal{M}_{4}$	91.65 +/- 0.1	70.94 +/- 1.2	91.39 +/- 0.1	92.46 +/- 0.3	86.90 +/- 0.3	94.26 +/- 0.2	83.79 +/- 0.2 / 83.94 +/- 0.3	35.25 +/- 32
$\mathcal{M}_{3}$	91.56 +/- 0.4	69.75 +/- 2.8	91.22 +/- 0.1	91.97 +/- 0.5	86.22 +/- 0.8	94.03 +/- 0.1	83.83 +/- 0.2 / 83.71 +/- 0.1	40.78 +/- 23
$\mathcal{M}_{2}$	90.51 +/- 0.1	70.00 +/- 2.5	91.33 +/- 0.0	91.78 +/- 0.3	85.90 +/- 1.2	93.53 +/- 0.3	83.45 +/- 0.3 / 83.54 +/- 0.3	50.83 +/- 5.8
$\mathcal{M}_{1}$	89.05 +/- 0.2	68.48 +/- 2.5	91.01 +/- 0.0	90.41 +/- 0.4	86.06 +/- 0.8	89.69 +/- 0.6	82.64 +/- 0.1 / 82.67 +/- 0.2	31.08 +/- 10
$\mathcal{M}_{\text{NP}}$	77.59 +/- 0.3	54.78 +/- 2.2	87.78 +/- 0.4	83.21 +/- 0.6	72.78 +/- 1.6	57.22 +/- 1.2	63.35 +/- 0.4 / 63.63 +/- 0.2	2.37 +/- 3.2
$\mathcal{M}_{\text{UG}}$	66.94 +/- 9.2	53.70 +/- 1.0	85.57 +/- 0.1	83.17 +/- 1.5	70.57 +/- 0.7	58.59 +/- 0.3	71.93 +/- 0.2 / 71.33 +/- 0.5	0.92 +/- 2.1
$\mathcal{M}_{\text{RI}}$	62.17 +/- 0.4	52.97 +/- 0.2	81.53 +/- 0.2	82.0 +/- 0.7	70.32 +/- 1.5	56.62 +/- 0.0	65.70 +/- 0.2 / 65.75 +/- 0.3	8.06 +/- 1.6

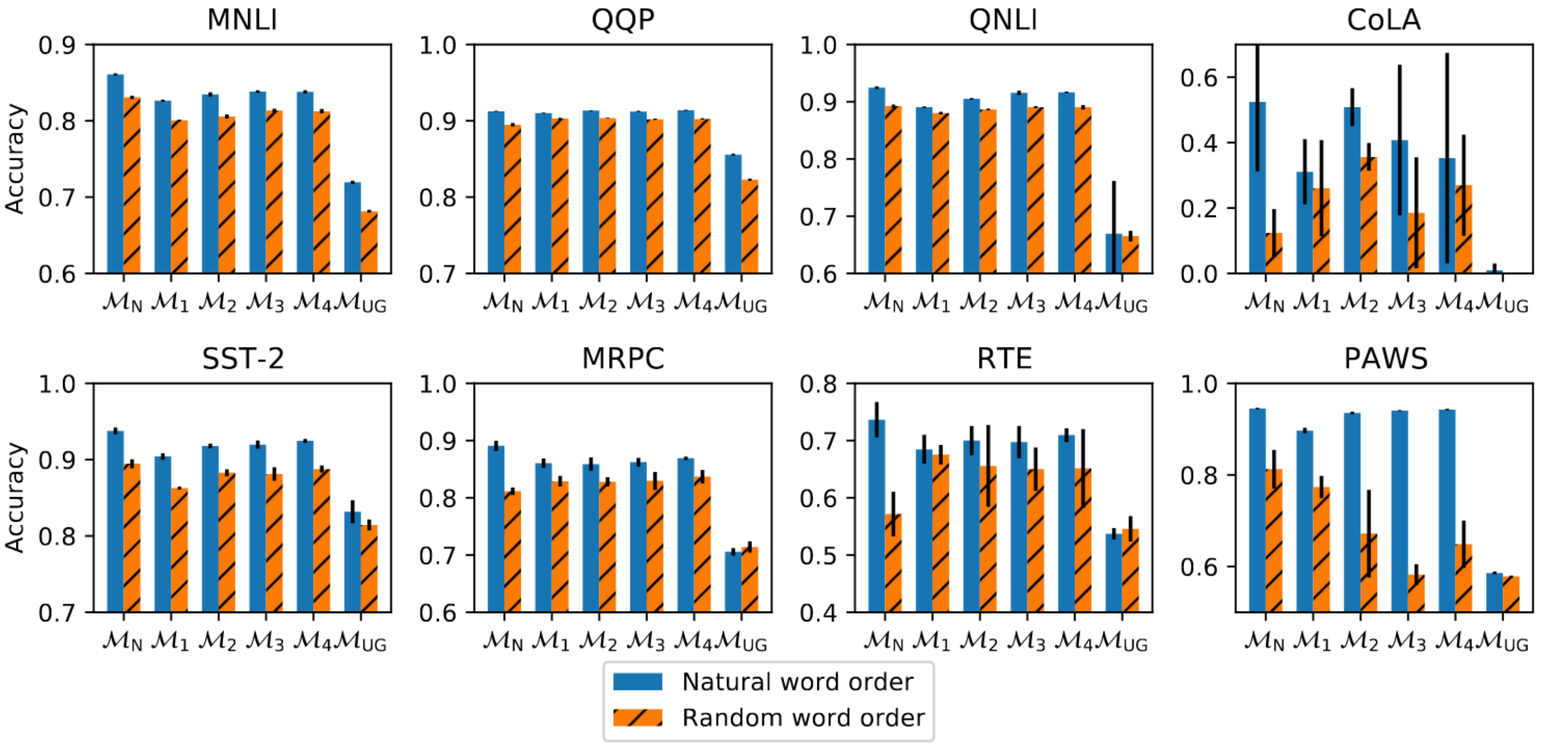
Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]

Adina Williams[†] **Douwe Kiela**[†]

[†] Facebook AI Research; [‡] McGill University / Mila - Quebec AI

{koustuvs,adinawilliams,dkiela}@fb.com



Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]

Adina Williams[†] **Douwe Kiela**[†]

[†] Facebook AI Research; [‡] McGill University / Mila - Quebec AI

{koustuvs,adinawilliams,dkiela}@fb.com

Model	UD EWT		PTB	
	MLP	Linear	MLP	Linear
$\mathcal{M}_{\text{N}}$	80.41 +/- 0.85	66.26 +/- 1.59	86.99 +/- 1.49	66.47 +/- 2.77
$\mathcal{M}_4$	78.04 +/- 2.06	65.61 +/- 1.99	85.62 +/- 1.09	66.49 +/- 2.02
$\mathcal{M}_3$	77.80 +/- 3.09	64.89 +/- 2.63	85.89 +/- 1.01	66.11 +/- 1.68
$\mathcal{M}_2$	78.22 +/- 0.88	64.96 +/- 2.32	84.72 +/- 0.55	64.69 +/- 2.50
$\mathcal{M}_1$	69.26 +/- 6.00	56.24 +/- 5.05	79.43 +/- 0.96	57.20 +/- 2.76
$\mathcal{M}_{\text{UG}}$	74.15 +/- 0.93	65.69 +/- 7.35	80.07 +/- 0.79	57.28 +/- 1.42

Table 2: Unlabeled Attachment Score (UAS) (mean and std) on the dependency parsing task (DEP) on two datasets, UD EWT and PTB, using the Pareto Probing framework (Pimentel et al., 2020a).

