Methods to Evaluate Verb Argument Structure Knowledge in Language Models and Embeddings

James v. Bruno, David Yi, Jiayu Han, Peter Zukerman

Probing **Linguistic Knowledge** in Neural Networks

Motivations:

- Does neural networks encode enough grammatical information?
- If so, what extent do the features learned by neural networks resemble the linguistic competence of humans?

Why do we want to probe?

- Would be helpful in downstream tasks
- Analyses of results can contribute to the scientific questions in linguistics: the role of prior grammatical bias in human language acquisition.

Roadmap

- Verb Argument Structure Alternations
- Probing Linguistics Knowledge of verbs in Embeddings
- Probing Linguistics Knowledge in Pretrained Language Models
- Our work
- Q & A

Verb Argument Structure Alternations

Levin (1993) comprehensively describes many classes of Verb Argument Structure Alternations.

1 example (out of many): The Spray/Load Alternation

1) a. Lucy sprayed the wall with paint.
   b. Lucy sprayed paint on the wall.

2) a. Lucy covered the towel with sand.
   b. *Lucy covered sand on the towel

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press

The Spray-Load Alternation (Arad, 2006)

(1) a. Lucy sprayed the paint on the wall.
    b. Lucy sprayed the wall with paint.

(2) a. Ben loaded hay on the truck.
    b. Ben loaded his truck with hay.

(3) a.

Jan	plant	bomen	in	de	tuin.
John	plants	trees	in	the	Garden.

b.

Jan	beplant	de	tuin	met	bomen.
John	be-plants	the	garden	with	trees.

(4) a.

kabe	ni	penki	o	nuru
wall	on	paint	ACC	Paint (VERB)

‘smear paint on the wall’

b.

kabe	o	penki	de	nuru
wall	ACC	paint	with	Paint (VERB)

‘smear the wall with paint’

(5) a.

János	rámázolta	a	festéket	a	falra.
John	onto-smeared-he-it	ACC	Paint-ACC	the	wall-onto

‘John smeared paint on the wall.’

b.

János	bemázolta	a	falat	festékkel.
John	in-smeared-he-it	ACC	The wall	Paint-with

‘John smeared the wall with paint.’

Arad, Maya. (2006). The Spray-Load Alternation. In The Blackwell Companion to Syntax (pp. 466–478). Blackwell Publishing.

Argument-structure based account (Levin and Rappaport, 1988)

(8) LOAD: <Agent, Locatum, Goal>

Lexical Entry

(9) a. LOAD: $x <y, P_{loc}z>$
b. LOAD: $x <y, P_{with}z>$

Linking Rules

(10) LOAD_a: <Agent, Locatum, Goal> (locative variant)
LOAD_b: <Agent, Theme, Locatum> (*with*-variant)

Or just two different entries altogether?

(14) a. LOAD: [*x* cause [*y* to come to be at *z*] / LOAD]
b. LOAD: [[*x* cause [*z* to come to be in a STATE]] BY MEANS OF [*x* cause [*y* to come to be at *z*]] / LOAD]

How about a richer semantic representation?

The Aspectual Interface Hypothesis (Tenny, 1987)

- The direct object “measures out the event”.
  - E.g. for *eat an apple*, the eating event is over when the apple is consumed.
- “Load verbs denote an event that can be measured out in two different ways – both by the Theme and by the Goal. Since measuring out is associated with direct objects, either the Theme or the Goal may be realized as direct objects”

- Lucy sprayed the wall with paint.
  - The spraying event is measured out according to the status of the wall.

- Lucy sprayed paint on the wall.
  - The spraying event is measured out according to the status of the paint.

The Aspectual Interface Hypothesis (Tenny, 1987)

- Lucy covered the towel with sand.
    - The covering event is measured out according to the status of the towel.

- *Lucy covered sand on the towel
    - The covering event cannot be measured out according to the status of the sand. The towel is either covered, or it's not.

What are the Lexical Properties of Spray-Load Verbs? (Pinker, 1989)

Ingredient 1: In general, the location has be construeable as undergoing a change of state.

3) a. Lucy sprayed the wall with paint.
   b. Lucy sprayed paint on the wall.

