

Are we there yet?

Meaning in the age of large language models

Jacopo Tagliabue, Director of AI (Coveo)
Computing Language: Love Letters, Large Models and NLP, CiE 2022



Ciao!



BUILDING

- Co-founder of Tooso and now Director of AI (TSX:CVO), after Tooso was acquired by Coveo
- OS contributor (>1000 Github stars) in MLOps and IR

STUDYING

- 25+ research papers in top NLP/ML venues, invited speaker (SIRIP, KDD), best paper NAACL21
- Adj. Prof. of MLSys at NYU

TALKING

- Co-organizer of SIGIR eCom, Sponsorship Chair CIKM, Committee Member for ECNLP and ECONLP
- Invited speaker at corporations (BBC, Walmart), startups (Tubi), tech companies (NVIDIA, Pinterest, Stitch Fix)

The golden era of NLP

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

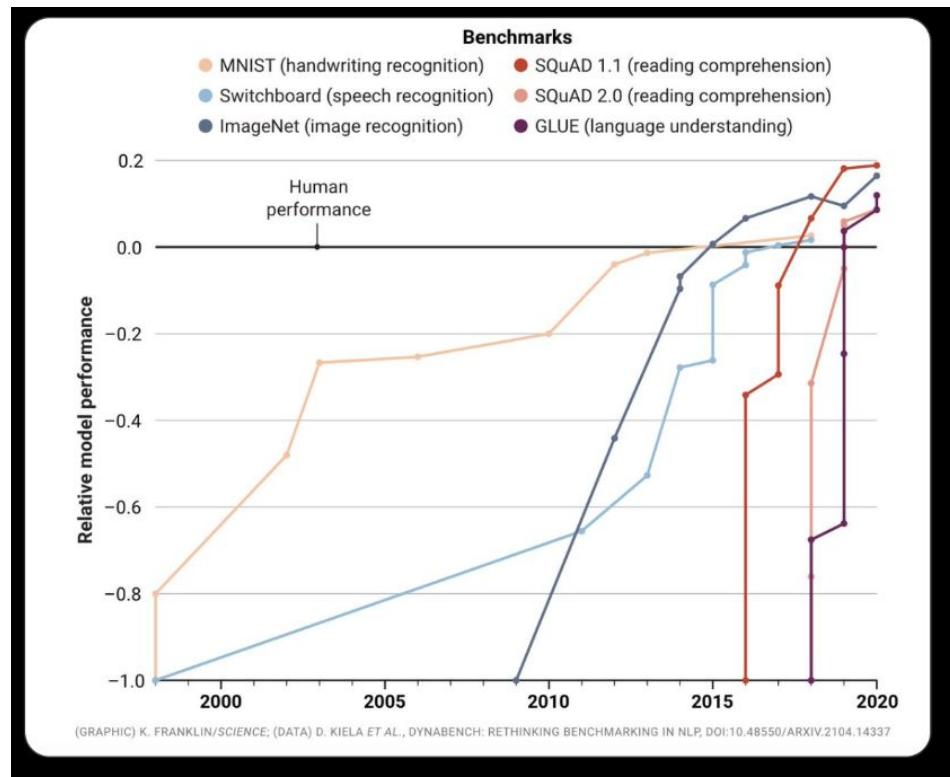
The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

A robot wrote this entire article. Are you scared yet, human?

GPT-3

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below



We are approaching a
point of “artificial fluency”
that is hard to ignore!

