



Multilingual Syntactic/Semantic Probes

Syntactic and Semantic Probes of NLMs in Counting Tasks

Drew Barker, Lexi Loessberg-Zahl, Devin Johnson,
Denise Mak, Vincent Soesanto

Agenda

- Are All Languages Equally Hard to Language-Model? (Cotterell, et al. 2018)
- Probing for Sentence Structure in Contextualized Word Representations (Tenney, et al. 2019)
- Probing for Semanting Evidence of Composition (Ettinger, et al. 2016)
- A discussion on how these papers are relevant to our task

Are All Languages Equally Hard to Language-Model?

Ryan Cotterell¹ and **Sebastian J. Mielke¹** and **Jason Eisner¹** and **Brian Roark²**

¹ Department of Computer Science, Johns Hopkins University ² Google

{ryan.cotterell@, sjmielke@, jason@cs.}jhu.edu roark@google.com

S(ituation) T A R

- How well should we expect language models to work on languages with differing typological profiles?
- It seems unlikely that all languages are equally easy to model, or that one method of modeling will be good for all languages.

S T(task) A R

Develop an evaluation framework for fair cross-linguistic comparison of language models:

- Use **morphological counting complexity** (MCC) (Sagot, 2013) to compare the degree of morphological inflection in each language
- A fairly crude metric that counts the number of **inflectional categories** distinguished by a language

S T(ask) A R

Morphological Counting Complexity: Inflectional Categories

- tense, case, voice, aspect, person, number, gender, mood, animacy, definiteness, comparison, evidentiality, politeness, possession, etc.
 - See unimorph.org for full list and methodology. Currently covers 110 languages.

S T(ask) A R

Morphological Counting

Complexity: Inflectional Categories

Ex:

English : 6

Lithuanian : 152

data (M)		
lang	wds / ch	MCC
bg	0.71/4.3	96
cs	0.65/3.9	195
da	0.70/4.1	15
de	0.74/4.8	38
el	0.75/4.6	50
en	0.75/4.1	6
es	0.81/4.6	71
et*	0.55/3.9	110
fi*	0.52/4.2	198
fr	0.88/4.9	30
hu*	0.63/4.3	94
it	0.85/4.8	52
lt	0.59/3.9	152
lv	0.61/3.9	81
nl	0.75/4.5	26
pl	0.65/4.3	112
pt	0.89/4.8	77
ro	0.74/4.4	60
sk	0.64/3.9	40
sl	0.64/3.8	100
sv	0.66/4.1	35

S T(ask) A R

So we have a way to compare the morphological complexity of languages, but how do we evaluate the performance of language models in a way that is fair across languages?

S T(ask) A R

First attempt: Bits per character (BPC)

$$\frac{1}{|\mathbf{c}|+1} \sum_{i=1}^{|\mathbf{c}|+1} \log p(c_i \mid \mathbf{c}_{<i})$$

$|\mathbf{c}|$ = length of the utterance, in characters

$c_{|\mathbf{c}|+1}$ is a distinguished end-of-string symbol EOS.

S T(ask) A R

First attempt: Bits per character (BPC)

- BPC relies on the vagaries of individual writing systems. Consider, for example, the difference in how Czech and German express the phoneme /tʃ/:

Czech: č

German: tsch

Consider the Czech word puč and its German equivalent putsch. Even if these words are both predicted with the same probability in a given context, German will end up with a lower BPC.

S T(ask) A R

A new metric for evaluating language models:

- Bits per English character (BPEC), a fair language model evaluation metric invariant to orthographic changes, and independent of utterance length:

$$\text{BPEC} = \frac{1}{|\mathbf{c}_{\text{English}}|+1} \sum_{i=1}^{|\mathbf{c}|+1} \log p(c_i \mid \mathbf{c}_{<i})$$

$|\mathbf{c}|$ = length of the utterance, $|\mathbf{c}_{\text{English}}|$ = length of that utterance in English,
 $c_{|\mathbf{c}|+1}$ is a distinguished end-of-string symbol EOS.

S T(ask) A R

Data: Europarl multi-text corpus

- 21 languages: all Indo-European except 3 Uralic languages (Finnish, Hungarian and Estonian)
- Potential confound: The characteristics of translated language has been widely studied, indicating that translated utterances are often simpler than the original (Baker, 1993).
 - This may have caused underestimation of the BPEC for other languages, since most of the data is translated from English. However, English still has the lowest BPEC score.