4) a. Lucy covered the towel with sand.
   b. *Lucy covered sand on the towel

What are the Lexical Properties of Spray-Load Verbs? (Pinker, 1989)

Ingredient 2: Content-Oriented vs. Container-Oriented

- Content-Oriented verbs obligatorily take a locatum, with an optional location.

(20) a. Lucy piled the books (on the shelf).
b. Lucy piled the shelf *(with books).

- Container-Oriented verbs obligatorily take a location, with an optional locatum.

(21) a. Lucy stuffed the turkey (with breadcrumbs).
b. Lucy stuffed the breadcrumbs *(into the turkey).

What are the Lexical Properties of Spray-Load Verbs? (Pinker, 1989)

Those ingredients allow us to say:

- Container-oriented verbs that alternate must specify not only the change of state in the container, but also the manner in which the substance is moved into the location
    - *Lucy covered sand on the towel
    - Lucy stuffed breadcrumbs into the turkey.
- Content-oriented verbs that alternate must specify not only the manner in which the substance is moved, but also the change of state in the location
    - Lucy piled the shelf with books.
    - *Lucy poured the glass with water.

So what’s really important here, in the context of 575?

- Verb argument structure alternations are a lexical property of the verb.
- Verb argument structure alternations are identifiable by the kinds of tokens in the neighborhood of the verb. *“You shall know a word by the company it keeps”* (Firth, 1957).

- That suggests that the alternation class may be encoded in embeddings.

Verb Argument Structure Alternations in Word and Sentence Embeddings

Kann et al. (2019)

Verb Alternation Classes

Verb Frame	Example Sentences		
Caus. Inch.	Jessica **dropped** the vase. The vase **dropped**.	Jessica blew the bubble. *The bubble blew.	
Dative-Prep. Dative-2-Obj.	Liz **gave** a gift to the boy. Liz **gave** the boy a gift.	Liz administered a test to the kid. *Liz administered the kid a test.	*Liz charged $50 to Jon. Liz charged Jon $50.
Spr.-Lo.-*with* Spr.-Lo.-Loc.	Sue **loaded** the truck with wood. Sue **loaded** wood onto the truck.	Sue coated the deck with paint. *Sue coated paint on the deck.	*Sue swept the bin with sand. Sue swept sand into the bin.
no-*there* *there*	Fear **remained** in my mind. There **remained** fear in my mind.	A girl focused on the quiz. *There focused on the quiz a girl.	
U.-Obj.-Refl. U.-Obj.-No-Refl.	Ada **clapped** her hands. Ada **clapped.**	Ada permed her hair. *Ada permed.	*Ada exercised herself. Ada exercised.

*Important Note: The sentences are formed in such a way that only the **main verb** alternation information determines grammaticality judgements.

LaVA Dataset

The LaVA (Lexical Verb-Frame Alternations) dataset includes 515 verbs annotated for membership in 10 verb frame classes.

Human annotations note 1 for membership, 0 for non-membership, and 'x' where membership is unknown (or non-existent).

verb	sl	sl_noloc	sl_nowith	inch	non_inch	there	non_there	dat_both	dative_to	dat_do	refl_op	refl_only
fed	0	0	0	0	1	0	1	1	0	0	0	0
served	0	0	0	0	x	0	x	1	0	0	0	0
gave	0	0	0	0	x	0	x	1	0	0	0	0
left	0	0	0	0	x	0	x	1	0	0	0	0

LaVA Dataset

The LaVA corpus presents 5 of the largest syntactic verb frame alternations provided by Levin (1993):

- Causative-Inchoative
- Dative
- Spray-Load (as seen earlier)
- *there*-Insertion
- Understood-Object

On Sparsity:
Due to how verb argument structure alternations function, negative samples can not always be obtained. For example, no English verbs can appear in the inchoative but not the causative. There are also no verbs that can only appear in the *there* frame but not the no-*there*. This leads to sparsity in annotations, which causes trivial word-level classifications.