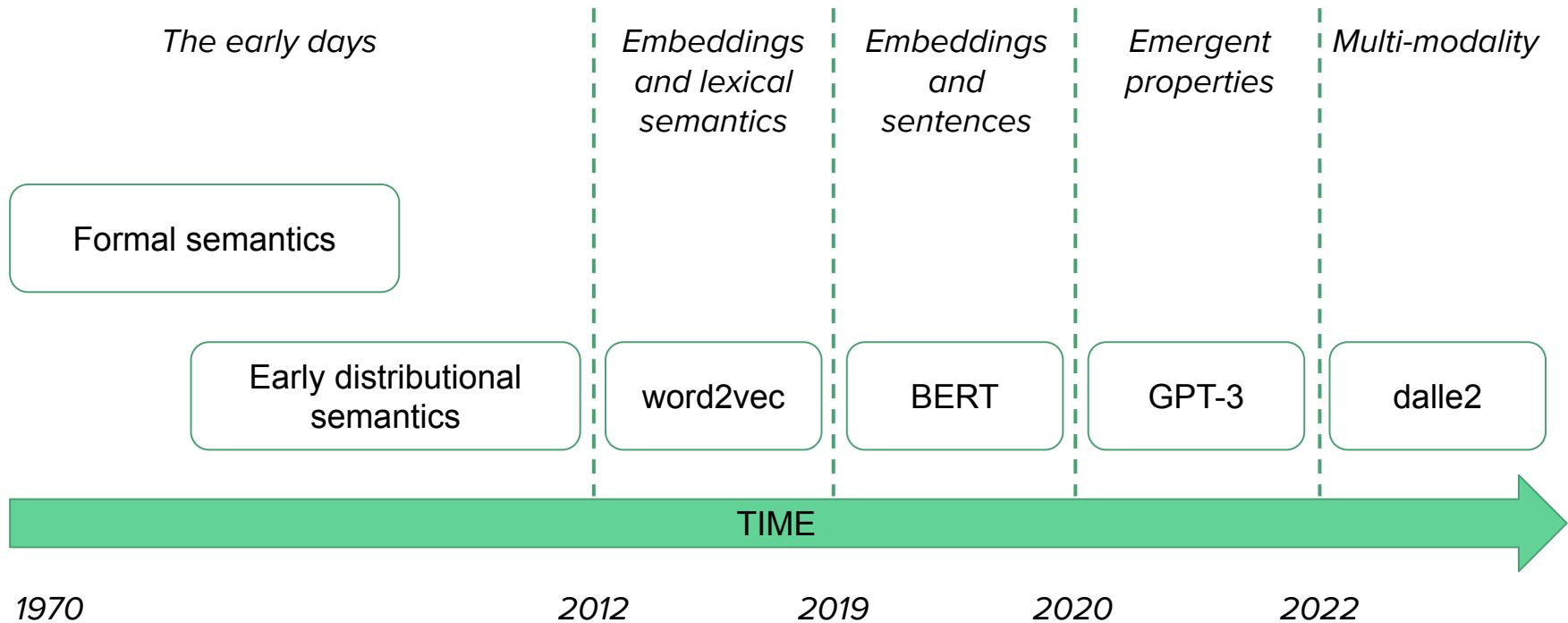
The golden era of NLP



“It is impossible to review the specifics of your tenure file without becoming enraptured by the vivid accounts of your life. However, it is not a life that will be appropriate for a member of the faculty at Indiana University, and **it is with deep regret that I must deny your application for tenure.** ... Your lack of diplomacy, your flagrant disregard for the feelings of others, (...), and, frankly, the fact that you often take the side of the oppressor, **leads us to the conclusion that you have used your tenure here to gain a personal advantage and have failed to adhere to the ideals of this institution.**”

TL;DR: meaning may be decoupled from (perceived) competence (was Searle right all along?)

Are we there yet?



Caveat

The literature is utterly *insane*: while I did my best to review even very recent papers, it's likely the views represented here are only a partial overview of the current landscape.

We aim to review in a fairly non-technical manner the current debate on *meaning*, and point to further readings when appropriate: generally speaking, we shall defend the *boring* view that large language models are both interesting and incomplete. As the wise man said:

“To say anything good about anyone is beyond the scope of this talk”

Semantics and good ol' NLP

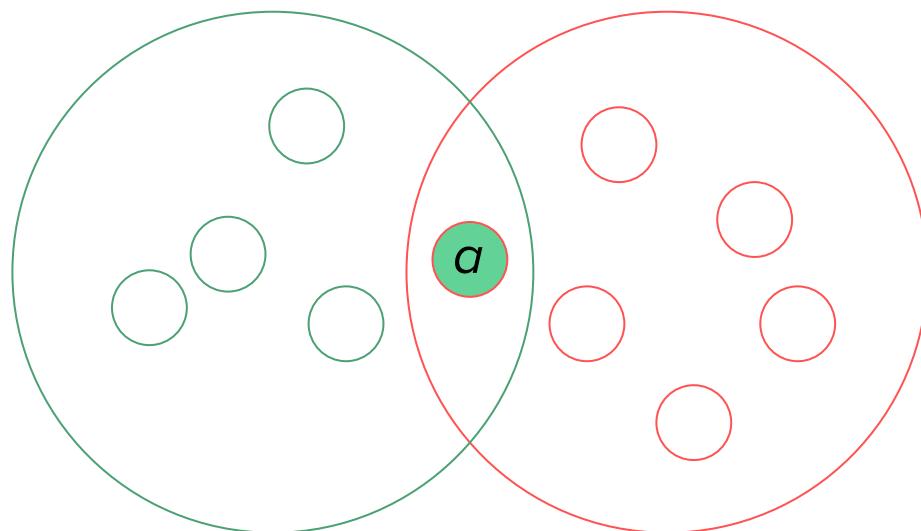
“Meaning” (“semantics”) means many things

- Lexical semantics
 - Words, e.g. the meaning of “cat”, which is different from (but related to) “dog”, and also different from (and *not* related to) “Rome”.
- Compositional semantics
 - Chunks, e.g. semantics for noun phrases: “dress”, “black dress”, “black long dress” - adding adjectives modifies the meaning of the noun (in this case, restrict its extension).
 - Full sentence, e.g. entailment: “every man is mortal”, “Socrates is a man” **entail** “Socrates is mortal”; pragmatic implicatures etc.
- There is also a bunch of related concepts:
 - Syntactic parsing used to be considered a prerequisite for semantics (“Dog bites man” vs “man bites dog”).
 - We are often asked whether “model X *understands* language”; usually understanding presupposes handling meaning correctly.

Meaning in Frege (Montague etc.): words are sets

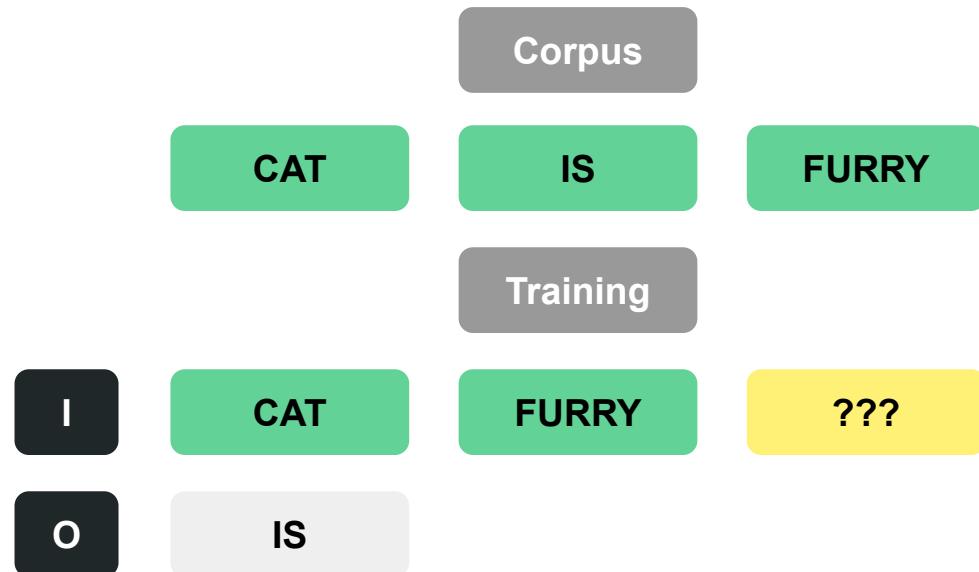
- Lexical meaning as reference provided by an interpretation function: **meaning requires words and a world**
- Semantics as (mainly) functional composition
- **Sentence > lexical semantics**
- Can do:
 - Zero-shot generalization
 - Entailment
 - Extensional vs Intensional
- Can't do:
 - How is the “interpretation function” learned in the first place?
 - Relations between words: “Rome” is more similar to “Berlin” than “cat”