S T A(ction) R

Build two open vocabulary language models: n-gram and LSTM

N-gram: hybrid word/character model

- Vocabulary is the union of unique words, unique characters in the training data, and special tokens {EOW, EOS}
- 7-gram model using standard Kneser and Ney (1995) training

S T A(ction) R

Build two open vocabulary language models: n-gram and LSTM

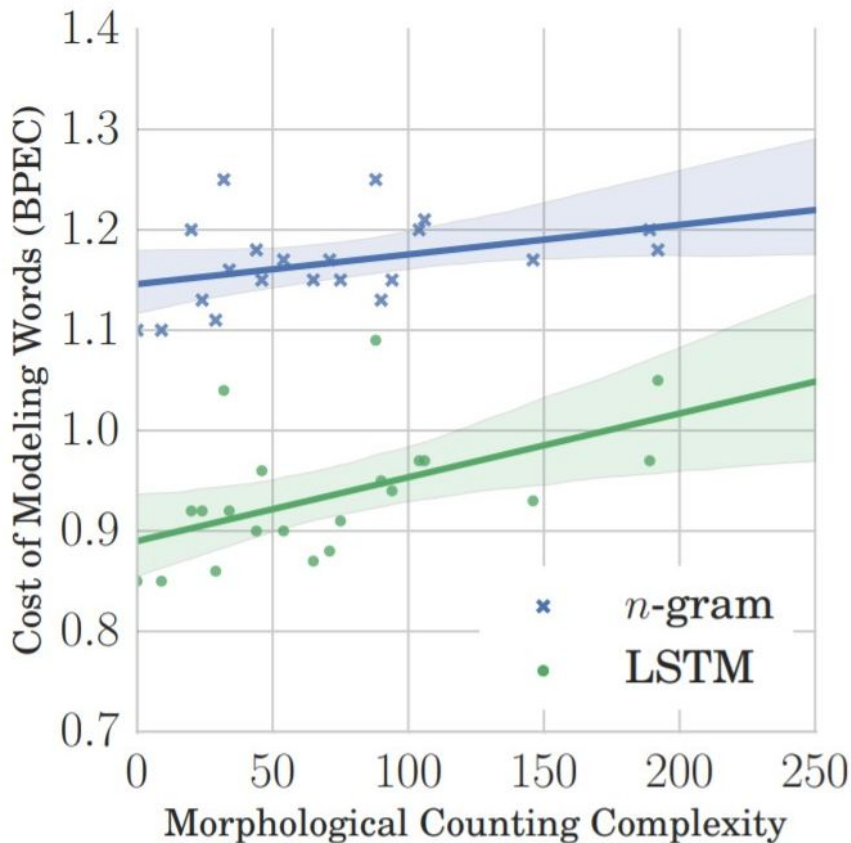
LSTM: full character level model

- Builds character level representations
- 2 hidden layers, size 1024
- Trained with SGD (Stochastic Gradient Descent)

S T A R(results)

BPEC performance of n-gram (blue) and LSTM (green) LMs over word sequences. Lower is better.

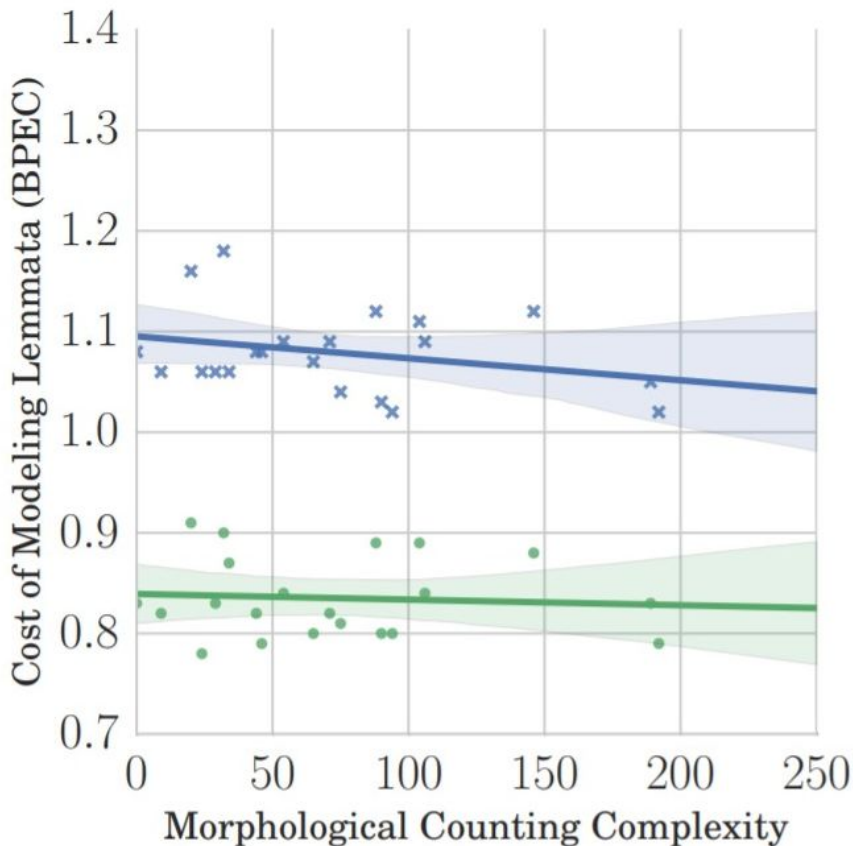
- LSTM is clearly better overall
- Languages with higher MCC are more difficult for both models
 - Spearman's rank correlation: $\rho = 0.59$, significant at $p < 0.005$



S T A R(results)

BPEC performance of n-gram (blue) and LSTM (green) LMs over *lemma* sequences. Lower is better.