Experiment 1: From Word Embeddings to Argument Structures

Objective

- For each alternation class, build a multi-label classifier that predicts whether a verb participates in a particular syntactic frame

$$p(s) = \boldsymbol{\sigma}(W_2(f(W_1x))$$

Modeling Details

- **Input** (x): Word embedding representation of verb *v*
- **Alternation Class**: causative-inchoative
- **Syntactic frame** (s): Inchoative
- **Output** p(s): Probability that verb *v* participates in frame *s*
- **Training**: Single-layer MLP with 4-fold Cross Validation

LaVA Dataset

[LaVA Dataset (Kann et al. 2019)](#)

Levin class	Caus.–Inch.		Dative		Spray–Load		*there*-Insertion		Understood-Object	
	Inch.	Caus.	Prep.	2-Obj.	*with*	Loc.	no-*there*	*there*	Refl.	No-Refl.
Positive	70	120	63	72	90	81	50	145	11	81
Negative	140	(0)	356	405	220	229	185	(0)	466	396
Total	210	120	419	477	310	310	235	145	477	477

Why does Causative (NEG) have 0 examples?

The vase **dropped** (inchoative) /Jessica **dropped** the vase (causative)
* The bubble **blew** (inchoative) / Jessica **blew** the bubble (causative)

Where do the word embeddings come from?

Word Embeddings

- **GloVe** Embeddings: 300d embeddings trained on 6B Tokens
- **Custom** Embeddings: Trained on **100M** tokens from the British National Corpus (BNC) using a single-directional LSTM w/ LM Objective

Why these Embeddings?

- Trained on similar amount of data that humans are exposed to during language acquisition
- Large pretrained models (i.e. BERT) trained on "several orders of magnitude more data than humans see in a lifetime" than custom embeddings
    - 3.3B tokens v.s. 100M

Evaluation: Matthew's Correlation Coefficient

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Why MCC?

- Special case of Pearson's Correlation Coefficient for Binary Classification
- Generalizes better to imbalanced distributions than accuracy/F1-score

-1: Complete disagreement between predictions and observations
0: Average score of two unrelated distributions
1: Perfect correlation between predictions and observations

Results

[LaVA Results (Kann et al. 2019)]

		CAUSATIVE–INCHOATIVE		DATIVE		SPRAY–LOAD		*there*-INSERTION		UNDERSTOOD-OBJECT	
		Inch.	(Caus.)	Prep.	2-Obj.	*with*	Loc.	no-*there*	(*there*)	Refl.	Non-Refl.
CoLA: Majority BL	Acc.	66.7	(100.0)	85.0	84.9	71.0	73.9	78.7	(100.0)	97.7	83.0
CoLA: MLP	MCC	**0.555**	0.0	0.32	**0.482**	**0.645**	0.253	**0.459**	0.0	0.0	0.219
	Acc.	81.0	(100.0)	86.6	88.3	85.8	72.9	84.3	(100.0)	97.7	79.0
GloVe: Majority BL	Acc.	66.8	(100.0)	85.0	85.3	71.0	74.6	79.1	(100.0)	97.6	81.5
GloVe: MLP	MCC	**0.672**	0.0	0.0	0.0	**0.585**	0.145	**0.536**	0.0	0.0	0.3
	Acc.	85.5	(100.0)	85.0	85.3	83.9	73.4	85.8	(100.0)	97.6	73.2

FAVA

The FAVA (Frame and Alternations of Verbs Acceptability) dataset consists of ~10,000 sentences containing the verbs in LaVA in different verb frames and labeled for grammaticality.

Annotations are 1 for accepted and 0 for unaccepted sentences.

dat	0		christopher tipped a week 's salary to james .
dat	1		christopher tipped james a week 's salary .
dat	0		jason tipped 20 pounds to rebecca .
dat	1		jason tipped rebecca 20 pounds .
inch	1		rebecca steered the car .
inch	1		the car steered .
inch	1		rebecca steered the bicycle .
inch	1		the bicycle steered .
inch	1		james steered the truck .

Detour: CoLA Dataset(Corpus of Linguistic Acceptability)

Label	Sentence	Source
*	The more books I ask to whom he will give, the more he reads.	Culicover and Jackendoff (1999)
✓	I said that my father, he was tight as a hoot-owl.	Ross (1967)
✓	The jeweller inscribed the ring with the name.	Levin (1993)
*	many evidence was provided.	Kim and Sells (2008)

Data: 10,657 sentences labeled for grammatical acceptability that analyze different types of linguistic phenomena

- 17 **in-domain**, 6 **out-of-domain**

A few examples:

- Comparatives (Culicover and Jackendoff, 1999)
- Islands (Ross, 1967)
- Verb Alternations (Levin, 1993)
- General syntax (Kim and Sells, 2008)

Warstadt et al. 2019

Experiment 2: Sentence Embedding Probing

- Linguists would classify a word by interrogating whether sentences with a given verb and frame are acceptable.
- Analogously, a MLP model is used to calculate the probability that a sentence S is acceptable.

$$p(S) = \boldsymbol{\sigma}(W_2(\tanh(W_1 x))$$

Where do these Sentence Embeddings come from?