Is Supervised Syntactic Parsing Beneficial for Language Understanding Tasks? An Empirical Investigation

Goran Glavaš
University of Mannheim
Data and Web Science Group
goran@informatik.uni-mannheim.de

Ivan Vulić
University of Cambridge
Language Technology Lab
iv250@cam.ac.uk

(RQ) *Is explicit structural language information, provided in the form of a widely adopted syntactic formalism (Universal Dependencies, UD) (Nivre et al., 2016) and injected in a supervised manner into LM-pretrained transformers beneficial for transformers' downstream LU performance?*

Transf.	Parsing FT	NLI	HANS	PAWS	SIQA
BERT	None	84.1	53.3	**92.4**	**60.7**
	Standard	**84.4**	**56.7**	91.9	58.8
	Adapter	84.1	53.3	**92.4**	58.3
RoBERTa	None	**88.4**	**67.4**	94.7	67.2
	Standard	87.7	64.5	**94.9**	66.5
	Adapter	87.9	66.3	94.7	**67.3**

Table 2: Downstream LU performance of monolingual EN transformers (BERT and RoBERTa). **None**: no IPT; **Standard**: IPT via standard fine-tuning; **Adapter**: IPT via adapter-based fine-tuning.

		EN (EWT)		DE (GSD)		FR (GSD)		TR (IMST)		ZH (GSD)	
Transformer	**Fine-tune**	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
BERT	Standard	91.9	89.3	–	–	–	–	–	–	–	–
	Adapter	90.1	87.3	–	–	–	–	–	–	–	–
RoBERTa	Standard	93.0	**90.5**	–	–	–	–	–	–	–	–
	Adapter	91.5	88.7	–	–	–	–	–	–	–	–
mBERT	Standard	91.5	88.9	76.3	72.0	94.1	91.3	75.5	67.5	87.0	**83.8**
	Adapter	89.6	86.8	75.1	70.1	92.8	89.7	66.4	57.8	81.0	77.4
XLM-R	Standard	**93.1**	**90.5**	**89.4**	**85.0**	**94.3**	**91.7**	**77.9**	**70.0**	79.0	75.6
	Adapter	91.4	88.6	88.3	83.8	93.1	90.3	72.1	64.1	73.8	70.3
Baseline: UDify (mBERT, Standard)		91.0	88.5	87.8	83.6	93.6	91.5	74.6	67.4	**87.9**	**83.8**

Table 1: Dependency parsing performance of our transformer-based biaffine parsers.

		XNLI				PAWS-X			XCOPA	
Transformer	**Parse FT**	DE	FR	TR	ZH	DE	FR	ZH	TR	ZH
mBERT	None	71.0	73.7	**63.0**	70.3	85.1	86.3	76.4	52.0	61.2
	Standard	71.4	72.9	61.5	**70.4**	85.4	86.9	**79.8**	**57.4**	**65.4**
	Adapter	**71.7**	**74.8**	62.5	70.2	**85.8**	**87.1**	78.7	50.4	61.6
XLM-R	None	77.1	**78.1**	73.4	73.8	**88.3**	**89.3**	**81.4**	**61.2**	66.4
	Standard	76.1	77.2	73.1	73.8	86.4	89.2	81.1	59.2	**67.4**
	Adapter	**77.8**	76.4	**73.9**	**74.7**	86.7	88.7	80.7	57.4	65.6

Table 3: Performance of multilingual transformers, mBERT and XLM-R, in zero-shot language transfer for downstream LU tasks, with and without prior intermediate dependency parsing training on target language treebanks.

Syntactic Structure Distillation Pretraining for Bidirectional Encoders

Adhiguna Kuncoro[*♠♢] **Lingpeng Kong**[*♠] **Daniel Fried**[*♣]
Dani Yogatama[♠] **Laura Rimell**[♠] **Chris Dyer**[♠] **Phil Blunsom**[♠♢]
[♠]DeepMind, London, UK
[♢]Department of Computer Science, University of Oxford, UK
[♣]Computer Science Division, University of California, Berkeley, CA, USA
{akuncoro,lingpenk,dyogatama,laurarimell,cdyer,pblunsom}@google.com
dfried@cs.berkeley.edu

Shallow Syntax in Deep Water

Swabha Swayamdipta[♠*] **Matthew Peters**[♣]
Brendan Roof[♣] **Chris Dyer**[♡] **Noah A. Smith**[♢♣]

[♠]Language Technologies Institute, Carnegie Mellon University
[♣]Allen Institute for Artificial Intelligence
[♢]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Google DeepMind
{swabhas,matthewp,brendanr}@allenai.org

(deep breath)

… so where does all of this leave us?

Have language models *learned* “syntax”?

Does syntax even matter for NLU tasks?

Does syntax even matter for NLU?

Does syntax even matter?

Some Considerations:

1. Coding properties are not “syntax”

1. Coding properties are not “syntax”

a model’s sensitivity to coding properties does not indicate that it has “learned syntax”

English coding properties are not “syntax”

Can LSTM Learn to Capture Agreement? The Case of Basque

Shauli Ravfogel[1] and **Francis M. Tyers**[2,3] and **Yoav Goldberg**[1,4]
[1] Computer Science Department, Bar Ilan University
[2] School of Linguistics, Higher School of Economics
[3] Department of Linguistics, Indiana University
[4] Allen Institute for Artificial Intelligence
{shauli.ravfogel, yoav.goldberg}@gmail.com, ftyers@prompsit.com

(1) *Kutxazain-ek bezeroa-ri*
cashier-PL.ERG customer-SG.DAT
liburu-ak eman dizkiote
book-PL.ABS gave they-them-to-her/him
The cashiers gave the books to the customer.

(2) *Kutxazain-ak hemen daude*
cashier-PL-ABS here they are-PL.ABS3
The cashiers are here.

(3) *Pertson-ak zuhaitz-ak*
person-SG.ERG tree-PL.ABS
ikusten ditu
he/she-sees-them
The person sees the trees.

