Transformers: Capabilities and Limitations

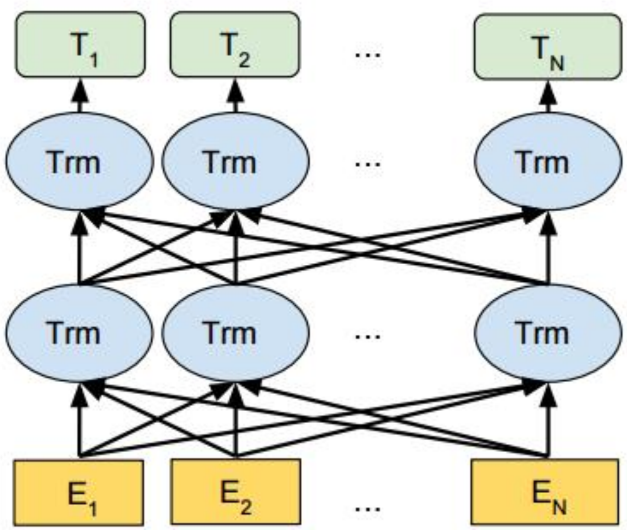
ECE324, Winter 2022

Slides from Jackie C.K. Cheung

Michael Guerzhoy

Pre-trained Language Models

- BERT (Devlin et al., 2019) and friends are the most popular starting point of current NLP systems



BERTology

- BERTology investigates what BERT-like models learn
  - Syntactic knowledge
  - Semantic knowledge
  - World knowledge

Syntactic knowledge

- Syntax: the rules according to which sentences are formed

Grammar	Lexicon
S → *NP VP*	*Det* → *that* | *this* | *the* | *a*
S → *Aux NP VP*	*Noun* → *book* | *flight* | *meal* | *money*
S → *VP*	*Verb* → *book* | *include* | *prefer*
NP → *Pronoun*	*Pronoun* → *I* | *she* | *me*
NP → *Proper-Noun*	*Proper-Noun* → *Houston* | *NWA*
NP → *Det Nominal*	*Aux* → *does*
Nominal → *Noun*	*Preposition* → *from* | *to* | *on* | *near* | *through*
Nominal → *Nominal Noun*	
Nominal → *Nominal PP*	
VP → *Verb*	
VP → *Verb NP*	
VP → *Verb NP PP*	
VP → *Verb PP*	
VP → *VP PP*	
PP → *Preposition NP*	

Figure 13.1 The $\mathscr{L}_1$ miniature English grammar and lexicon.