Sentence Encoder trained by Warstadt et al. (2018) on “Real/Fake” discrimination task for downstream CoLA task

Input: ELMo-style Word Embeddings

Training (12M sentences)

- Real: 6M sentences from BNC(British National Corpus)
- Fake: 3M generated by LSTM + 3M permuted from BNC

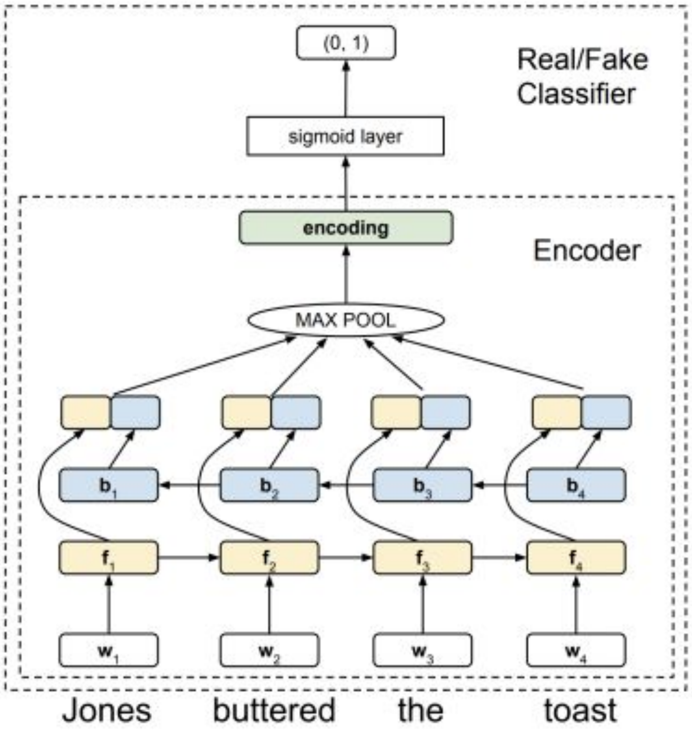

Figure 1: Real/fake model. w_i = word embeddings, f_i = forward LSTM hidden state, b_i = backward LSTM hidden state. Figure from Warstadt et al. (2018).

Warstadt et al. (2019)

Results

		Comb.	CAUSATIVE–INCHOATIVE	DATIVE	SPRAY–LOAD	*there*-INSERTION	UNDERSTOOD-OBJECT
w/o CoLA	MCC	0.290	**0.603**	0.413	0.323	**0.528**	**0.753**
	Acc.	64.6	85.4	76.0	66.2	72.9	87.4
w/ CoLA	MCC	0.361	**0.464**	0.329	0.261	**0.523**	**0.638**
	Acc.	68.7	81.2	59.0	63.4	72.5	81.8
Majority BL	MCC	0.0	0.0	0.0	0.0	0.0	0.0
	Acc.	66.6	77.6	82.1	60.3	77.5	53.7

1. Easiest alternation was **UNDERSTOOD-OBJECT** alternation
Blink -> her eyes, Clap -> his hands, etc..

2.Alterations involving **only transitive** verb frames (i.e. SPRAY-LOAD) were generally **more difficult** than those with **at least one intransitive frames** (i.e.CAUSATIVE-INCHOATIVE)

3.**No relationship** between **# training examples** and performance

4. CoLA will help when in ‘comb’ situation

Verb Frame	Example Sentences		
Caus. Inch.	Jessica **dropped** the vase. The vase **dropped**.	Jessica blew the bubble. *The bubble blew.	
Dative-Prep. Dative-2-Obj.	Liz **gave** a gift to the boy. Liz **gave** the boy a gift.	Liz administered a test to the kid. *Liz administered the kid a test.	*Liz charged $50 to Jon. Liz charged Jon $50.
Spr.-Lo.-*with* Spr.-Lo.-Loc.	Sue **loaded** the truck with wood. Sue **loaded** wood onto the truck.	Sue coated the deck with paint. *Sue coated paint on the deck.	*Sue swept the bin with sand. Sue swept sand into the bin.
no-*there* *there*	Fear **remained** in my mind. There **remained** fear in my mind.	A girl focused on the quiz. *There focused on the quiz a girl.	
U.-Obj.-Refl. U.-Obj.-No-Refl.	Ada **clapped** her hands. Ada **clapped.**	Ada permed her hair. *Ada permed.	*Ada exercised herself. Ada exercised.