(4) *Zuhaitz-ak pertson-ak*
tree-SG.ERG person-PL.ABS
ikusten ditu
seeing it-is-them
The tree sees the people.

Condition	**Ergative** A / R	**Absolutive** A / R	**Dative** A / R
Base	87.1 / 80.0	93.8 / 100	98.0 / 54.9
Suffixes only	69.0 / 40.3	83.7 / 100	97.0 / 26.0
No suffixes	83.8 / 80.0	87.8 / 100	97.3 / 34.7
Neutralized case	86.0 / 79.3	93.3 / 100	97.3 / 38.1
Single verb	90.6 / 89.0	96.04 / 100	98.9 / 74.7
No *-ak*	90.9 / 81.1	96.6 / 100	98.6 / 67.7
Sing. verb, no *-ak*	92.6 / 83.4	97.2 / 100	99.1 / 75.4

Table 2: Summary of verb number prediction results for accuracy (A) and recall (R).

1. Coding properties are not “syntax”

a model’s sensitivity to coding properties does not indicate that it has “learned syntax”

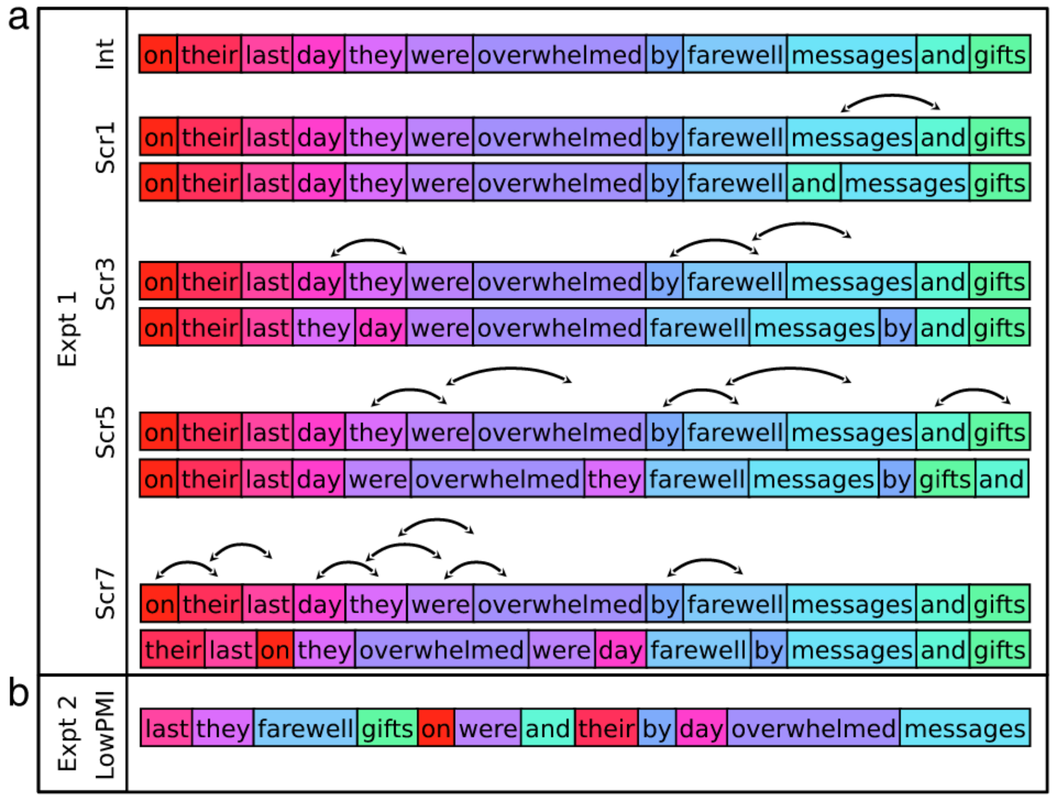
English coding properties are not “syntax”

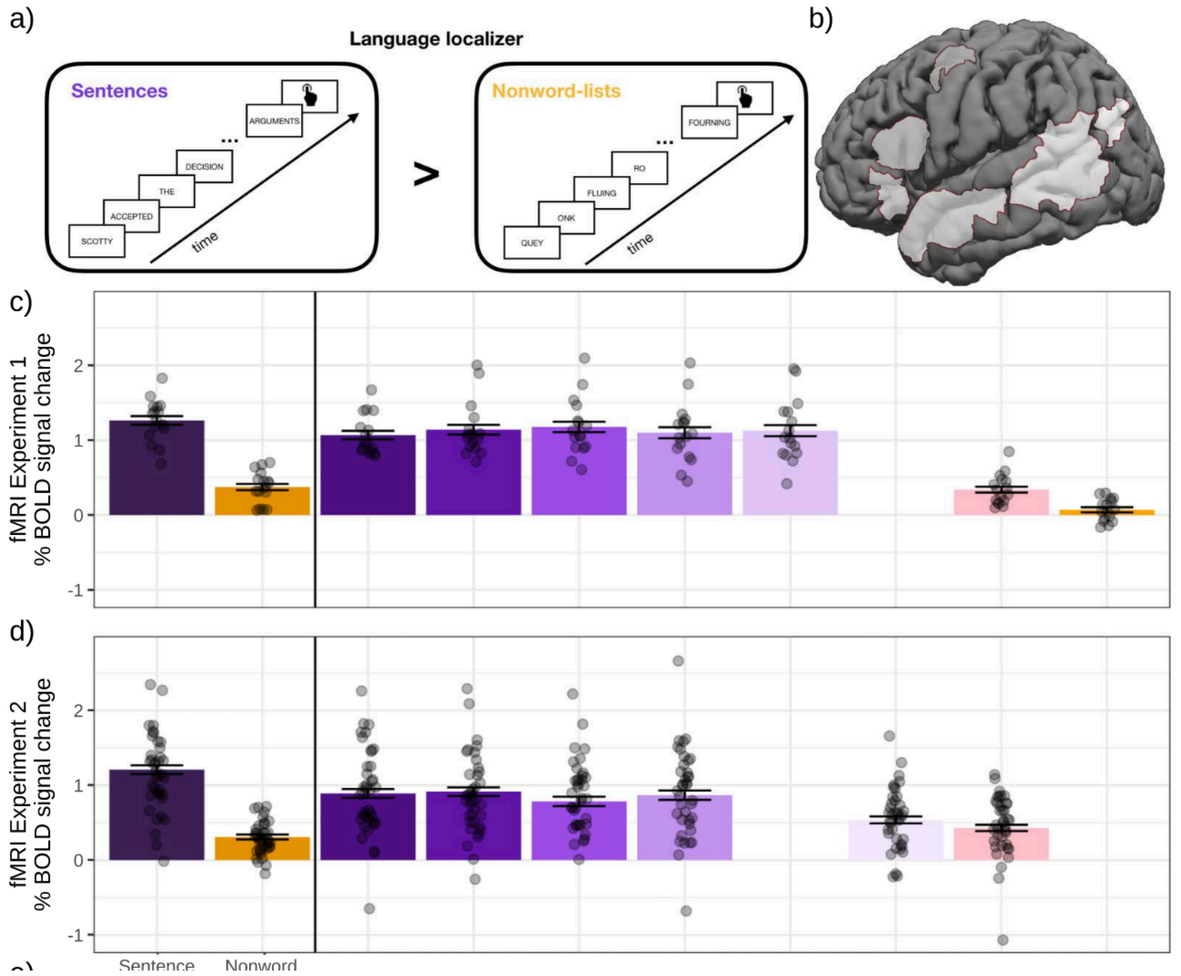
a model’s *insensitivity* to coding properties (downstream) does not mean it is devoid of syntactic knowledge

