What do you learn from context?
Probing for sentence structure in contextualized word representations

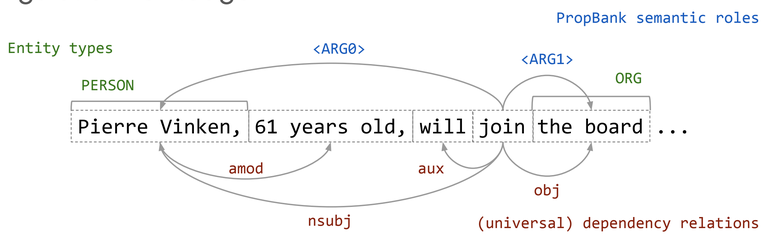
Ian Tenney[*1], Patrick Xia[2], Berlin Chen[3], Alex Wang[4], Adam Poliak[2], R. Thomas McCoy[2], Najoung Kim[2],*
Benjamin Van Durme[2], Samuel R. Bowman[4], Dipanjan Das[1], and Ellie Pavlick[1,5]
[1]Google AI Language, [2]Johns Hopkins University, [3]Swarthmore College, [4]New York University, [5]Brown University

Overview

Question: Contextual encoders like ELMo and BERT outperform traditional word embeddings - why?

- What kind of linguistic information is modeled?
- Mostly syntax, mostly semantics, or both?
- Long range, or just local?
- Are there qualitative differences between encoders?

Our approach: use traditional intermediate tasks - tagging, parsing, etc. - to *probe* contextual representations for linguistic knowledge.



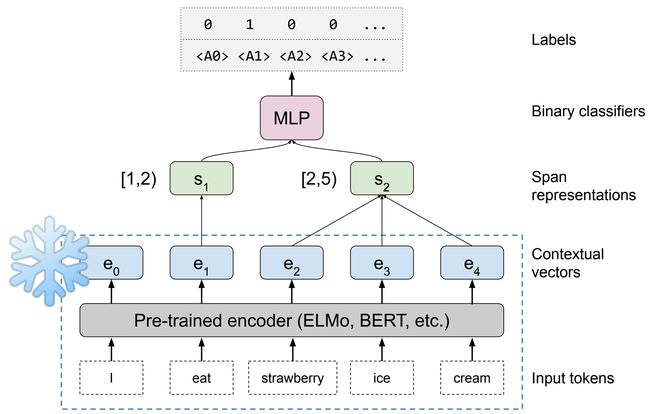
Takeaways:

- Pretrained LM encoders (ELMo, BERT, GPT) give strong representations for syntax, but weaker for semantic phenomena.
- Encoders are able to capture long-range dependencies, and trained models outperform random baselines.

Edge Probing

Multi-label classifier over pairs of spans. Given contextual vectors $E = [e_0, e_1, ..., e_n]$, predict:

- **Unary:** label(s) for span1 = $[i_1, j_1)$
- **Binary:** label(s) for (span1 = $[i_1, j_1)$, span2 = $[i_2, j_2)$)



Classifier sees *inside* of spans only - no other context. Must use encoder representations to infer role of phrase(s) in a sentence.

- Encoder weights are frozen - don't learn new features.
- ELMo-style scalar mixing across encoder layers.
- Common metric (Micro F1) for comparison across tasks.

Results

Labeling versions of 10 tasks, ranging from syntax to semantics. Lexical baseline gives performance from word priors, independent of the rest of the sentence.

	ELMo			BERT (base)		BERT (large)	
	Lexical *(char CNN)*	Full context *(mix all)*	Abs. Δ vs. Lexical	Full context *(mix all)*	Abs. Δ vs. ELMo full	Full context *(mix all)*	Abs. Δ vs. BERT base
Part-of-Speech	90	97	6.3	97	0.0	**97**	0.2
Constituents	69	85	**15.4**	87	**2.1**	**87**	**0.4**
Dependencies	80	94	**13.6**	95	1.1	**95**	0.3
Entities	92	96	3.5	96	0.6	**96**	0.3
SRL (all)	74	90	**16.0**	91	1.2	**92**	1.0
Core	*74*	*93*	***19.0***	*94*	*1.0*	***95***	*1.0*
Non-core	*75*	*84*	*8.8*	*86*	*1.8*	***87***	*1.0*
OntoNotes coref.	75	84	8.7	90	**6.3**	**91**	1.2
SPR1	80	85	4.7	**86**	1.3	86	-0.3
SPR2	82	83	1.0	84	0.7	**84**	0.3
Winograd coref.	54	54	-0.8	55	**1.4**	**61 ± 6**	**6.5**
Relations	58	78	**22.1**	82	**4.2**	**82**	0.5

- **ELMo is great at syntactic tasks, less so at semantics...**
- **BERT improves on ELMo by around 20% (relative) on most tasks, over 40% on coreference**
- GPT is approximately on-par with ELMo; CoVe somewhat lower.

Probing Data

Task	Example target	
Part-of-Speech	The important thing about Disney is that it is a global [brand]$_1$.	`NN` (noun)
Constituents	The important thing about Disney is that it [is a global brand]$_1$.	`VP` (verb phrase)
Dependencies (UD EWT)	The important [thing]$_2$ about Disney [is]$_1$ that it is a global brand.	`nsubj` (nominal subject)
Entities	The important thing about [Disney]$_1$ is that it is a global brand.	`Organization`
Semantic roles (SRL)	[The important thing about Disney]$_2$ [is]$_1$ that it is a global brand.	`ARG1` (agent)
Semantic proto-roles (SPR)	[It]$_1$ [endorsed]$_2$ the White House strategy ...	`{awareness, existed_after, ...}`
OntoNotes coreference	The important thing about [Disney]$_1$ is that [it]$_2$ is a global brand.	`True`
Winograd coreference (DPR)	[Characters]$_2$ entertain audiences because [they]$_1$ want people to be happy. Characters entertain [audiences]$_2$ because [they]$_1$ want people to be happy.	`True` `False`
SemEval relations	The [burst]$_1$ has been caused by water hammer [pressure]$_2$.	`Cause-Effect(`e_2, e_1`)`

Ablations on ELMo

How much is just due to architecture? (versus training)

- ☐ ELMo outperforms random-ELMo (orthonormal LSTM weights)

What about local features?

- ☐ ELMo outperforms lexical baseline + local CNN (±1 or ±2 tokens)
- ☐ Real ELMo performance drops only slowly on long-range edges