Takeaways

- Models achieved moderate correlation (0.5-0.7) in **5/12** acceptability experiments, all except one achieved > 0.3

- Easiest alternation was **UNDERSTOOD-OBJECT** alternation
    - *Blink -> her eyes*, *Clap -> his hands*, etc..

- Alterations involving only ***transitive*** verb frames (i.e. SPRAY-LOAD) were generally **more difficult** than those with at least one ***intransitive*** frames (i.e. CAUSATIVE-INCHOATIVE)

- **No relationship** between **# training examples** and performance

Takeaways

- Pros:
  - word-level and sentence-level datasets: LaVA, FAVA
  - Probing word embeddings
  - Probing sentence embeddings
- Cons:
  - FAVA is not from natural sentences
  - Not easy to tell linguistic knowledge is in neural networks or in the probing models

BliMP: The Benchmark of Linguistic Minimal Pairs for English

Warstadt et al. (2020)

BliMP: The **B**enchmark of **Li**nguistic **M**inimal **P**airs

Motivations:

- Existed evaluating datasets only focus on a small set of linguistic phenomena
- Probing by additional models cannot tell whether the linguistic knowledge is in the Neural Networks

Minimal Pairs + New probing paradigm

Minimal Pairs: Pairs of minimally different sentences that contrast in grammatical acceptability and isolate specific phenomenon in syntax, morphology, or semantics.

a. The cats annoy Tim. (grammatical)

b. * The cats annoys Tim. (ungrammatical)

New probing paradigm: probing LMs without an additional supervised model

- Observe whether LMs assign a higher probability to the acceptable sentence in each minimal pairs

BliMP Dataset Overview

- 12 linguistic phenomenon categories, 67 individual datasets(different linguistic paradigms), each containing 1000 minimal pairs.

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Table 1: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category.

Warstadt et al. (2020)