Pa & Qa



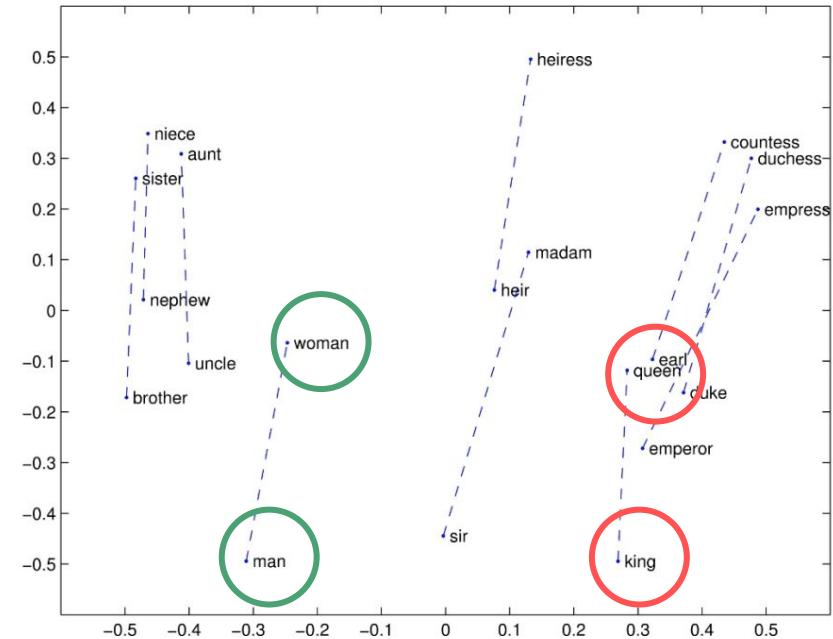
Meaning in Mikolov et al.: *words are vectors*

- Lexical meaning as points in a vector space (“embeddings”): **no reference to external world**
- Semantics as (mainly) vector composition
- **Lexical semantics > sentence**



Meaning in Mikolov et al.: *words are vectors*

- Can do:
 - Learn from corpora
 - Vector semantics is rich: analogies, synonyms etc.
- Can't do:
 - Zero-shot generalization
 - Logical symbols (e.g. NOT)



$man : woman = king : queen$

Doing things with ~~words~~ meanings

- People interested in Montague-style semantics are typically logicians, linguists, philosophers: they are building explanations around language (the notion of truth, logical consequence etc.)
- People interested in embeddings are typically computer scientists / NLP practitioners: they are building applications through computation (text classification, sentiment analysis, entity extraction - corpora is the natural input)

All models are wrong, maybe some will be useful

- Even just restricting our attention to lexical semantics, “words as vectors” and “words as sets” fall short of capturing two fundamental dimensions of meaning:
 - Referential: point, recognize, name etc. **nothing in word2vec is about referential knowledge**
 - Inferential: paraphrase, explain, entail etc. **very little in model-theory is about word-level inference (outside of special logical words)**



The best of both worlds?

- While not surprising that different “scientific tribes” may have different interest, some practitioners have been tried to combine the two views in a principled way, but it’s fair to say that practical impact has been limited.

Frege in Space: A Program for Compositional Distributional Semantics

MARCO BARONI,¹ RAFFAELLA BERNARDI¹ AND
ROBERTO ZAMPARELLI¹

To Emanuele Pianta,
in memoriam

The lexicon of any natural language encodes a huge number of distinct word meanings. Just to understand this article, you will need to know what thousands of words mean. The space of possible sentential meanings is infinite: In this article alone, you will encounter many sentences that are valid, and many more that are false or dubious. Statistical

We lack a good model of what “meaning” is. If syntax is a parse tree, what is *meaning*?