humans are surprisingly robust to word order permutations

Composition is the Core Driver of the Language-selective Network

Francis Mollica[1*], Matthew Siegelman[2*], Evgeniia Diachek[3], Steven T. Piantadosi[4], Zachary Mineroff[5], Richard Futrell[6], Hope Kean[7], Peng Qian[7], and Evelina Fedorenko[7,8,9]





Word Order Does Matter
(And Shuffled Language Models Know It)

***Vinit Ravishankar**[†] ***Mostafa Abdou**[‡] **Artur Kulmizev**[§] **Anders Søgaard**[‡]

[†]Language Technology Group, Department of Informatics, University of Oslo
[‡]Department of Computer Science, University of Copenhagen
[§]Department of Linguistics and Philology, Uppsala University

[†]vinitr@ifi.uio.no
[‡]{abdou,soegaard}@di.ku.dk

Some Considerations:

1. Coding properties are not “syntax”

2. Syntactic representations are not linguistic data

2. Syntactic representations are not linguistic data

need to clarify if studies assume specific syntactic representations

how do models fare with different representations for the same stimuli?

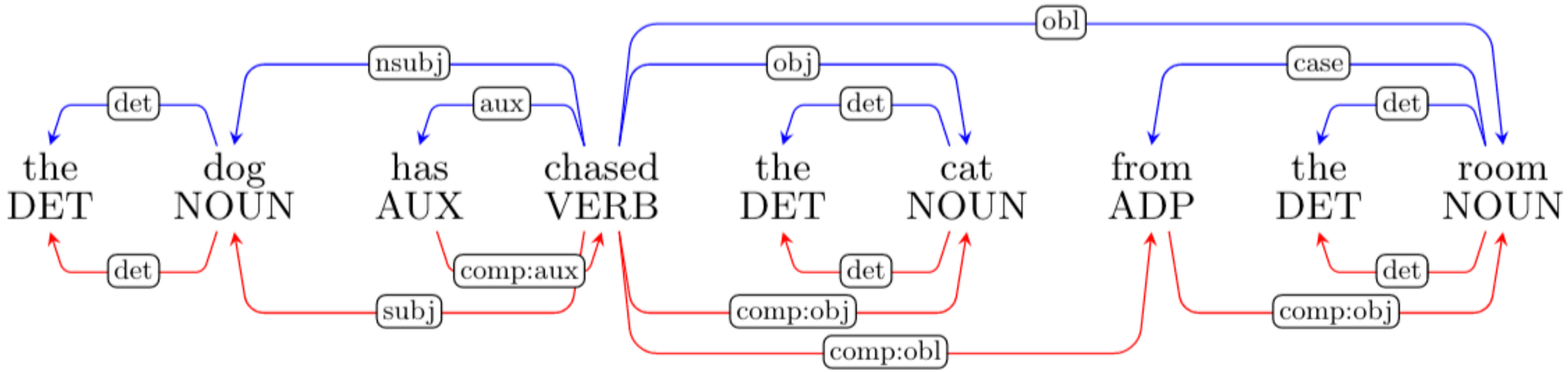
Do Neural Language Models Show Preferences for Syntactic Formalisms?

Artur Kulmizev
Uppsala University
artur.kulmizev@lingfil.uu.se

Vinit Ravishankar
University of Oslo
vinitr@ifi.uio.no

Mostafa Abdou
University of Copenhagen
abdou@di.ku.dk

Joakim Nivre
Uppsala University
joakim.nivre@lingfil.uu.se



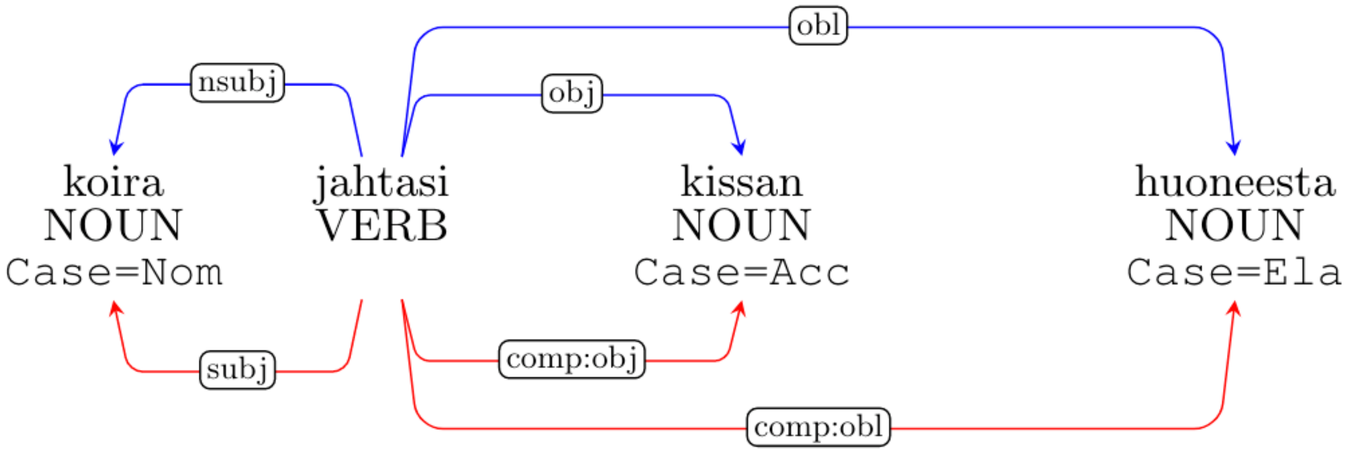
Do Neural Language Models Show Preferences for Syntactic Formalisms?