BliMP Dataset Overview

Phenomenon	UID	5-gram	LSTM	TXL	GPT-2	Human	Acceptable Example	Unacceptable Example	1pfx	2pfx
ANAPHOR AGREEMENT	anaphor_gender_agreement	44	88	91	99	96	Katherine can't help **herself**.	Katherine can't help **himself**.	✓	
	anaphor_number_agreement	52	95	97	100	99	Many teenagers were helping **themselves**.	Many teenagers were helping **herself**.	✓	
ARGUMENT STRUCTURE	animate_subject_passive	54	68	58	77	98	Amanda was respected by some **waitresses**.	Amanda was respected by some **picture**.	✓	
	animate_subject_trans	72	79	70	80	87	Danielle **visited** Irene.	The eye **visited** Irene.		✓
	causative	51	65	54	68	82	Aaron breaks the glass.	Aaron appeared the glass.		
	drop_argument	68	79	67	84	90	The Lutherans couldn't skate around.	The Lutherans couldn't disagree with.		
	inchoative	89	72	81	90	95	A screen was fading.	A screen was cleaning.		
	intransitive	82	73	81	90	86	Some glaciers are vaporizing.	Some glaciers are scaring.		
	passive_1	71	65	76	89	99	Jeffrey's sons are insulted by Tina's supervisor.	Jeffrey's sons are smiled by Tina's supervisor.		
	passive_2	70	72	74	79	86	Most cashiers are disliked.	Most cashiers are flirted.		
	transitive	91	87	89	49	87	A lot of actresses' nieces have toured **that** art gallery.	A lot of actresses' nieces have coped **that** art gallery.		✓
BINDING	principle_A_c_command	58	59	61	100	86	A lot of actresses that thought about Alice healed **themselves**.	A lot of actresses that thought about Alice healed **herself**.	✓	
	principle_A_case_1	100	100	100	96	98	Tara thinks that **she** sounded like Wayne.	Tara thinks that **herself** sounded like Wayne.	✓	
	principle_A_case_2	49	87	95	73	96	Stacy imagines herself **praising** this actress.	Stacy imagines herself **praises** this actress.	✓	
	principle_A_domain_1	95	98	99	99	95	Carlos said that Lori helped **him**.	Carlos said that Lori helped **himself**.	✓	
	principle_A_domain_2	56	68	70	73	75	Mark imagines Erin might admire **herself**.	Mark imagines Erin might admire **himself**.	✓	
	principle_A_domain_3	52	55	60	82	83	Nancy could say every guy hides **himself**.	Every guy could say Nancy hides **himself**.		✓
	principle_A_reconstruction	40	46	38	37	78	It's herself who Karen criticized.	It's herself who criticized Karen.		
CONTROL/ RAISING	existential_there_object_raising	84	66	76	92	90	William has declared **there** to be no guests getting fired.	William has obliged **there** to be no guests getting fired.		✓
	existential_there_subject_raising	77	80	79	89	88	There was bound to be a fish escaping.	There was unable to be a fish escaping.		
	expletive_it_object_raising	72	63	72	58	86	Regina wanted it to be **obvious** that Maria thought about Anna.	Regina forced it to be **obvious** that Maria thought about Anna.		✓
	tough_vs_raising_1	33	34	45	72	75	Julia wasn't fun to talk to.	Julia wasn't unlikely to talk to.		
	tough_vs_raising_2	77	93	86	92	81	Rachel was apt to talk to Alicia.	Rachel was exciting to talk to Alicia.		
DETER-MINER-NOUN AGR.	determiner_noun_agreement_1	88	92	92	100	96	Craig explored that **grocery store**.	Craig explored that **grocery stores**.	✓	
	determiner_noun_agreement_2	86	92	81	93	95	Carl cures those **horses**.	Carl cures that **horses**.		✓
	determiner_noun_agreement_irregular_1	85	82	88	94	92	Phillip was lifting this **mouse**.	Phillip was lifting this **mice**.	✓	
	determiner_noun_agreement_irregular_2	90	86	82	93	85	Those ladies walk through those **oases**.	Those ladies walk through that **oases**.		✓
	determiner_noun_agreement_with_adj_1	50	86	78	90	96	Tracy praises those lucky **guys**.	Tracy praises those lucky **guys**.	✓	
	determiner_noun_agreement_with_adj_2	53	76	81	96	94	Some actors buy these gray **books**.	Some actors buy this gray **books**.		✓
	determiner_noun_agreement_with_adj_irregular_1	55	83	77	88	85	This person shouldn't criticize this upset **child**.	This person shouldn't criticize this upset **children**.	✓	
	determiner_noun_agreement_with_adj_irregular_2	52	87	86	93	95	That adult has brought that purple **octopus**.	That adult has brought those purple **octopus**.		✓
ELLIPSIS	ellipsis_n_bar_1	23	68	65	88	92	Brad passed one big museum and Eva passed several.	Brad passed one museum and Eva passed several big.		
	ellipsis_n_bar_2	50	67	89	86	78	Curtis's boss discussed four sons and Andrew discussed five sick sons.	Curtis's boss discussed four happy sons and Andrew discussed five sick.		
FILLER GAP	wh_questions_object_gap	53	79	61	84	85	Joel discovered the vase that Patricia might take.	Joel discovered what Patricia might take the vase.		