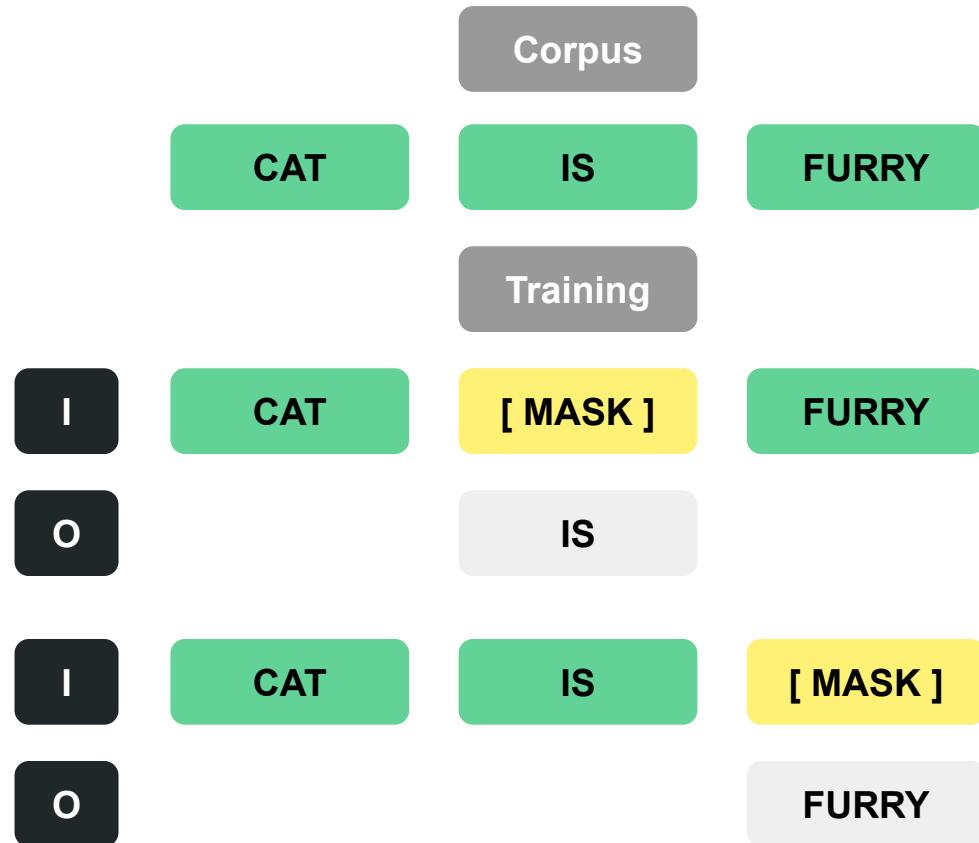
There's more to read - Part 1

- A great intro talk (with slides) on distributional methods and their relation to structuralism, [by Piero Molino](#)
- [A long survey](#) on word meaning in human and machines
- The SEP entry on word meaning, by [Luca Gasparri](#)
- On the lexical information encoded in embeddings, and how it aligns with human judgment, [a recent article in Nature Human Behavior](#); a nice article on how meaning of words change over time and vectors can be “re-aligned” by [Federico Bianchi](#)
- On recovering compositionality for noun phrases from vectors, see the classic [here](#) from Baroni and Zamparelli, and our own work on [learning NP semantics through data collected on a search engine](#).

From words to sentences

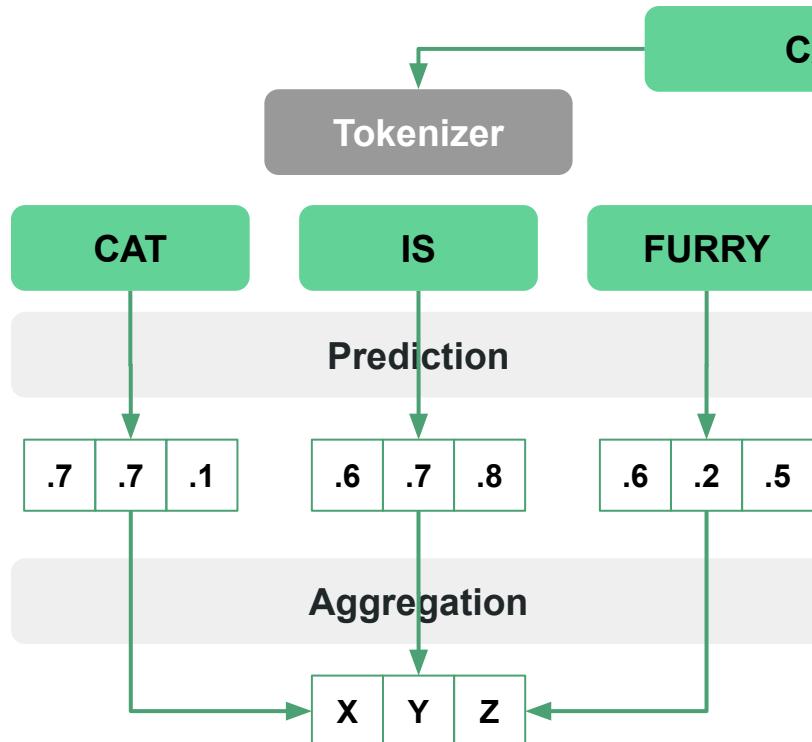
Transformers

- Classical word embeddings are static (i.e. “bank” has one vector)
- BERT combines *contextual* embeddings (see also ELMo) with a new sequential architecture: “bank of the river” and “bank with the atm” receive a different representation
- As **Transformers** shift focus to sentences, **sentence > lexical semantics but != from formal semantics**

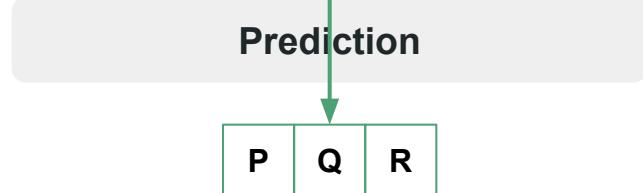


Transformers

word2vec is mostly about words



*BERT is mostly about sentences**



* Note: *it is possible to recover word vectors.*

Transformers
improved
performance in most
NLP tasks

BERTology

- From BERT, a long list of “children” have spawned: [RoBERTa](#), [DeBERTa](#), [DistillBERT](#), etc.
- While BERT is small by today’s standards (it can fit on this laptop comfortably), its complexity and “black-box-ness” motivated several studies to understand what “it knows”.

Although it is clear that BERT works remarkably well, it is less clear *why*, which limits further hypothesis-driven improvement of the architecture. Unlike CNNs, the Transformers have little cognitive motivation, and the size of these models limits our ability to experiment with pre-training and perform ablation studies. This explains a large number of studies over the past year that attempted to understand the reasons behind BERT’s performance.

BERTology

Syntax (UNCLEAR)

BERT latent space encodes information about syntactic parse trees (in the same sense as word2vec encodes analogies)... **BUT** researchers showed that explicit syntactic information during training doesn't help and that the actual word order doesn't matter much ... **BUT** indeed the order does matter when lexical semantics is not enough.

BERTology

Lexical semantics (YES)

- BERT outperforms word2vec on word similarity and relatedness datasets

Probable completions (MOSTLY)

- “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 ____” (BERT: **lipstick**)

Logical symbols (NO)

- Negation is poorly handled

Context	BERT _{LARGE} predictions
<i>A robin is a ____</i>	<i>bird, robin, person, hunter, pigeon</i>
<i>A daisy is a ____</i>	<i>daisy, rose, flower, berry, tree</i>
<i>A hammer is a ____</i>	<i>hammer, tool, weapon, nail, device</i>
<i>A hammer is an ____</i>	<i>object, instrument, axe, implement, explosive</i>
<i>A robin is not a ____</i>	<i>robin, bird, penguin, man, fly</i>
<i>A daisy is not a ____</i>	<i>daisy, rose, flower, lily, cherry</i>

BERTology

Common-sense in two sentences (NO)

- “In each case, BERT provides completions that are sensible in the context of the second sentence, but that fail to take into account the context provided by the first sentence”

Context	BERT _{LARGE} predictions
<p><i>Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her _____</i></p>	<i>car, house, room, truck, apartment</i>
<p><i>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 _____</i></p>	<i>note, letter, gun, blanket, newspaper</i>

Old dog, old tricks

- While BERT constitutes a decisive step forward in language processing, its semantics is still primitive upon closer examination.
- Even using **understanding** as a spectrum, and adopting the shallowest possible notion of **semantics**, a model that fails with negation, changes predictions between paraphrases, and fails to generalize outside of training can hardly be considered semantically proficient.