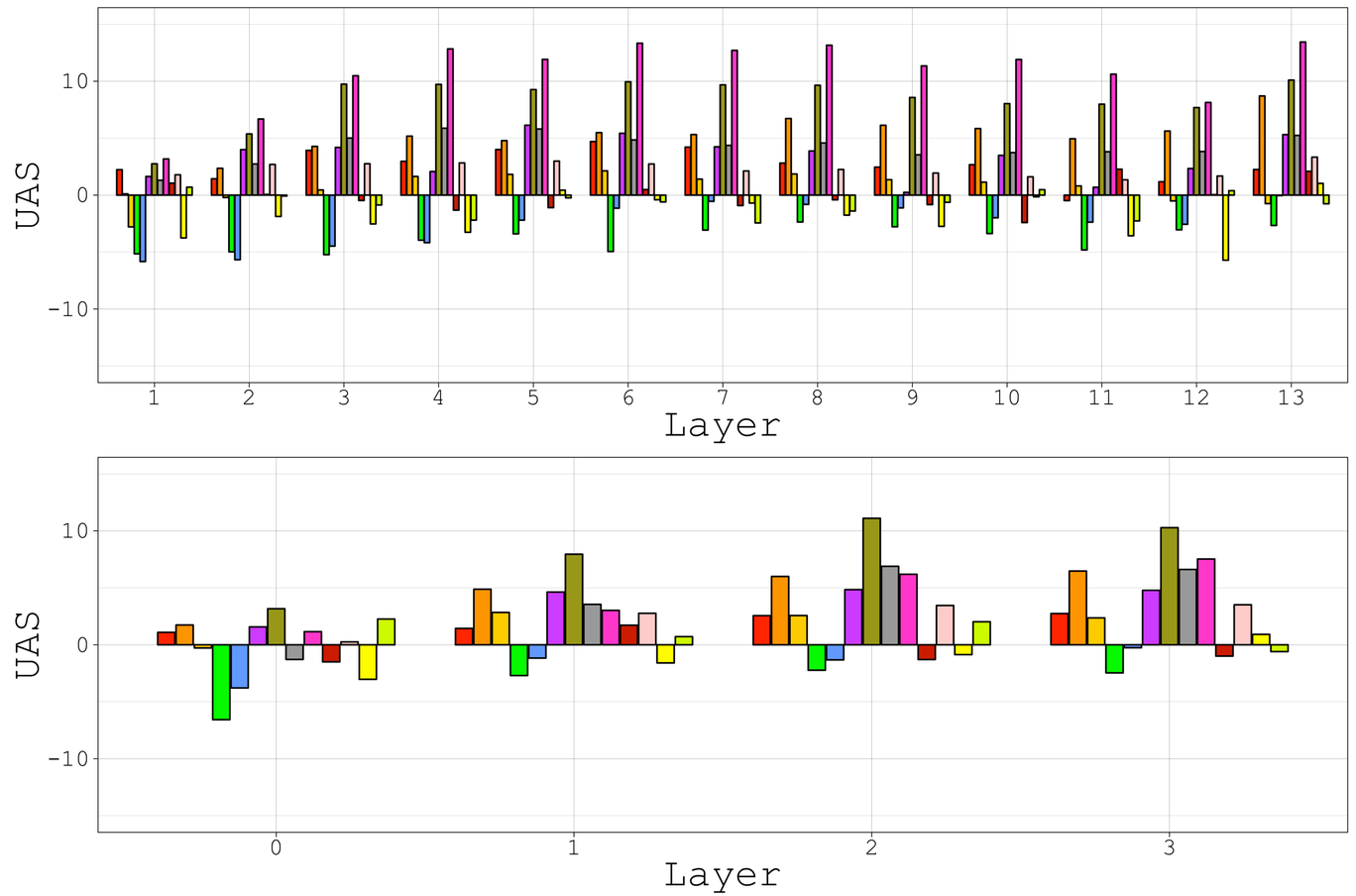
Artur Kulmizev
Uppsala University
artur.kulmizev@lingfil.uu.se

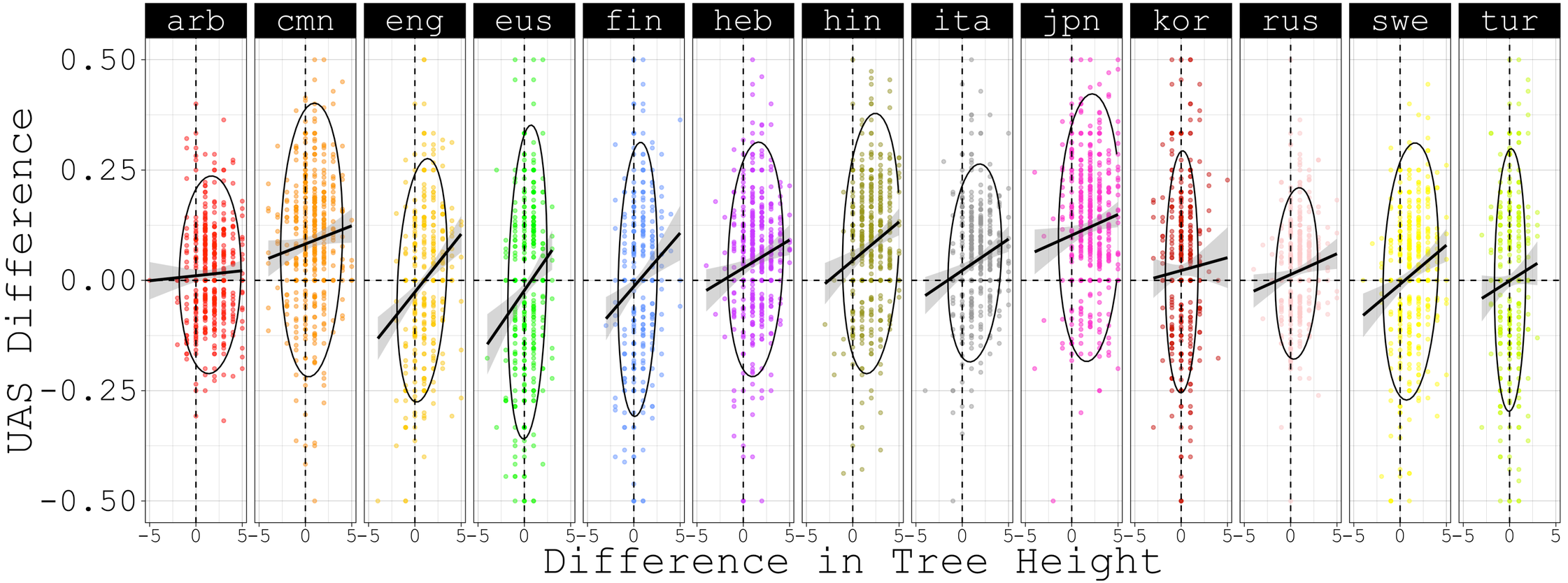
Vinit Ravishankar
University of Oslo
vinitr@ifi.uio.no

Mostafa Abdou
University of Copenhagen
abdou@di.ku.dk

Joakim Nivre
Uppsala University
joakim.nivre@lingfil.uu.se




UAS
10
0
-10
1
2
3
4
5
6
7
8
9
10
11
12
13
Layer
UAS
10
0
-10
0
1
2
3
Layer
lang
arb
cmn
eng
eus
fin
heb
hin
ita
jpn
kor
rus
swe
tur


arb
cmn
eng
eus
fin
heb
hin
ita
jpn
kor
rus
swe
tur
UAS Difference
0.50
0.25
0.00
−0.25
−0.50
−5
0
5
Difference in Tree Height

2. Syntactic representations are not linguistic data

need to clarify if studies assume specific syntactic representations

how do models fare with different representations for the same stimuli?

how does choice of representation affect “structure injection” methods?

what privileges one method over another, w.r.t. conclusions about “syntax”?