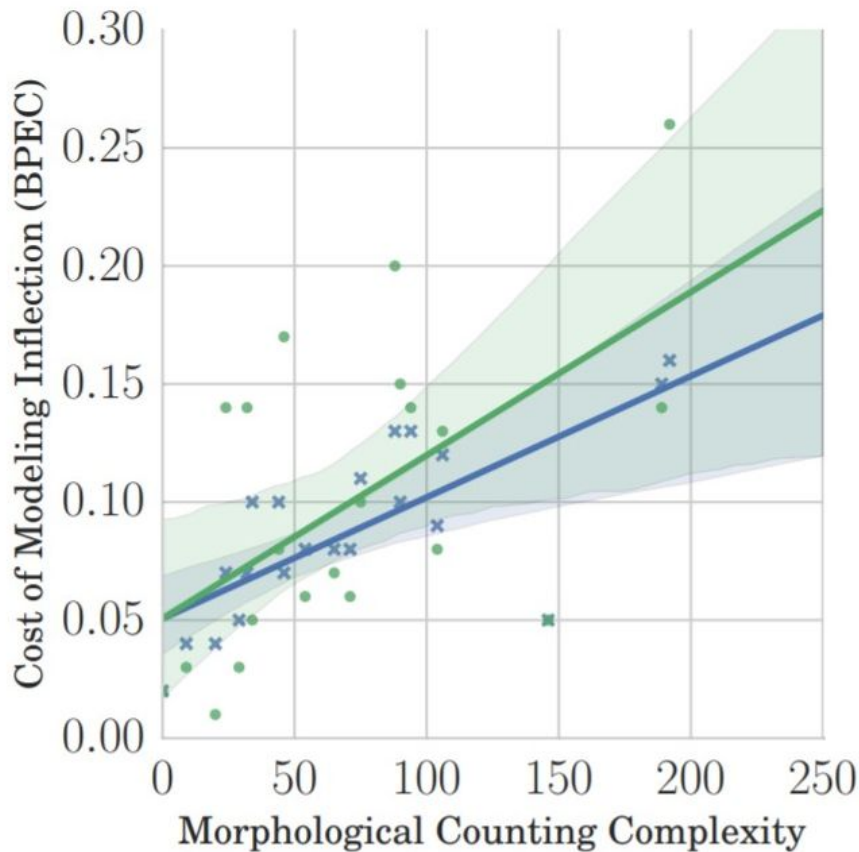
- We can see that the correlation becomes insignificant and slightly negative ($\rho = -0.13$, $p \approx 0.56$)



S T A R(results)

Difference in BPEC performance of n-gram (blue) and LSTM (green) LMs between words and lemmata.

- the LM penalty for modeling inflectional endings is greater for languages with higher counting complexity.
- Authors argue this penalty is a more appropriate measure of the complexity of the inflectional system, as compared to MCC.



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}**

¹Google AI Language, ²Johns Hopkins University, ³Swarthmore College,

⁴New York University, ⁵Brown University

S(ituation) T A R

- A need to understand where contextualized word representations improve over conventional representations
- What do contextualized representations encode that conventional representations do not?

S T(task) A R

Design a suite of probing tasks:

- Pos tagging: OntoNotes 5.0
- Constituent labeling: OntoNotes 5.0
- Dependency labeling: English Web Treebank
- Named entity labeling: OntoNotes 5.0
- Semantic role labeling: OntoNotes 5.0
- Coreference: OntoNotes 5.0, Winograd
- Semantic proto-role: SPR1 (PTB), SPR2 (English Web Treebank)
- Relation classification: SemEval 2010 Task 8 dataset

S T(ask) A R

Examples

POS	The important thing about Disney is that it is a global [brand] ₁ . → NN (Noun)
Constit.	The important thing about Disney is that it [is a global brand] ₁ . → VP (Verb Phrase)
Depend.	[Atmosphere] ₁ is always [fun] ₂ → nsubj (nominal subject)
Entities	The important thing about [Disney] ₁ is that it is a global brand. → Organization
SRL	[The important thing about Disney] ₂ [is] ₁ that it is a global brand. → Arg1 (Agent)
SPR	[It] ₁ [endorsed] ₂ the White House strategy. . . → {awareness, existed_after, . . . }
Coref. ^O	The important thing about [Disney] ₁ is that [it] ₂ is a global brand. → True
Coref. ^W	[Characters] ₂ entertain audiences because [they] ₁ want people to be happy. → True Characters entertain [audiences] ₂ because [they] ₁ want people to be happy. → False
Rel.	The [burst] ₁ has been caused by water hammer [pressure] ₂ . → Cause-Effect(e_2, e_1)

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

S T(task) A R

Models probed:

CoVe	Top-level activations of a 2-layer biLSTM trained on English-German translation, concatenated with Glove vectors
ELMo	2-layer biLSTM, over CNN character layer, trained on Billion Word Benchmark newswire text.
OpenAI GPT	12-layer Transformer encoder trained as a left-to-right language model over Toronto Books corpus
BERT	Deep transformer encoder trained jointly as a masked model and on next-sentence prediction, on concatenation of Toronto Books and English Wikipedia. 12-layer (base) and 24-layer models (large) are probed

S T A(ction) R

Classifier: MLP labels the spans

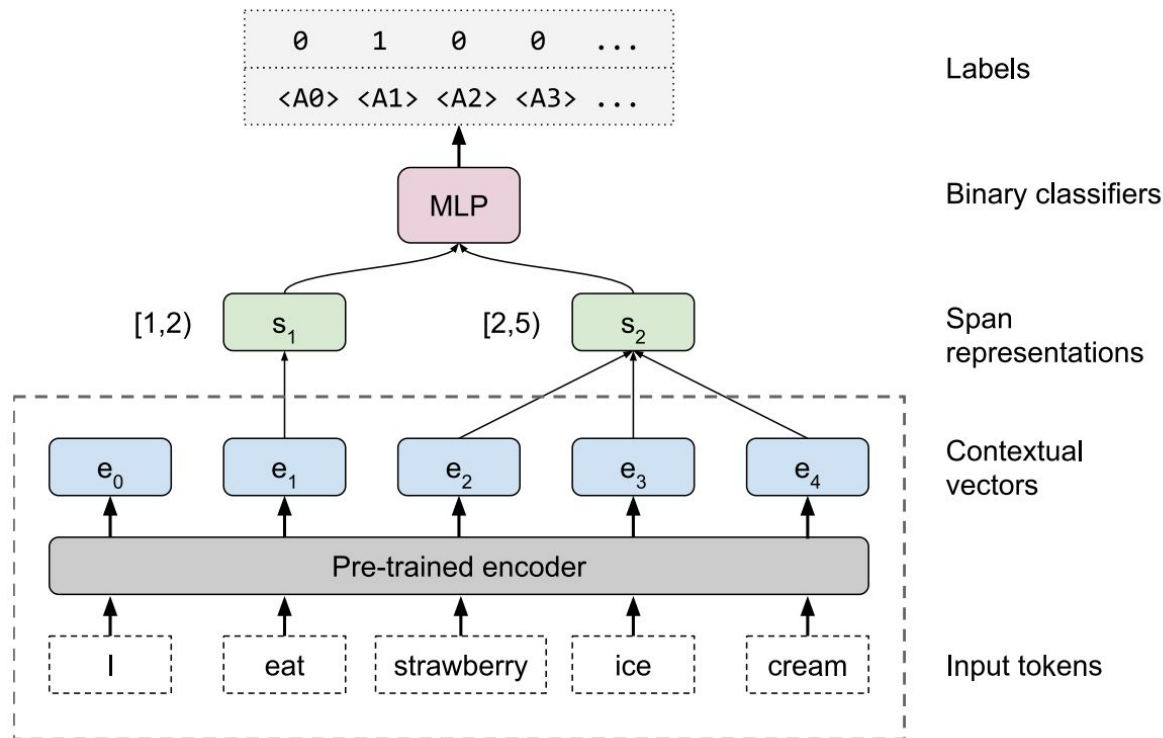
We see predicate-argument role-labeling in this example.

[1,2) "eat"

=> Predicate

[2,5) "strawberry ice cream"