Failure Rate (⌚)	Example Test cases (with expected behavior and ⌚ prediction)
20.0	<p>C: Victoria is younger than Dylan. Q: Who is less young? A: Dylan ⌚: Victoria</p>
91.3	<p>C: Anna is worried about the project. Matthew is extremely worried about the project. Q: Who is least worried about the project? A: Anna ⌚: Matthew</p>

Old dog, old tricks

- Aside from specific shortcomings, more general arguments have been made against the idea that BERT *understands English*.
- Crucially, they all go back to the very first idea we described today, i.e. the **referential** nature of language: since BERT (like word2vec before) only learns from textual data, it will never generalize a semantics: “you can’t learn Finnish from the radio alone”.

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

the structure and use of language and the ability to ground it in the world. While large neural LMs may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge. We argue in this paper that gen-

Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?

William Merrill* Yoav Goldberg*† Roy Schwartz‡ Noah A. Smith*§
*Allen Institute for AI, United States †Bar Ilan University, Israel
‡Hebrew University of Jerusalem, Israel §University of Washington, United States
{willm,yoavg,roys,noah}@allenai.org

Abstract

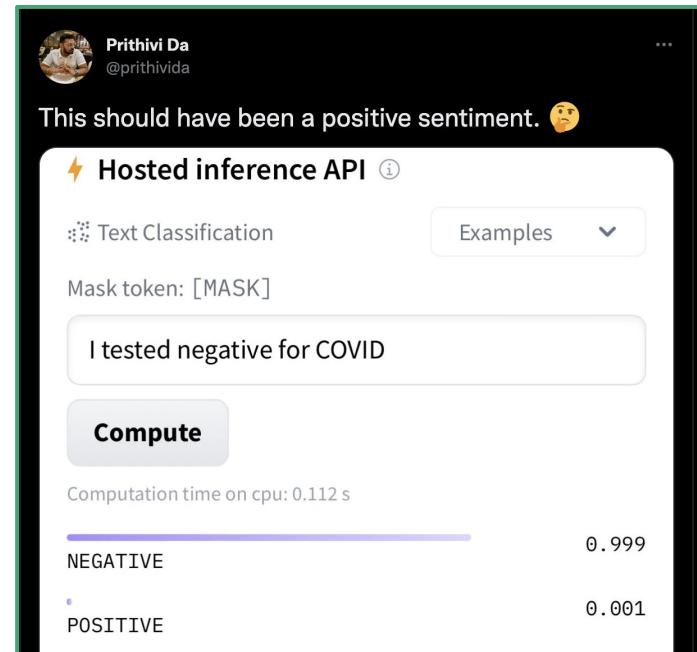
Language models trained on billions of tokens have recently led to unprecedented results on many NLP tasks. This success raises the question of whether, in principle, a system can ever “understand” raw text without access to some form of grounding. We formally investigate the abilities of ungrounded systems to acquire

semantic dependencies can emerge *without explicit supervision* (Rogers et al., 2020; Tenney et al., 2019). This knowledge can then be transferred to a variety of downstream NLP tasks.

Yet, today’s NLP systems built on large language models still fall short of human-level general understanding (Yogatama et al., 2019; Zhang et al., 2020). Brown et al. (2020) discuss the limi-

Does anybody care?

- While it is definitely a worthwhile endeavor to counteract the hype (and point out that many tests currently overestimate BERT ability), *most* practitioners just welcomed the added performance and relatively ease-of-use of Transformer methods.
- In the last month alone, BERT has been downloaded 22 M times and it's heavily used in social sciences:
 - Best case scenario: for many practical applications, shallow understanding is enough for the target use case.
 - Worst case scenario: we are drawing many real-world conclusion from shaky foundations.



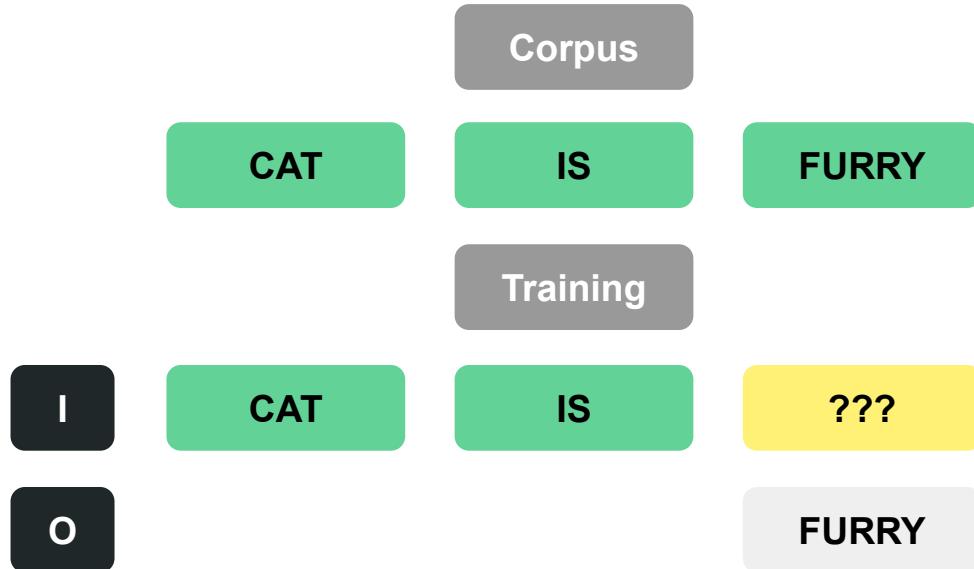
There's more to read - Part 2

- The original [transformer paper](#) and the [annotated transformer](#) tutorial
- [BERTology](#) is a great introduction to recent attempts at explaining BERT behavior
- Testing models on out-of-distribution samples is critical: behavioral testing with [CheckList](#) is a recent, well-thought approach.
- [A fantastic work](#) on the shortcomings of BERT when learning logical reasoning (hint: statistical features fail to generalize)

We need a bigger boat

GPT-3: is scale all you need?