Infusing Finetuning with Semantic Dependencies

Zhaofeng Wu[♠] **Hao Peng**[♠] **Noah A. Smith**[♠♢]

[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♢]Allen Institute for Artificial Intelligence
{zfw7,hapeng,nasmith}@cs.washington.edu

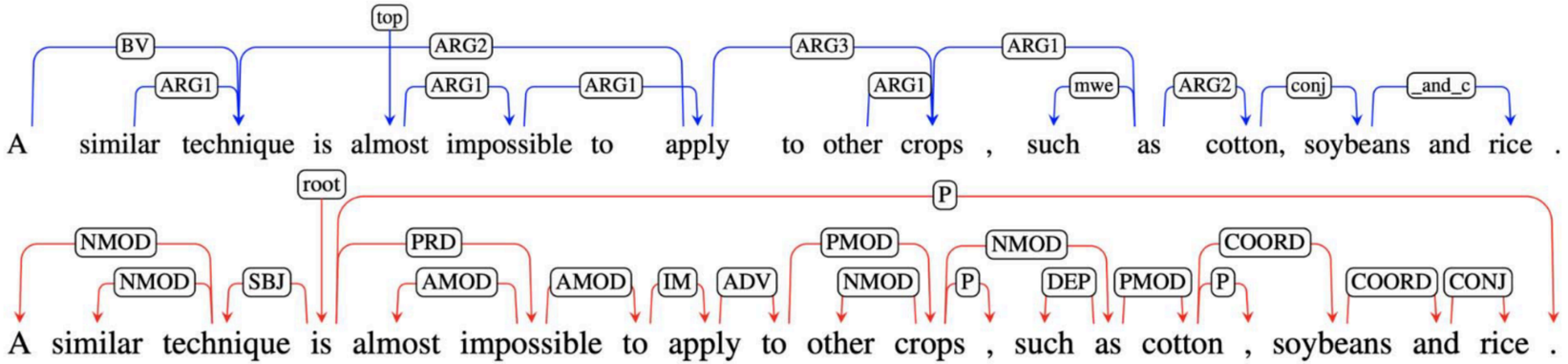


Figure 1: An example sentence in the DM (top, blue) and Stanford Dependencies (bottom, red) format, taken from Oepen et al. (2015) and Ivanova et al. (2012).

Infusing Finetuning with Semantic Dependencies

Zhaofeng Wu[♠] **Hao Peng**[♠] **Noah A. Smith**[♠♢]

[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♢]Allen Institute for Artificial Intelligence
{zfw7,hapeng,nasmith}@cs.washington.edu

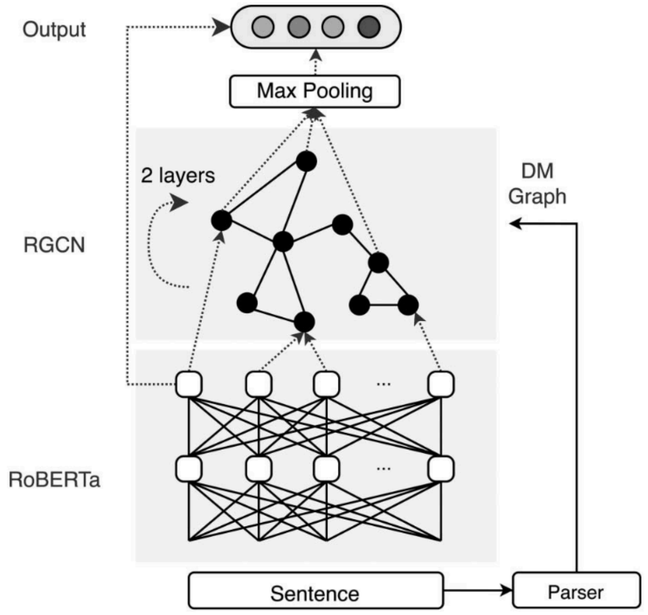


Figure 2: SIFT architecture. The sentence is first contextualized using RoBERTa, and then parsed. RGCN encodes the graph structures on top of RoBERTa. We max-pool over the RGCN's outputs for onward computation.

Models	CoLA	MRPC	RTE	SST-2	STS-B	QNLI	QQP	MNLI ID.	MNLI OOD.	Avg.
RoBERTa	$63.1_{\pm 0.9}$	$90.1_{\pm 0.8}$	$79.0_{\pm 1.6}$	$94.6_{\pm 0.3}$	$91.0_{\pm 0.0}$	$93.0_{\pm 0.3}$	$91.8_{\pm 0.1}$	$87.7_{\pm 0.2}$	$87.3_{\pm 0.3}$	86.4
SIFT	$\mathbf{64.8}_{\pm 0.4}$	$90.5_{\pm 0.7}$	$81.0_{\pm 1.4}$	$95.1_{\pm 0.4}$	$\mathbf{91.3}_{\pm 0.1}$	$93.2_{\pm 0.2}$	$91.9_{\pm 0.1}$	$87.9_{\pm 0.2}$	$\mathbf{87.7}_{\pm 0.1}$	87.0
SIFT-Light	$64.1_{\pm 1.3}$	$90.3_{\pm 0.5}$	$80.6_{\pm 1.4}$	$94.7_{\pm 0.1}$	$\mathbf{91.2}_{\pm 0.1}$	$92.8_{\pm 0.3}$	$91.7_{\pm 0.0}$	$87.7_{\pm 0.1}$	$87.6_{\pm 0.1}$	86.7
Syntax	$63.5_{\pm 0.6}$	$90.4_{\pm 0.5}$	$80.9_{\pm 1.0}$	$94.7_{\pm 0.5}$	$91.1_{\pm 0.2}$	$92.8_{\pm 0.2}$	$91.8_{\pm 0.0}$	$87.9_{\pm 0.1}$	$\mathbf{87.7}_{\pm 0.1}$	86.7

(a) Base.