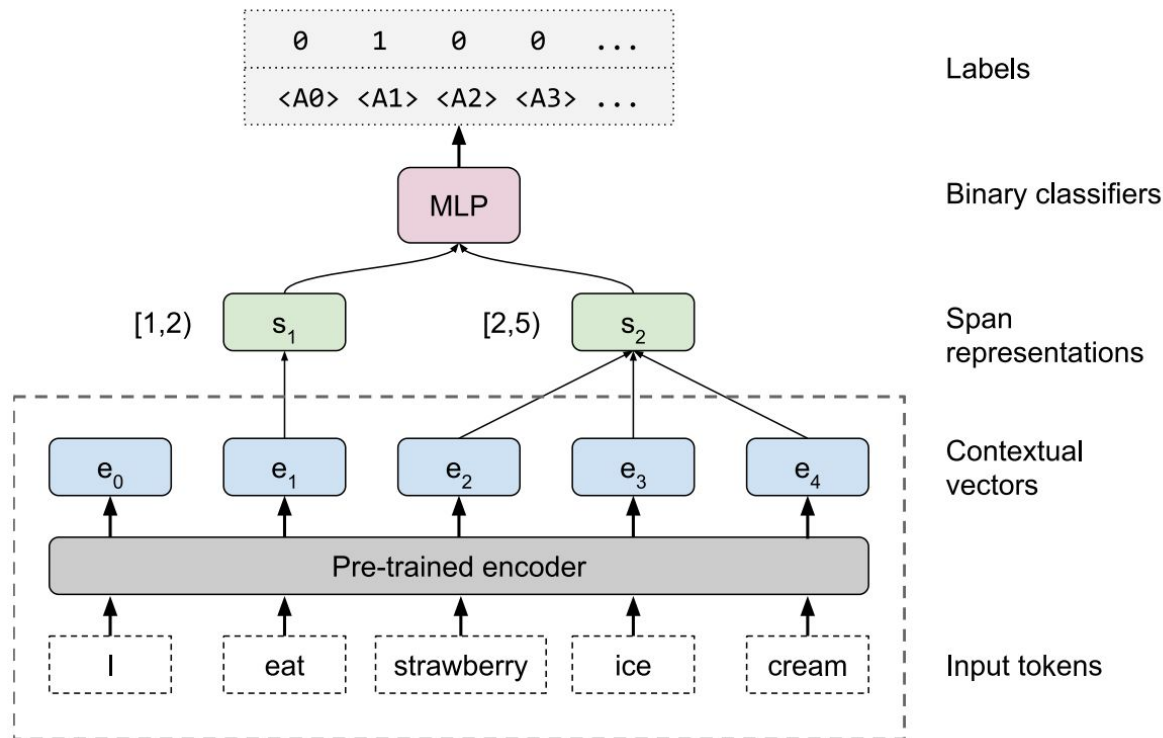
=> Argument



S T A(ction) R

Projection Layer (Spans)

- The only info about the rest of the sentence comes from embeddings within a span.



ST A(ction) R

Lexical baselines

- CoVe: Glove
- ELMo: Layer 0 char CNN
- GPT/Bert: subword embeddings

Randomized Elmo

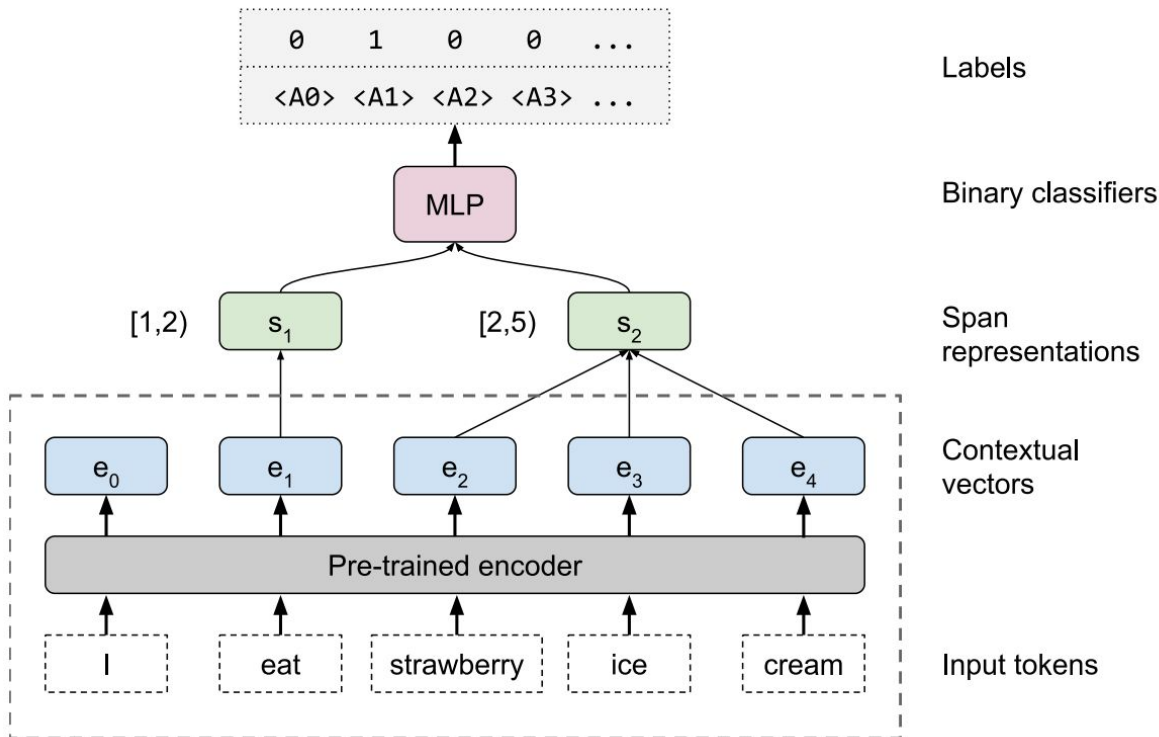
- Replace weights above layer zero with random orthonormal matrixes

Word-Level CNN

- 1 or 2 tokens around the center word

BERT and GPT:

- Scalar mixing vs concatenation



STAR(results)

Additional baselines for ELMO:

Baseline < Randomized Elmo < CNN1 < CNN2 < Full

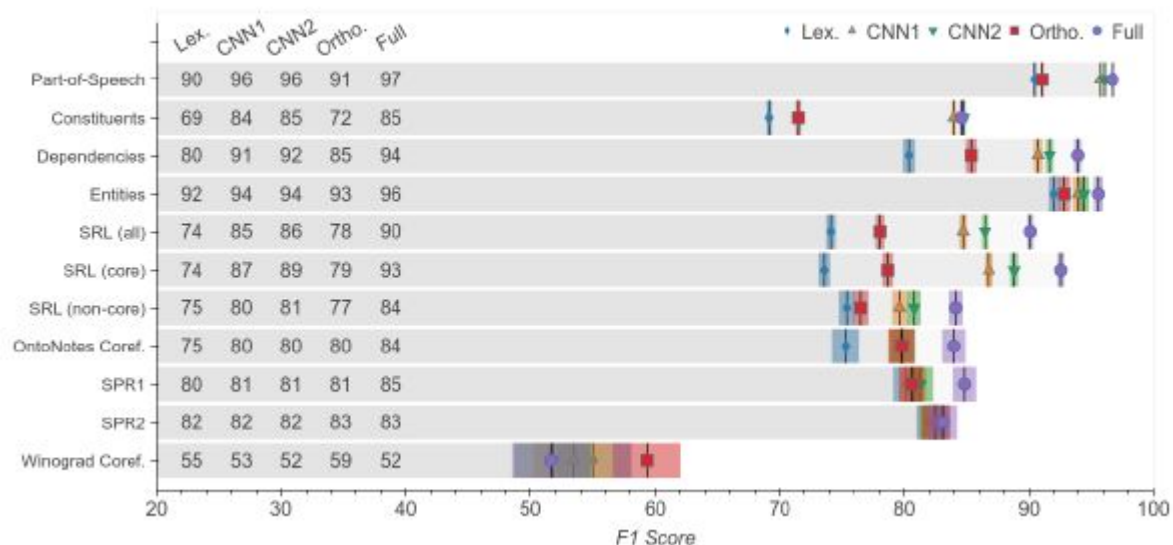


Diagram from: Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S.R., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. *ArXiv, abs/1905.06316*.

STAR(results)

- Bigger gains on syntax vs semantics

	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

Table from: Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S.R., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. *ArXiv, abs/1905.06316*.

S T A R(results)

- Deeper models might help learn semantics

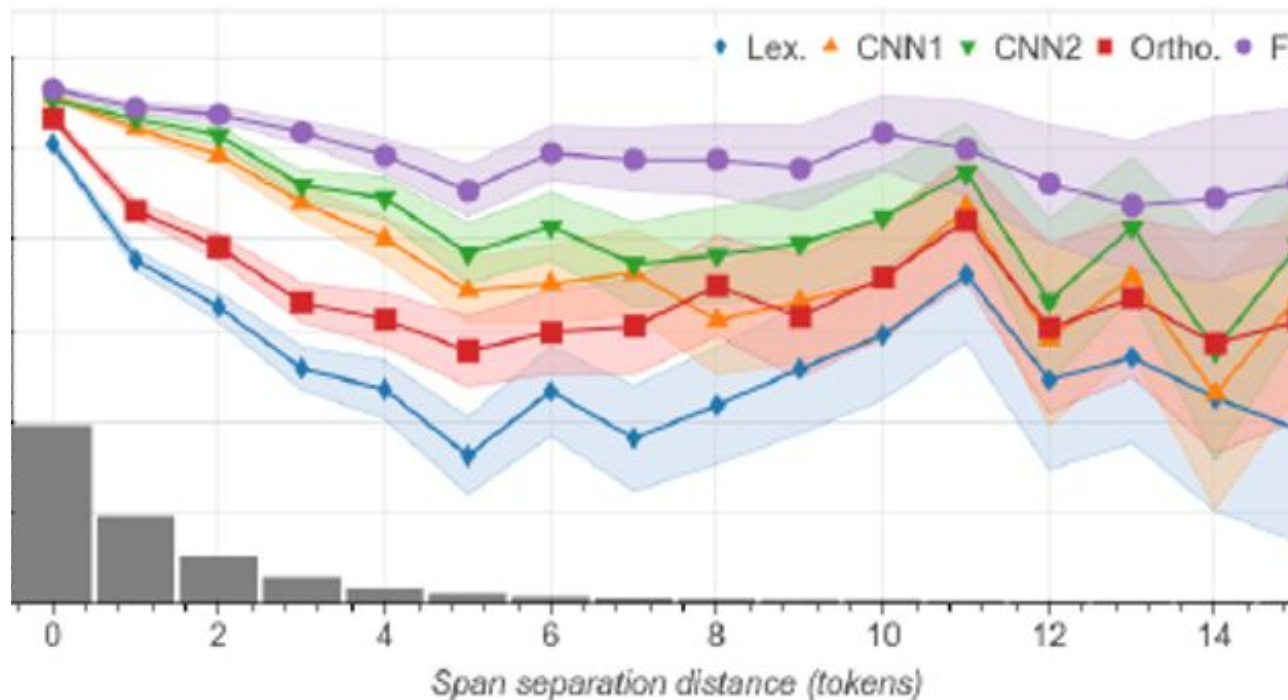
	BERT-base				BERT-large				
	F1 Score			Abs. Δ	F1 Score			Abs. Δ	
	Lex.	cat	mix	ELMo	Lex.	cat	mix	(base)	ELMo
Part-of-Speech	88.4	97.0	96.7	0.0	88.1	96.5	96.9	0.2	0.2
Constituents	68.4	83.7	86.7	2.1	69.0	80.1	87.0	0.4	2.5
Dependencies	80.1	93.0	95.1	1.1	80.2	91.5	95.4	0.3	1.4
Entities	90.9	96.1	96.2	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 from: Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S.R., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. *ArXiv, abs/1905.06316*.