- In 2020, OpenAI trained a new version of their language model, but at unprecedented scale:
 - BERT 345M vs GPT-3 175B params
 - 40GB of English web text available on the internet
- Just training to *predict the next word* gives the model impressive learning abilities with scale.

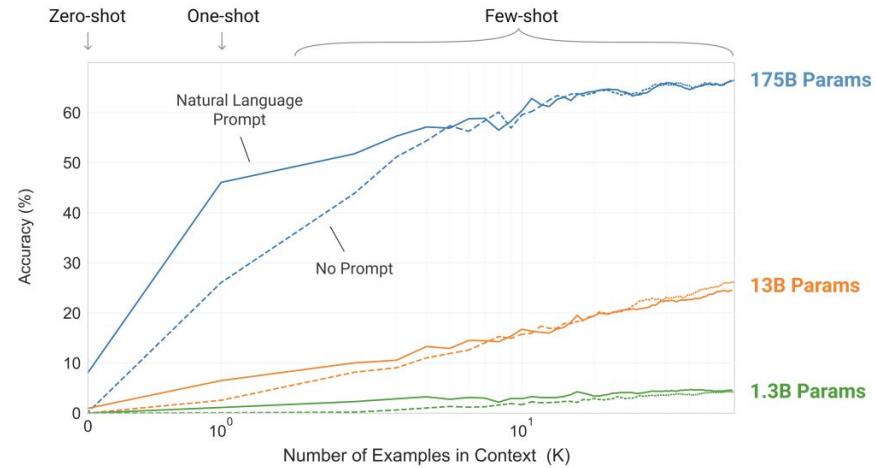


Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

GPT-3: is scale all you need?

- In 2020, OpenAI trained a new version of their language model, but at unprecedented scale:
 - BERT 345M vs GPT-3 175B params
 - 40GB of English web text available on the internet
- Just training to *predict the next word* gives the model impressive learning abilities with scale.



Language Models are Few-Shot Learners

Learning from instructions

BERT

GPT-3

Traditional fine-tuning (not used for GPT-3)

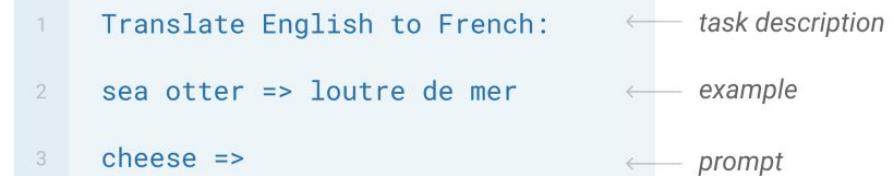
Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Learning from instructions

Convert movie titles into emoji.

Back to the Future: 🎈🕒🚗🕒

Batman: 🦇🦇

Transformers: 🚗🚗

Star Wars:

Submit ⌘ Enter

Submit



Write a restaurant review based on these notes:

Name: The Blue Wharf

Lobster great, noisy, service polite, prices good.

Review:

Submit



New dog, new trick

- Remember the common sense test on multiple sentences that BERT failed? **GPT3 doesn't**

Playground

Load a preset... ▾

Save View code Share ... ⚙

Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her

saw.

New dog, new trick

- Remember the common sense test on multiple sentences that BERT failed? **GPT3 doesn't**

The image shows a screenshot of the Hugging Face Playground interface. The main window has a title "Playground" and contains a text input field with the placeholder "Load a preset...". Below the input field are five buttons: "Save", "View code", "Share", "...", and a gear icon. A large text area displays a generated sentence: "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 shovel." The word "shovel" is highlighted with a green rectangular box. On the left side of the interface, there is a sidebar with a title "Playground" and three buttons: "Load a preset", "Save", and "View code". Below these buttons, there is some partially visible text: "Pablo wan" and "neighbor if". At the bottom of the sidebar, there is a button with the text "saw." followed by a green rectangular box.

New dog, new trick

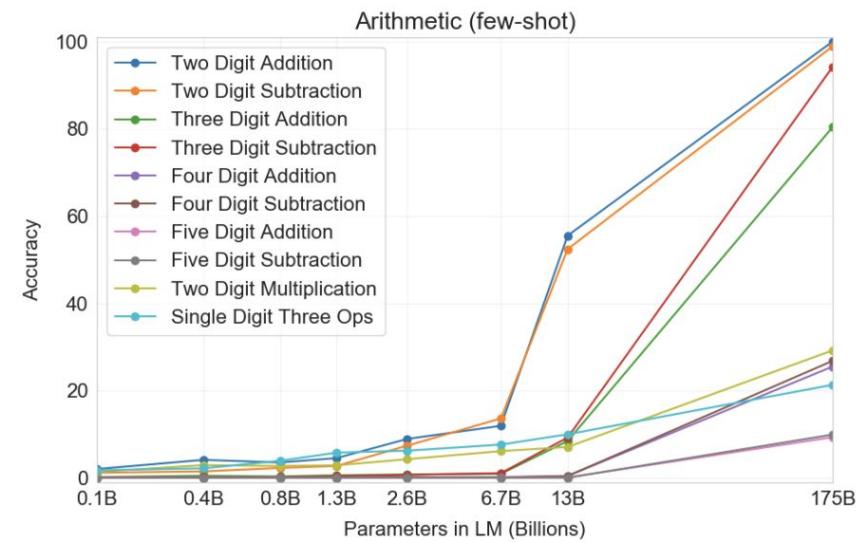
- Remember the superlative test that BERT failed? **GPT3 doesn't**

The image shows a screenshot of the Hugging Face Playground interface. At the top, there is a header with the word "Playground" and several buttons: "Load a preset...", "Save", "View code", "Share", and "...". Below the header, there is a text input area containing the following text:
Anna is worried about the project. Matthew is extremely worried about the project.