Models	CoLA	MRPC	RTE	SST-2	STS-B	QNLI	QQP	MNLI ID.	MNLI OOD.	Avg.
RoBERTa	$68.0_{\pm 0.6}$	$90.1_{\pm 0.8}$	$85.1_{\pm 1.0}$	$96.1_{\pm 0.3}$	$92.3_{\pm 0.2}$	$94.5_{\pm 0.2}$	$91.9_{\pm 0.1}$	$90.3_{\pm 0.1}$	$89.8_{\pm 0.3}$	88.7
SIFT	$\mathbf{69.7}_{\pm 0.5}$	$\mathbf{91.3}_{\pm 0.4}$	$\mathbf{87.0}_{\pm 1.1}$	$96.3_{\pm 0.3}$	$\mathbf{92.6}_{\pm 0.0}$	$94.7_{\pm 0.1}$	$\mathbf{92.1}_{\pm 0.1}$	$90.4_{\pm 0.1}$	$90.1_{\pm 0.1}$	89.3
Syntax	$69.6_{\pm 1.2}$	$91.0_{\pm 0.5}$	$86.0_{\pm 1.6}$	$95.9_{\pm 0.3}$	$92.4_{\pm 0.1}$	$94.6_{\pm 0.1}$	$\mathbf{92.0}_{\pm 0.0}$	$90.4_{\pm 0.3}$	$90.0_{\pm 0.2}$	89.1

(b) Large.

Infusing Finetuning with Semantic Dependencies

Zhaofeng Wu[♠] **Hao Peng**[♠] **Noah A. Smith**[♠♢]

[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♢]Allen Institute for Artificial Intelligence
{zfw7,hapeng,nasmith}@cs.washington.edu

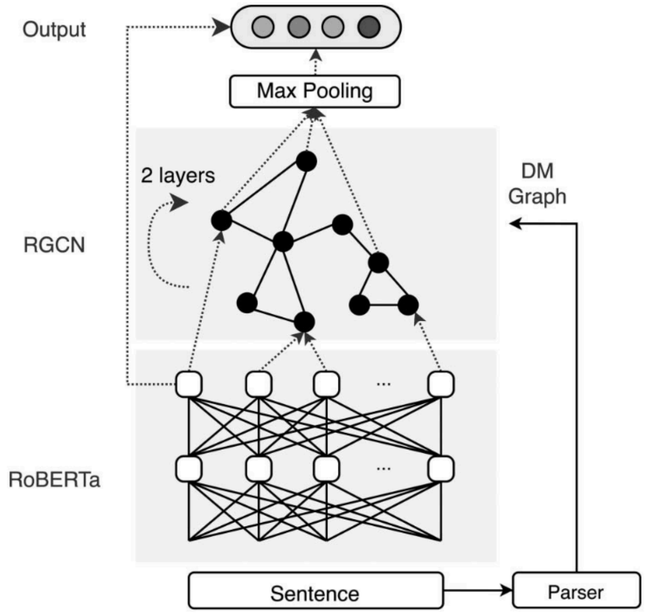


Figure 2: SIFT architecture. The sentence is first contextualized using RoBERTa, and then parsed. RGCN encodes the graph structures on top of RoBERTa. We max-pool over the RGCN's outputs for onward computation.

								MNLI		
Models	**CoLA**	**MRPC**	**RTE**	**SST-2**	**STS-B**	**QNLI**	**QQP**	**ID.**	**OOD.**	**Avg.**
RoBERTa	63.1	90.1	79.0	94.6	91.0	93.0	91.8	87.7	87.3	86.4
GCN	**65.2**	90.2	80.2	94.8	91.1	92.9	91.8	87.8	**87.7**	86.8
GAT	63.4	90.0	79.4	94.7	91.2	92.9	91.8	87.7	87.6	86.5
Hidden	64.2	90.2	79.7	94.5	91.0	92.8	91.8	87.1	86.7	86.4
Scaffold	62.5	**90.5**	71.1	94.3	91.0	92.6	91.7	87.7	87.6	85.5
SIFT	64.8	**90.5**	**81.0**	**95.1**	**91.3**	**93.2**	**91.9**	**87.9**	**87.7**	**87.0**
SIFT-Light	64.1	90.3	80.6	94.7	91.2	92.8	91.7	87.7	87.6	86.7

Table 7: GLUE development set results for different architectures for incorporating semantic information. The settings and metrics are identical to Table 3a. All models use the base size variant.

Some Considerations:

1. Coding properties are not “syntax”

2. Syntactic representations are not linguistic data

3. Theory, model, and task

3. Theory, model, and task

studies should be concerned with evaluating syntactic phenomena / representations directly

absent this, need to have a concrete understanding of how models behave

Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement

Nina Poerner, Benjamin Roth & Hinrich Schütze
Center for In
p

Evaluating Recurrent Neural Network Explanations

Leila Arras[1], **Ahmed Osman**[1], **Klaus-Robert Müller**[2,3,4], **and Wojciech Samek**[1]
[1]Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[2]Machine Lear
[3]Department of B
[4]Max Pl
{leila.a

A Diagnostic Study of Explainability Techniques for Text Classification

Pepa Atanasova Jakob Grue Simonsen Christina Lioma Isabelle Augenstein
De
{pepa, simonse

Aligning Faithful Interpretations with their Social Attribution

Alon Jacovi
Bar Ilan University
alonjacovi@gmail.com

Yoav Goldberg
Bar Ilan University and
Allen Institute for AI
yoav.goldberg@gmail.com

3. Theory, model, and task

studies should be concerned with evaluating syntactic phenomena / representations directly

absent this, need to have a concrete understanding of how models behave

how much emphasis should we put on the **model** over the **task**?

how much emphasis should we put on the **model** over the **dataset**?

UnNatural Language Inference

Koustuv Sinha[1,2,3], **Prasanna Parthasarathi**[1,2], **Joelle Pineau**[1,2,3] **and Adina Williams**[3]
[1] School of Computer Science, McGill University, Canada
[2] Montreal Institute of Learning Algorithms (Mila), Canada
[3] Facebook AI Research (FAIR)
{koustuv.sinha, prasanna.parthasarathi, jpineau, adinawilliams}
@{mail.mcgill.ca, mail.mcgill.ca, cs.mcgill.ca, fb.com}

Coupled with the finding that humans cannot perform UNLI at all well, the high rate of permutation acceptance that we observe leads us to conclude that current models do not yet "know syntax" in the fully systematic and humanlike way we would like them to.

Out of Order: How Important Is The Sequential Order of Words in a Sentence in Natural Language Understanding Tasks?

Thang M. Pham[1] thangpham@auburn.edu
Trung Bui[2] bui@adobe.com
Long Mai[2] malong@adobe.com
Anh Nguyen[1] anh.ng8@gmail.com
[1]Auburn University [2]Adobe Research

Despite their superhuman scores, most GLUE-trained models behave similarly to Bag-of-Words (BOW) models, which are prone to naive mistakes (Fig. 1b–d). Our results also suggest that GLUE does not necessarily require syntactic information or complex reasoning.