	wh_questions_subject_gap	82	92	83	95	98	Cheryl thought about some dog that upset Sandra.	Cheryl thought about who some dog upset Sandra.		
	wh_questions_subject_gap_long_distance	86	96	86	88	85	Bruce knows that person that Dawn likes that argued about a lot of guys.	Bruce knows who that person that Dawn likes argued about a lot of guys.		
	wh_vs_that_no_gap	83	97	86	97	97	Danielle finds out that many organizations have alarmed Chad.	Danielle finds out who many organizations have alarmed Chad.		
	wh_vs_that_no_gap_long_distance	81	97	91	94	92	Christina forgot that all plays that win worry Dana.	Christina forgot who all plays that win worry Dana.		
	wh_vs_that_with_gap	18	43	42	56	77	Nina has learned who most men sound like.	Nina has learned that most men sound like.		
	wh_vs_that_with_gap_long_distance	20	14	17	56	75	Martin did find out what every cashier that shouldn't drink wore.	Martin did find out that every cashier that shouldn't drink wore.		
IRREGULAR FORMS	irregular_past_participle_adjectives	79	93	91	78	99	The forgotten **newspaper article** was bad.	The forgot **newspaper article** was bad.		✓
	irregular_past_participle_verbs	80	85	66	90	95	Edward **hid** the cats.	Edward **hidden** the cats.	✓	
ISLAND EFFECTS	adjunct_island	48	67	65	91	94	Who has Colleen aggravated before kissing Judy?	Who has Colleen aggravated Judy before kissing?		
	complex_NP__island	50	47	58	72	80	Who hadn't some driver who would fire Jennifer's colleague embarrassed?	Who hadn't Jennifer's colleague embarrassed some driver who would fire?		
	coordinate_structure_constraint_complex_left_branch	32	30	36	42	90	What lights could Spain sell and **Andrea** discover?	What could Spain sell lights and **Andrea** discover?		✓
	coordinate_structure_constraint_object_extraction	59	71	74	88	91	Who will Elizabeth and Gregory cure?	Who will Elizabeth cure and Gregory?		
	left_branch_island_echo_question	96	32	63	77	91	David would cure what **snake**?	What would David cure **snake**?		✓
	left_branch_island_simple_question	57	36	36	82	99	Whose hat should Tonya wear?	Whose should Tonya wear hat?		
	sentential_subject_island	61	43	37	35	61	Who have many women's touring Spain embarrassed.	Who have many women's touring embarrassed Spain.		
	wh_island	56	47	20	77	73	What could Alan discover **he** has run around?	What could Alan discover **who** has run around?	✓	
NPI LICENSING	matrix_question_npi_licensor_present	1	2	1	67	98	Should Monica **ever** grin?	Monica should **ever** grin.		✓
	npi_present_1	47	54	61	55	83	Even these trucks have **often** slowed.	Even these trucks have **ever** slowed.	✓	
	npi_present_2	47	54	48	62	98	Many skateboards **also** roll.	Many skateboards **ever** roll.	✓	
	only_npi_licensor_present	57	93	80	100	92	Only Bill would **ever** complain.	Even Bill would **ever** complain.		✓
	only_npi_scope	30	36	45	85	72	Only those doctors who Karla respects **ever** conceal many snakes.	Those doctors who only Karla respects **ever** conceal many snakes.		✓
	sentential_negation_npi_licensor_present	93	100	99	89	93	Those banks had not **ever** lied.	Those banks had really **ever** lied.		✓
	sentential_negation_npi_scope	45	23	53	95	81	Those turtles that are boring April could not **ever** break those couches.	Those turtles that are not boring April could **ever** break those couches.		✓
QUANTIFIERS	existential_there_quantifiers_1	91	96	94	99	94	There aren't many lights darkening.	There aren't all lights darkening.		
	existential_there_quantifiers_2	62	16	14	24	76	Each book is there disturbing Margaret.	There is each book disturbing Margaret.		
	superlative_quantifiers_1	45	63	84	84	91	No man has revealed more than five forks.	No man has revealed at least five forks.		
	superlative_quantifiers_2	17	83	85	78	85	An actor arrived at at **most** six lakes.	No actor arrived at at **most** six lakes.		✓
SUBJECT-VERB AGR.	distractor_agreement_relational_noun	24	76	77	83	81	A sketch of lights **doesn't** appear.	A sketch of lights **don't** appear.	✓	
	distractor_agreement_relative_clause	22	63	60	68	86	Boys that aren't disturbing Natalie **suffer**.	Boys that aren't disturbing Natalie **suffers**.		✓
	irregular_plural_subject_verb_agreement_1	73	81	78	95	95	This goose **isn't** bothering Edward.	This goose **weren't** bothering Edward.	✓	
	irregular_plural_subject_verb_agreement_2	88	89	83	96	94	The woman **cleans** every public park.	The women **cleans** every public park.		✓
	regular_plural_subject_verb_agreement_1	76	89	73	97	95	Jeffrey **hasn't** criticized Donald.	Jeffrey **haven't** criticized Donald.	✓	
	regular_plural_subject_verb_agreement_2	81	83	85	96	95	The dress **crumples**.	The dresses **crumples**.		✓