STAR(results)

Lexical baselines do worse on dependency labeling when the spans are distant!

Baseline < Randomized Elmo < CNN1 < CNN2 < Full



STAR(results)

Did results indicate what type of syntactic and semantic information each model encodes, at each layer?

- Lexical layer: Lexical representations used by ELMO and BERT outperform GloVE on all tasks. Especially on constituent and semantic role-labeling, maybe due to handling of morphology by character-level or subword representations.
- Intermediate layers: Mixing, rather than concatenating gave better performance. The authors conjecture that top layers of BERT and GPT became specialized for next-word prediction.
- Syntactic vs semantic tasks: Contextual models have a bigger impact on dependency and constituent labeling, and smaller on tasks that require more semantics, like SPR and Winograd. Deeper models like BERT-large may help difficult semantic tasks.
- Long-distance spans: Contextual models help with long-distance relationships between words.

Probing for semantic evidence of composition by means of simple classification tasks

Allyson Ettinger¹, Ahmed Elgohary², Philip Resnik^{1,3}

¹Linguistics, ²Computer Science, ³Institute for Advanced Computer Studies

University of Maryland, College Park, MD

{aetting, resnik}@umd.edu, elgohary@cs.umd.edu

S(ituation) T A R

- A need for evaluating sentence meaning representations
- How to represent meaning?
- Principle of compositionality!
The meaning of complex expression is determined by its structure and the meanings of its constituents

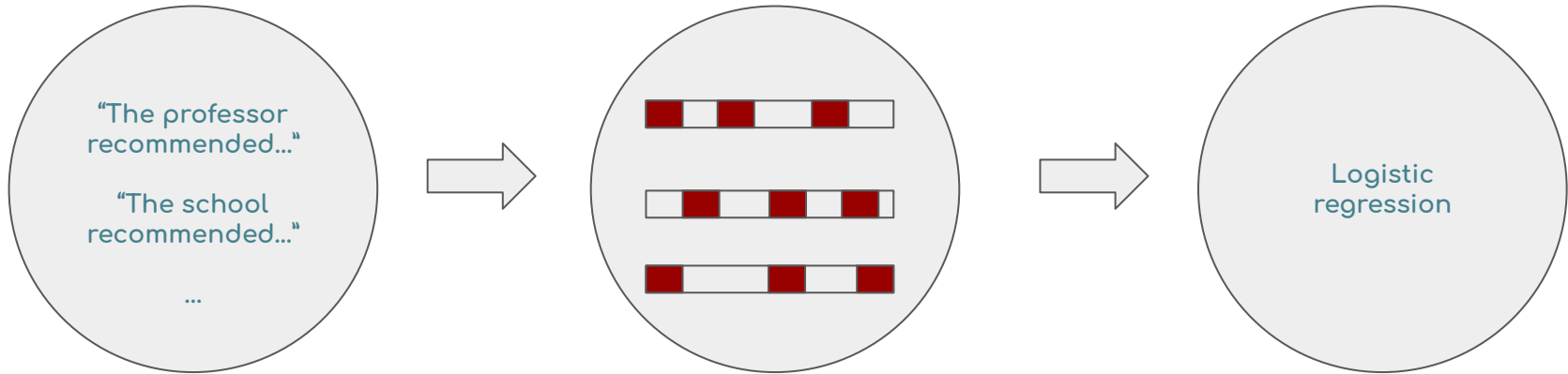
S T(task) A R

Design an evaluation method to measure how well a model composes the meanings of the constituents of a sentence and make this model generalizable for any task

S T A(ction) R

- Construct dataset containing sentences
- Obtain vector representations of sentences from dataset
- Identify semantic information of interest
- Perform binary classification based on semantic information

S T A(ction) R



S T A(ction) R

More on sentence representations...

- Averaging GloVe vectors
- Paraphrastic word averaging embeddings
- Skip-Thought embeddings

S T A(ction) R

More on classification...

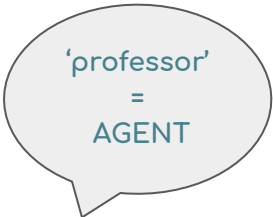
- Logistic regression
- train=1000 sentences
- test=500 sentences
- 5-fold cross validation for tuning

S T A(ction) R

More on classification...