Some Considerations:

1. Coding properties are not “syntax”

2. Syntactic representations are not linguistic data

3. Theory, model, and task

4. What are the research questions?

4. What are the research questions?

I. To what degree *does* M learn A when trained on D to perform T?

II. To what degree *can* M learn A when trained on D to perform T?

III. To what degree does M *need* learn A when trained on D to perform T?

4. What are the research questions?

I. To what degree *does* M learn A when trained on D to perform T?

* straightforward to answer, given reliable method to measure the degree to which M learns A with respect to D and T

4. What are the research questions?

II. To what degree *can* M learn A when trained on D to perform T?

* indirectly answers questions of type I
* modal in nature
* fundamental asymmetry between positive and negative findings
* positive results can establish that something is *possible*
* negative results are inconclusive

4. What are the research questions?

III. To what degree does M *need* learn A when trained on D to perform T?

* causality: does learning A improve performance of M on T?
* modality: is learning A necessary to achieve better performance?

Is Supervised Syntactic Parsing Beneficial for Language Understanding Tasks? An Empirical Investigation

Goran Glavaš
University of Mannheim
Data and Web Science Group
goran@informatik.uni-mannheim.de

Ivan Vulić
University of Cambridge
Language Technology Lab
iv250@cam.ac.uk

6 Conclusion

We thoroughly examined the effects of leveraging formalized syntactic structures (UD) in state-of-the-art neural language models (e.g., RoBERTa, XLM-R) for downstream language understanding (LU) tasks, both in monolingual and language transfer settings. The key results, obtained through intermediate parsing training (IPT) based on a state-of-the-art-level dependency parser, indicate that explicit syntax, at least in our extensive experiments, provides negligible impact on LU tasks.

Can we conclude that explicit syntax is not beneficial for language understanding tasks, without showing that:

a) the fine-tuned model actually learned some aspects of syntax?

b) does this knowledge causally affect the model's performance downstream?

Some Considerations:

1. Coding properties are not “syntax”
2. Syntactic representations are not linguistic data
3. Theory, model, and task
4. What are the research questions?
5. Aggregate Metrics may be misleading, but are necessary

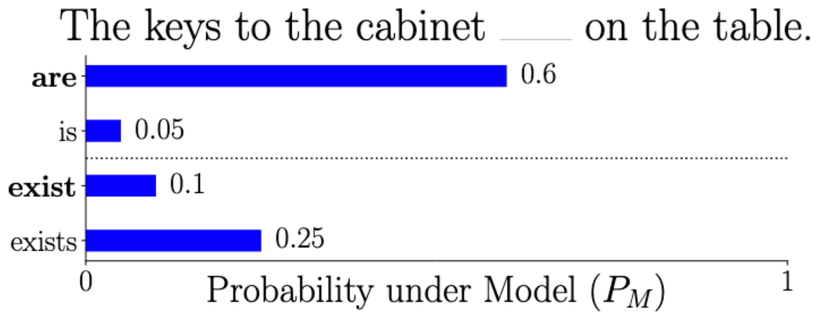
5. Aggregate Metrics may be misleading, but are necessary

what do aggregate metrics really represent in terms of syntactic knowledge? when is the glass half full or half empty?

aggregate metrics obfuscate important variation in data; over-represent frequent phenomena

Refining Targeted Syntactic Evaluation of Language Models

Benjamin Newman **Kai-Siang Ang** **Julia Gong** **John Hewitt**
Department of Computer Science
Stanford University
{blnewman,kaiang,jxgong,johnhew}@cs.stanford.edu



Metric	Computation	Score
TSE	**are** $>$ is	1.0
EW (systematicity)	**are** $>$ is exists $>$ **exist**	0.5
MW (likely behavior)	$\frac{\textbf{are} + \textbf{exist}}{\textbf{are} + \textbf{exist} + \text{is} + \text{exists}}$	0.7

Templates	BERT cased			BERT uncased			RoBERTa			GPT2		
	MW	EW	TSE	MW	EW	TSE	MW	EW	TSE	MW	EW	TSE
Simple	0.99	0.94	**1.00**	0.98	0.90	**1.00**	0.98	0.93	**1.00**	0.90	0.86	**1.00**
In a sentential complement	0.92	0.67	**0.89**	0.92	0.60	**0.86**	0.92	0.67	**0.88**	0.96	0.65	**0.89**
VP coordination	0.91	0.89	**0.90**	0.93	**0.90**	**0.90**	0.93	0.90	**0.93**	0.89	0.87	**0.97**
Across prepositional phrase	0.91	0.83	**0.93**	0.83	0.75	**0.85**	0.87	0.83	**0.89**	0.84	0.76	**0.96**
Across subject relative clause	0.87	**0.84**	**0.84**	0.88	0.84	**0.85**	0.76	0.72	**0.80**	0.82	0.77	**0.97**
Across object relative clause	0.91	0.88	**0.91**	0.86	0.80	**0.85**	0.88	0.85	**0.91**	0.95	0.89	**0.99**
Across object relative (no that)	0.92	0.88	**0.90**	0.79	0.72	**0.81**	0.86	0.82	**0.89**	0.95	0.89	**0.99**
In object relative clause	0.93	0.95	**0.97**	0.95	0.97	**0.99**	0.89	0.91	**0.97**	0.91	0.88	**0.98**
In object relative (no that)	0.90	0.91	**0.92**	0.81	**0.82**	**0.82**	0.82	0.83	**0.90**	0.91	0.88	**0.97**
BLiMP	0.81	0.73	**0.90**	0.78	0.69	**0.85**	0.70	0.66	**0.78**	0.82	0.75	**0.91**

Table 2: MW, EW, and TSE evaluations on various models and syntactic constructions (See Warstadt et al. (2020); Marvin and Linzen (2018) for descriptions). MW is colored differently because its score is based directly on the model's probability mass, while EW and TSE are based on 0/1 judgements, so they are not directly comparable.

5. Aggregate Metrics may be misleading, but are necessary

what do aggregate metrics really represent in terms of syntactic knowledge? when is the glass half full or half empty?

aggregate metrics obfuscate important variation in data; over-represent frequent phenomena

syntactic benchmarks cannot enumerate every single phenomena

should all (collected) phenomena have a uniform weight, in aggregate?

Conclusions:

measuring “syntactic knowledge” (of humans or machines) is difficult

varying methodologies, hypotheses, and conclusions make the big picture fuzzy

exercising clarity and caution will lead to a more nuanced and focused research agenda

pursuing these questions further can shed valuable insights on the nature of “syntax”

Thank you!