Q: Who is least worried about the project? A: Anna

GPT3ology

- GPT3 exhibits *emergent* properties: even if trained *only* to predict the next word, it learns to solve many different tasks without being explicitly taught to do so: e.g.
 - Performing arithmetic operations (“What is $2+2$?”)
 - Translate from English to French
 - Answering factual questions (“what is the capital of Italy?”)
 - Data cleaning



A ladder to the moon?

- As usual, not all that glitters is gold: GPT3 struggles with many common-sense questions, and its impressive performance depends also on the data it is trained on, which at least partially invalidates the claim that “it knows arithmetics” (being addition defined over an infinite set of inputs, generalizing the rule should make training frequency irrelevant)

Biological reasoning

- You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So **you drink it**.

A ladder to the moon?

- The philosophical question therefore remains: as impressive as GPT3 is, aren't we "just" building a bigger ladder (word2vec -> BERT -> GPT3), that **will never get us to the moon?**
- Note that the "radio" argument applies also here: GPT3 has been trained **only on text**, with no reference to the outside world.

Is it possible for language models to achieve language understanding?

I was invited to deliver a few remarks at an HAI symposium on OpenAI's GPT-3 project, at the end of October. I chose for my title "Is it possible for language models to achieve language understanding?" I've had many lively discussions of this question with different groups at Stanford recently, and I am finding that my

A ladder to the moon?

- Yes and no
- Yes, as the philosophical arguments still stand
- No, as, in all fairness, we are *basically scratching the surface* of what large language models can do:
 - Reason “step by step”
 - Learn spatial grounding (left, right...) with few examples
 - Process visual input implicitly
- In particular, **emergent behavior** is surprising, and totally not predictable, **which is in itself an interesting fact that calls for a scientific explanation.**

TL;DR: It is unfair to brush off GPT3 achievements as just a “stochastic parrot”

A ladder to the moon?

- Emergent properties make it very hard to predict the behavior of models as scale increases: “further scaling will likely endow even-larger language models with new emergent abilities”
- Consider the semantic failures for BERT, solved by GPT3, or the failures for GPT3 (**right**), solved by PaLM. Scale may indeed be very surprising!

Camacho-Collados, 2019) shown in Figure 2H, as a historical example. Here, scaling GPT-3 to around $3 \cdot 10^{23}$ training FLOPs (175B parameters) failed to unlock above-random one-shot prompting performance.³ Regarding this negative result, Brown et al. (2020) cited the model architecture of GPT-3 or the use of an autoregressive language modeling objective (rather than using a denoising training objective) as potential reasons, and suggested training a model of comparable size with bidirectional architecture as a remedy. However, later work found that further scaling a decoder-only language model was actually enough to enable above-random performance on this task. As is shown in Figure 2H, scaling PaLM (Chowdhery et al., 2022) from $3 \cdot 10^{23}$ training FLOPs (62B parameters) to $3 \cdot 10^{24}$ training FLOPs (540B parameters) led to a significant jump in performance, without the significant architectural changes suggested by Brown et al. (2020).

A broken dynamics

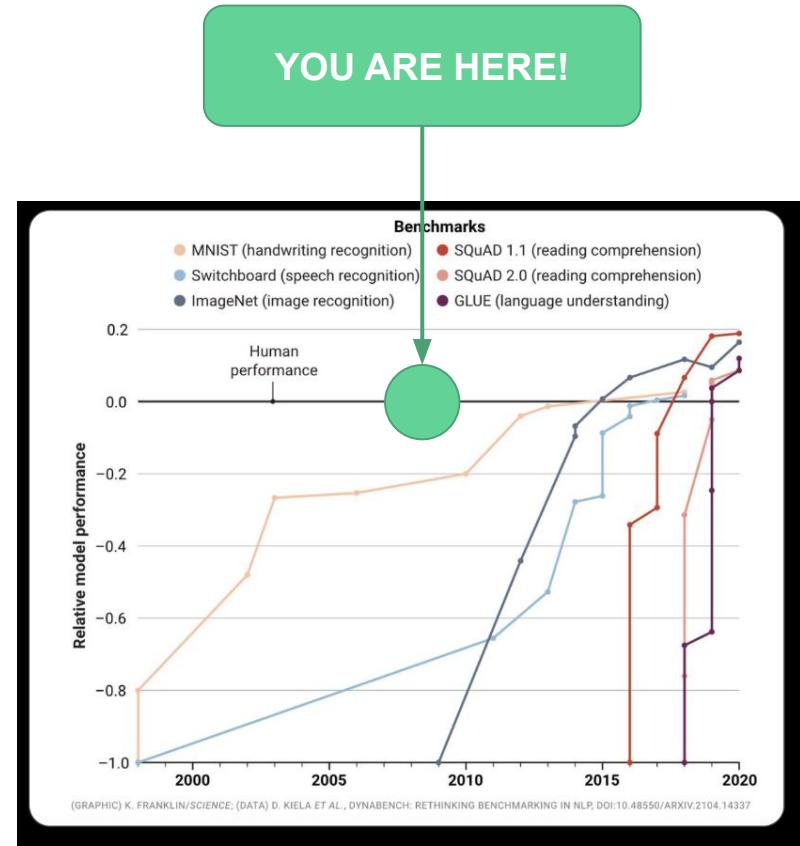
- **The optimist vs the pessimist**
 - Optimist trains (say) BERT and is excited about how many new things it can do, including some “semantic tasks”
 - Pessimist points out that BERT, however impressive, cannot indeed solve example X, so does it really understand?
 - Optimist trains a new model, GPT3, that indeed solves example X, but now pessimist points out example Y, **and so on**

At this rate of progress, **pointing out specific failures is not a safe bet for pessimist**: you can always get refuted in 6 months!

Shifting goalposts in the pessimist camp implies something unsettling: having decent performance *in many cases* may not indeed require semantics at all - that is why it is increasingly hard to find examples where GPTX would fail.