Warstadt et al. (2020)

Data Generation

Datasets

- All minimal pairs are *artificially generated* from a vocabulary of 3,000 words, each lexical item annotated with morphological, syntactic, and semantic features

Example: Causative Frame

{
sentence_good: "Aaron breaks the glass."
sentence_bad: "Aaron appeared the glass.",
Linguistics_term (major): "argument_structure",
UID (minor): "causative"
}

Source

Comparing FAVA/CoLA and BliMP

CoLA/FAVA

- Supervised *Binary* Acceptability Judgments
- No "generally accepted method" to obtain acceptability predictions from unsupervised model -> need to use something like Logistic Reg. / MLP
- Sentences are pulled directly from wide variety of Linguistic corpora for CoLA (not the case for FAVA)

BliMP

- Unsupervised Acceptance *Probabilities* using LM objective
- Can use unsupervised LMs like GPT-2, Transformer-XL, LSTMs, etc.. directly to model probability
- Sentences are artificially generated, acceptability judgments from authors and validated through Amazon Mechanical Turk

Results

Phenomenon	UID	5-gram	LSTM	TXL	GPT-2	Human
ANAPHOR AGREEMENT	anaphor_gender_agreement	44	88	91	99	96
	anaphor_number_agreement	52	95	97	100	99
ARGUMENT STRUCTURE	animate_subject_passive	54	68	58	77	98
	animate_subject_trans	72	79	70	80	87
	causative	51	65	54	68	82
	drop_argument	68	79	67	84	90
	inchoative	89	72	81	90	95
	intransitive	82	73	81	90	86
	passive_1	71	65	76	89	99
	passive_2	70	72	74	79	86
	transitive	91	87	89	49	87

Performance on Verb Argument Structure Classes

Warstadt and Bowman (2019): "Performance is also high on sentences with marked argument structures, indicating that argument structure is relatively easy to learn"

- Analyzing the performance of BERT, GPT, etc.. on CoLA

Warstadt et al. (2020): "We note that the reported difficulty of these phenomena contradicts Warstadt and Bowman's (2019) conclusion that argument structure is one of the strongest domains for neural models."

Hypotheses:

- Supervised v.s. Unsupervised datasets
- Disproportionate amount of "Argument Structure" related sentences in CoLA

Other Interesting Takeaways

Figure 1: Heatmap showing the correlation between models' accuracies in each of the 67 paradigms.

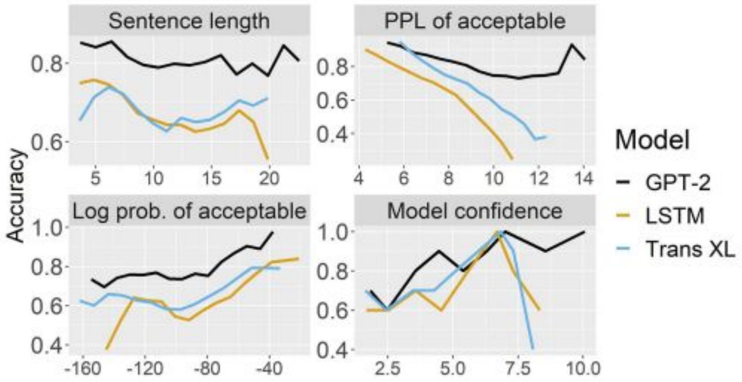

Figure 2: Models' performance on BLiMP as a function of sentence length, perplexity, log probability of the acceptable sentence, and model confidence (calculated as $|\log P(S_1) - \log P(S_2)|$).

]

Our Work

We are reproducing the experiments of Kann et al. (2019) by analyzing BERT (and other LLM) embeddings

- Without the constraints of studying “to what extent do the features learned by ANNs resemble the linguistic competence of humans”
- Probe linguistic knowledge of frozen BERT representations **without** additional finetuning for both word/sentence-level embeddings

Essential idea: Use “better” embeddings (static -> contextual) and dumb down the classifier to tackle “Probe Confounder Problem” (Hewitt and Liang, 2019)

- Classifier: Simple Logistic Regression (LR) Classifier
- Control Task: Compare between LR, MLP-1, MLP-2

Q & A