- Semantic information of interest: Semantic Roles
- AGENT = 'professor'
- EVENT = 'recommend'


S T A(ction) R



'professor'
=
AGENT

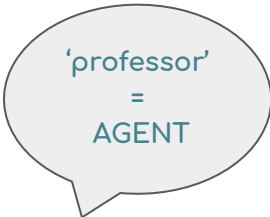
✓ The *professor recommended* the student.

✗ The *student recommended* the professor.



'professor'
≠
AGENT


S T A(ction) R



'professor'
=
AGENT

✓ The *professor* that liked the school *recommended* the researcher.

✗ The *school* that hired the professor *recommended* the researcher.



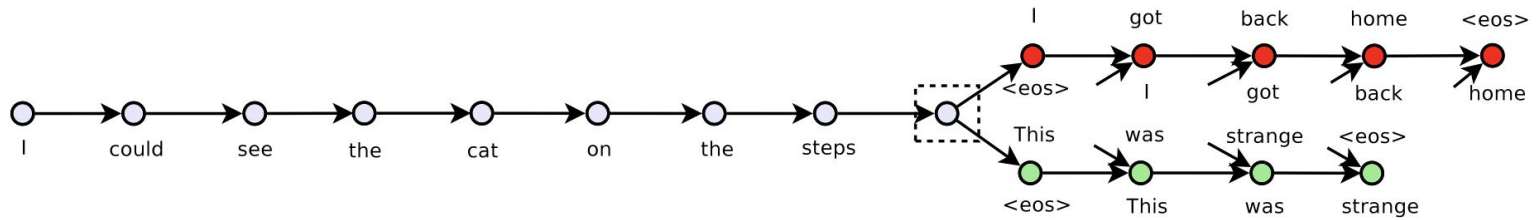
'professor'
≠
AGENT

S T A R(results)

- has-school: correctly detects 'school'
- has-human: correctly detects token as human
- school-as-agent: correctly detects school as an agent

Task	GloVe	Paragram	ST
Has-school	100.0	100.0	100.0
Has-human	99.9	90.5	99.0
School-as-agent	47.98	48.57	91.15

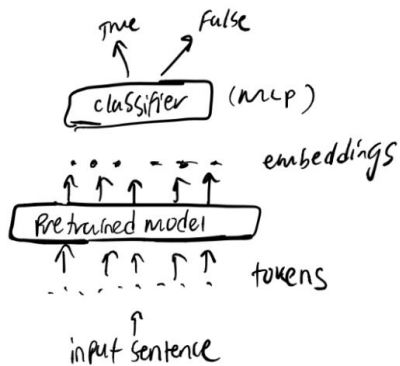
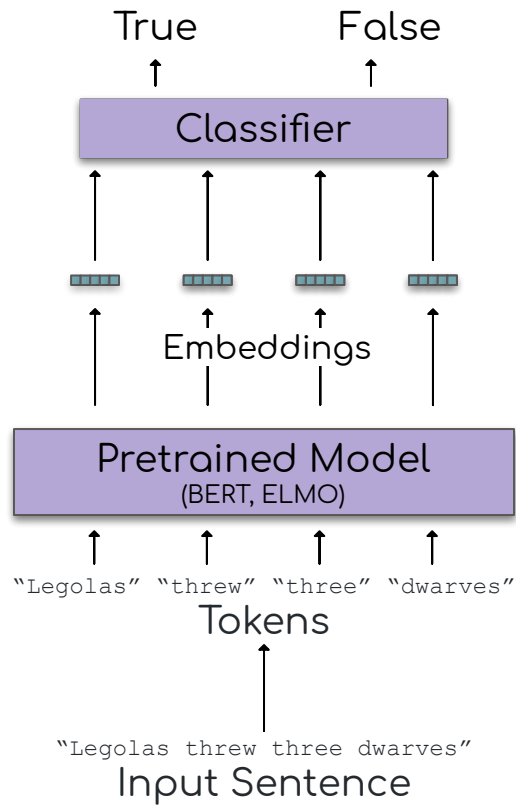
Percentage correct on has-school, has- human, and has-school-as-agent tasks



S T A R(results)

- Skip-Thought embeddings achieve high performance on detecting semantic roles (91.15% accuracy)
- Skip-Thought retains order information
- (Other two models are averaging-based)

How do these
papers motivate
our project?



Cotterell, et al. (1st presentation)

- Paper gives a nice measure of complexity (MCC) that we hope to incorporate in ranking our languages based on transparency vs. opaqueness
- This is important to our hypothesis that more opaque counting systems will be harder to learn

Tenney, et al. (2nd presentation)

- Presents a pipeline for probing language models
- Our model will be very similar but not using the “spans” idea for now (possible future implementation)
- Our model will include a linear classifier directly on top of the pre-trained model
- We can investigate whether contextual representations show similar patterns of improvement over lexical baselines

Ettinger, et al. (3rd presentation)

- Provides inspiration on how to make linguistically-informed probes with classifiers
- Paper probes for different types of information but overall idea is similar; we are making classifiers to probe for semantic and syntactic understanding of numbers

Q&A