https://web.stanford.edu/~jurafsky/slp3/13.pdf

Parse trees

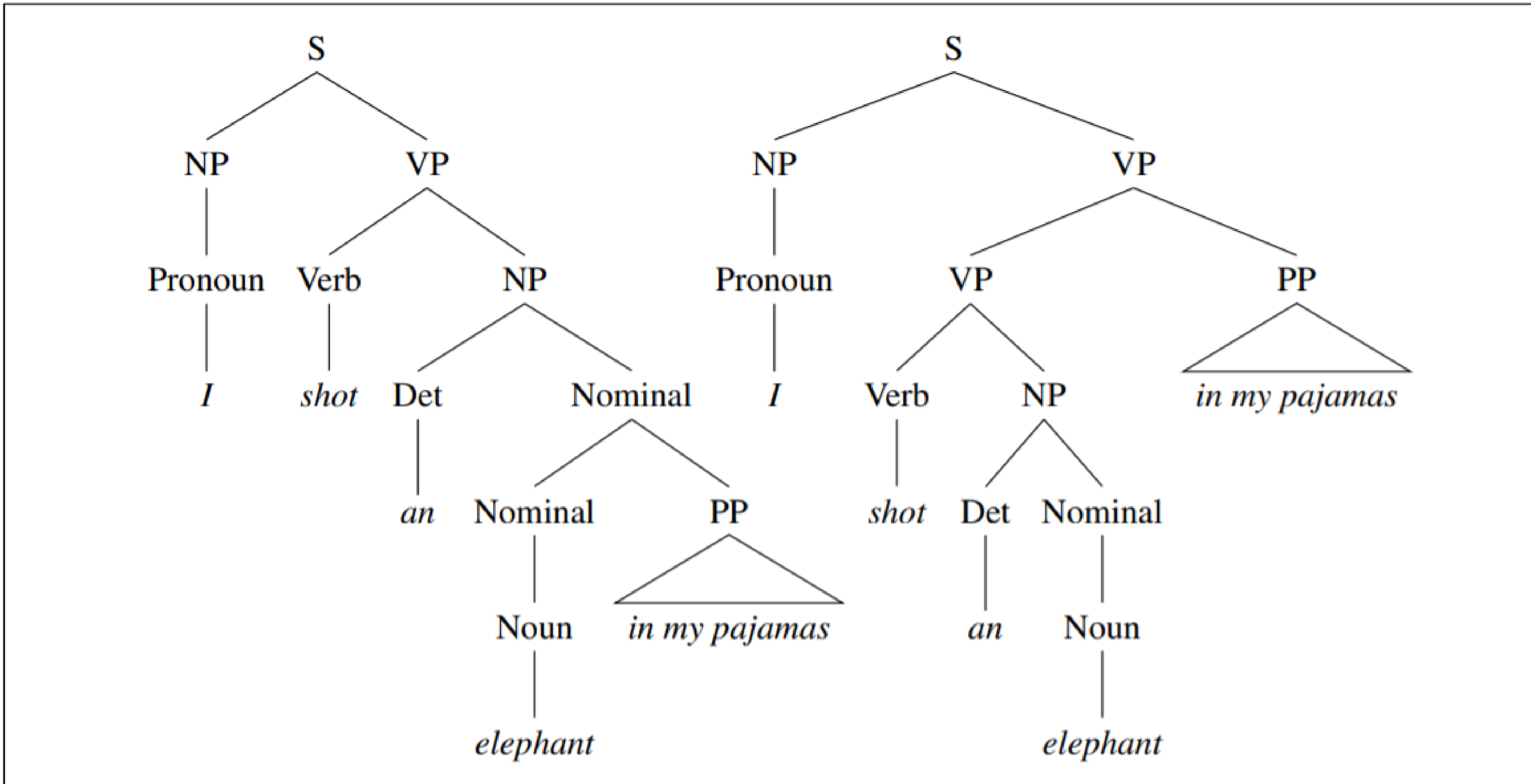


Figure 13.2 Two parse trees for an ambiguous sentence. The parse on the left corresponds to the humorous reading in which the elephant is in the pajamas, the parse on the right corresponds to the reading in which Captain Spaulding did the shooting in his pajamas.

BERT's syntactic representation study

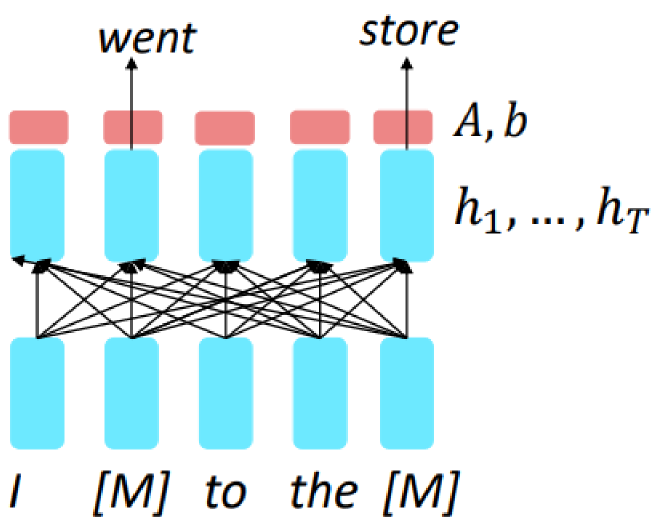
A Structural Probe for Finding Syntax in Word Representations

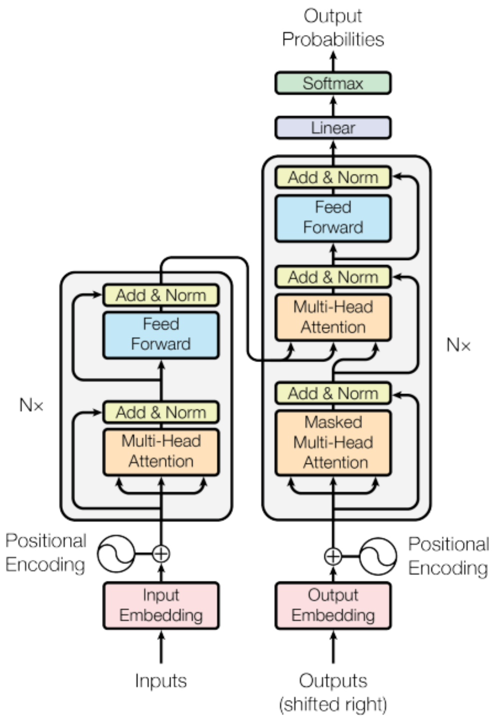
John Hewitt
Stanford University
johnhew@stanford.edu

Christopher D. Manning
Stanford University
manning@stanford.edu

- Want to determine if BERT embeddings contain syntactic information
- Idea: can we predict the distance in the parse tree from the embeddings

Aside: BERT embeddings





The last layer serves as an encoding because we train by making h_i predict the i-th word
The i-th vector in an inner layer is related to the i-th word because of residual connections
Can concatenate the i-th vectors from different layers

- Let the embedding of the i-th word in sentence ℓ be h_i^ℓ
- Define the distance as

$$d_B(h_i^\ell, h_j^\ell)^2 = \left(B(h_i^\ell - h_j^\ell)\right)^T \left(B(h_i^\ell - h_j^\ell)\right)$$

B is a matrix learned from the data by minimizing the following over all pairs of word in all sentences

$$\min_B \sum_\ell \frac{1}{|s^\ell|^2} \sum_{i,j} \left| d_{T^\ell}(w_i^\ell, w_j^\ell) - d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)^2 \right|$$

Distance in the parse tree

Results

- Computing the distance using the middle layer of BERT produces distances that are very similar to parse tree distances

Semantic knowledge

- Semantics: the meaning of words

What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models

Allyson Ettinger

- Idea: study how BERT predicts masked words ("cloze task")

Context	Expected	Inappropriate
He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ___	*lipstick*	*mascara* | *bracelet*
He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ___	*football*	*baseball* | *monopoly*

Context	Match	Mismatch
A robin is a ___	*bird*	*tree*
A robin is not a ___	*bird*	*tree*

Context	$\text{BERT}_{\text{LARGE}}$ predictions
Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her __	*car, house, room, truck, apartment*
The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a __	*note, letter, gun, blanket, newspaper*
At the zoo, my sister asked if they painted the black and white stripes on the animal. I explained to her that they were natural features of a __	*cat, person, human, bird, species*

Context	$\text{BERT}_{\text{BASE}}$ predictions	$\text{BERT}_{\text{LARGE}}$ predictions
the camper reported which girl the bear had __	*taken, killed, attacked, bitten, picked*	*attacked, killed, eaten, taken, targeted*
the camper reported which bear the girl had __	*taken, killed, fallen, bitten, jumped*	*taken, left, entered, found, chosen*
the restaurant owner forgot which customer the waitress had __	*served, hired, brought, been, taken*	*served, been, delivered, mentioned, brought*
the restaurant owner forgot which waitress the customer had __	*served, been, chosen, ordered, hired*	*served, chosen, called, ordered, been*

Semantic knowledge: summary

- BERT's overall good performance sometimes relies on shortcuts – statistical patterns that are not directly connected to meaning
- BERT is good at identifying objects as belonging to categories
  - E.g. robin is a bird
- BERT is bad at dealing with negation

World knowledge

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London

	Relation	Query	Answer
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence
	P20	Adolphe Adam died in ____.	Paris
	P279	English bulldog is a subclass of ____.	dog
	P37	The official language of Mauritius is ____.	English
	P413	Patrick Oboya plays in ____ position.	midfielder
	P138	Hamburg Airport is named after ____.	Hamburg
	P364	The original language of Mon oncle Benjamin is ____.	French
	P54	Dani Alves plays with ____ .	Barcelona
	P106	Paul Toungui is a ____ by profession .	politician
	P527	Sodium sulfide consists of ____.	sodium
	P102	Gordon Scholes is a member of the ____ political party.	Labor
	P530	Kenya maintains diplomatic relations with ____.	Uganda
	P176	iPod Touch is produced by ____.	Apple
	P30	Bailey Peninsula is located in ____.	Antarctica
	P178	JDK is developed by ____.	Oracle
	P1412	Carl III used to communicate in ____.	Swedish
	P17	Sunshine Coast, British Columbia is located in ____.	Canada
	P39	Pope Clement VII has the position of ____ .	pope
	P264	Joe Cocker is represented by music label ____.	Capitol
	P276	London Jazz Festival is located in ____.	London
	P127	Border TV is owned by ____.	ITV
	P103	The native language of Mammootty is ____.	Malayalam
	P495	The Sharon Cuneta Show was created in ____.	Philippines

Results competitive with other systems

Can Transformer-like architectures *understand* language?

- Argument for “yes”: remarkable performance on cloze tasks, remarkable ability to generate language
- Arguments for “no”
  - Mistakes on cloze tasks show that the good performance is due merely to learning statistical patterns
  - Humans can to attribute meaning to generated language even when it’s meaningless
  - A bunch of matrix multiplications can’t understand anything

Climbing towards NLU:
On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

Abstract

The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has *a priori* no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

Meaning

Definitions in *Climbing toward NLU*

- Form: any observable realization of language: marks on a page, pixels or bytes in memory, movements of articulators…
- Meaning: the relation between the form and something external to language
- Understanding: retrieving the communicative intent from an expression
    - Communicative intent is about something outside the language (e.g. "Open the window!" is about a the window)

Communication

“The speaker has a certain communicative intent i, and chooses an expression e with a standing meaning s which is fit to express i in the current communicative situation. Upon hearing e, the listener then reconstructs s and uses their own knowledge of the communicative situation and their hypotheses about the speaker’s state of mind and intention in an attempt to deduce i.”

This active participation of the listener is crucial to human communication For example, to make sense of (8) and (9) (from Clark, 1996, p.144), the listener has to calculate that Napoleon refers to a specific pose (hand inside coat flap) or that China trip refers to a person who has recently traveled to China.

(8) The photographer asked me to do a Napoleon for the camera.

(9) Never ask two China trips to the same party

“We argue that a model of natural language that is trained purely on form will not learn meaning: if the training data is only form, there is not sufficient signal to learn the relation M between that form and the non-linguistic intent of human language users, nor C between form and the standing meaning the linguistic system assigns to each form.”

Aside: Searle's Chinese Room Experiment

- A person is in a large room containing instructions for how to transform notes in Chinese to responses in Chinese
- The person doesn't speak Chinese but follow the instructions
- Argument: the person doesn't understand Chinese
- Counterargument:
    - The person + the room (+ whatever energy is needed to go through the instructions in a short amount of time) is a complex system that might be said to understand Chinese

The Octopus test

A and B, both fluent speakers of English, are independently stranded on two uninhabited islands*. They soon discover that previous visitors to these islands have left behind telegraphs and that they can* ***communicate with each other via an underwater cable****. A and B start happily typing messages to each other. Meanwhile****, O, a hyper-intelligent deep-sea octopus who is unable to visit or observe the two islands, discovers a way to tap into the underwater cable and listen in on A and B's conversations****. O knows nothing about English initially, but is very good at detecting statistical patterns.* ***Over time, O learns to predict with great accuracy how B will respond to each of A's utterances****. O also observes that* ***certain words tend to occur in similar contexts****, and perhaps learns to generalize across lexical patterns by hypothesizing that they can be used somewhat interchangeably. Nonetheless,* ***O has never observed these objects, and thus would not be able to pick out the referent of a word when presented with a set of (physical) alternatives****. At some point, O starts feeling lonely.* ***He cuts the underwater cable and inserts himself into the conversation, by pretending to be B and replying to A's messages. Can O successfully pose as B without making A suspicious****? This constitutes a weak form of the Turing test (weak because A has no reason to suspect she is talking to a nonhuman); the interesting question is whether O fails it because he has not learned the meaning relation, having seen only the form of A and B's utterances*

Argument: the Octopus would not be able to fake a conversation where world knowledge is required

Now say that A has invented a new device, say a coconut catapult. She excitedly sends detailed instructions on building a coconut catapult to B, and asks about B's experiences and suggestions for improvements. Even if O had a way of constructing the catapult underwater, he does not know what words such as rope and coconut refer to, and thus can't physically reproduce the experiment

Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O has no idea what A "means". Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test, if A hadn't been eaten by the bear before noticing the deception.7

Learning programming language semantics without grounding

Imagine that we were to train an LM on all of the well-formed Java code published on Github. The input is only the code. It is not paired with bytecode, nor a compiler, nor sample inputs and outputs for any specific program. We can use any type of LM we like and train it for as long as we like. We then ask the model to execute a sample program, and expect correct program output.

Not just language modeling

What about systems which are trained on a task that is not language modeling — say, **semantic parsing, or reading comprehension tests** — and that use word embeddings from BERT or some other large LM as one component? Numerous papers over the past couple of years have shown that using such pretrained embeddings can boost the accuracy of the downstream system drastically, even for tasks that are clearly related to meaning. **Our arguments do not apply to such scenarios**: reading comprehension datasets include information which goes beyond just form, in that they specify semantic relations between pieces of text, and thus a sufficiently sophisticated neural model might learn some aspects of meaning when trained on such datasets. It also is conceivable that whatever information a pretrained LM captures might help the downstream task in learning meaning, without being meaning itself.

Counterarguments

- Perhaps a tiny bit of grounding (digits of pi?) is enough if there is a lot of data
  - Is missing just this tiny bit really important?
- Perhaps using Occam's Razor is enough
  - To explain a whole lot of text, it's efficient to reinvent all of Physics
- "Meaning"/"Understanding" are properties of complex systems
  - A large enough model is complex enough that it can *understand* in the sense that language models understand