A broken dynamics

- While the pessimist rightly points out that being “better than human” at dataset X likely means X is not useful anymore, the optimist (rightly) replies that benchmarks are to some extent essential to the discipline and, more interestingly, that if we overgrow a test we considered impossible 3 years ago, *that is certainly some progress.*



A broken dynamics

- Chomsky famously argued against sequential models of language (e.g. HMM), as they could not account for long-range dependencies.
- **The same argument is now empirically much weaker**, as syntax seems “within grasp” (esp. vs semantics).
- **Crucially**, nobody - especially pessimists - expected scale to work so well: *this* is a surprising fact that no theory explains.

GPT3 achieved nothing. Zero.



A broken dynamics

- Is it really possible to make *so much practical progress* while solving the **wrong** problem?
- It certainly is (logically) possible, but somehow equivalent of flying a rocket to the moon without learning *anything* interesting about orbits and their physics.
 - Even if we didn't solve language *per se*, it is hard to believe that there is *no lesson* to be learned.

GPT3 achieved nothing. Zero.



A broken dynamics

- The argument that GPT3 learns (almost) everything (“anything goes”) is *correct* as an argument to the effect that transformers inductive biases are *weak* compared to humans.
- The same argument made **after training** is less convincing, as GPT3 has a decent (not perfect!) idea of what is possible and what is not in English (**right**).

Playground

Load a preset... ▾

Save View code Share ... ⚙

Mark the sentences as grammatical or ungrammatical according to English grammar.

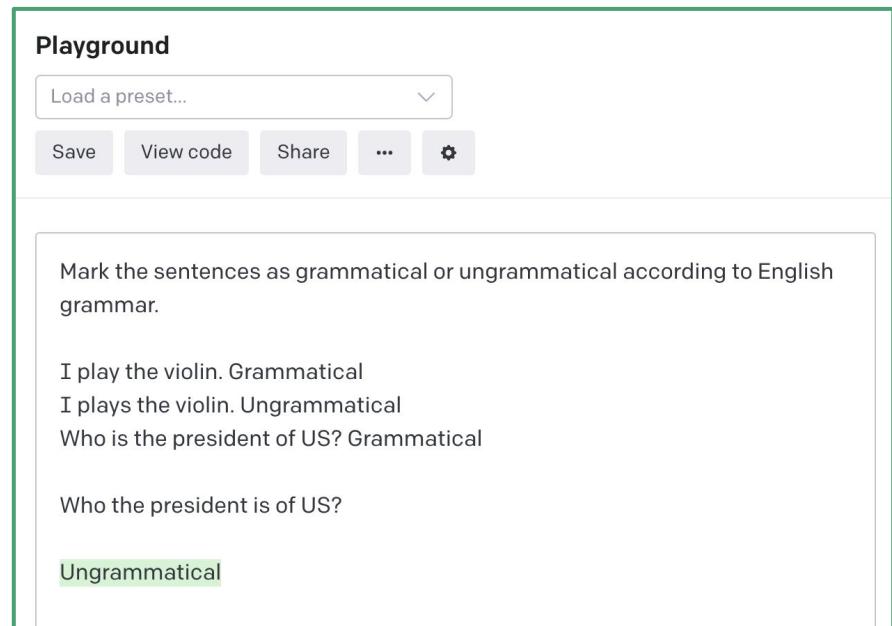
I play the violin. Grammatical

I plays the violin. Ungrammatical

Who is the president of US? Grammatical

Who the president is of US?

Ungrammatical



Isn't language just
much easier to
imitate than what we
previously thought?

There's more to read - Part 3

- A balanced account on the importance of reference for meaning
- Are we under-hyping AI?
- A critical view on GPT3
- Some even more recent large language models: PaLM (540 Bn parameters) and LaMDA for dialogue
- A fantastic overview of large language models across many tasks and architectures, Big Bench
- Our own work on neural networks learning a language through interactions: can meaning emerge implicitly from collective problem solving (e.g. *Convention 1969*)?

A brave new world

A whole new era for fake news

"A photo of Totoro standing bravely in front of a large tank on the road"



"Photograph of Apes attending the World Economic Forum in Davos"



"Photograph of a Banksy graffiti about Totoro holding a flower on a wall in Shinjuku"



It is increasingly hard
to spot AI generated
content!

A Tweet-size Turing
test is now almost
meaningless

What happens now?

- OpenAI and Google recently released new models working with *text and images simultaneously*
- Multi-modality is not just practically useful, but it's conceptually interesting as images provide a natural *reference* for language - **you can't learn Finnish from the radio, but maybe you can from Netflix!**

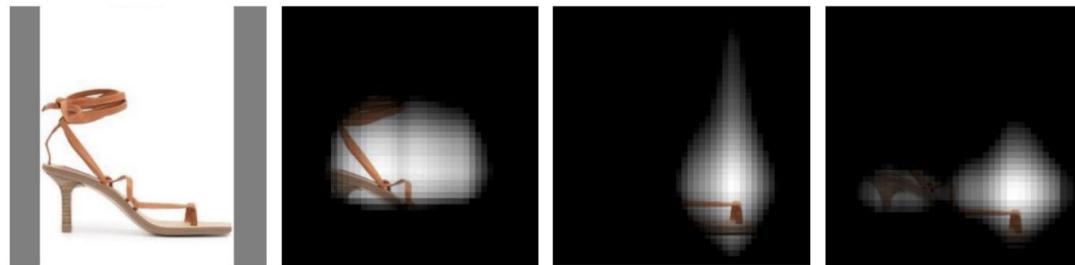


Figure 3: **Grounding and compositionality.** Localization maps for a product retrieved with the query “ankle strap sandals with high heels”: left-to-right, the product, “ankle strap”, “sandals”, “high heels”).

What happens now?

- Multi-modality opens new areas of investigation, and makes possible researching compositionality through image grounding.
- While humans (**right**) are notoriously good at generating new concepts on the fly, “generalization vs memorization” is always a concern in large neural models: what if we could produce controllable inputs that are *by definition* impossible to find in the training data?



Open questions and food for thoughts

- As you may have guessed, the pessimist found examples of failures for multi-modality as well!
- We saw that testing in BERT often overestimated the model; larger models are even more complex as slight changes in input would drastically change their behavior (of course, this may in itself be an argument against *understanding*). **Better testing is crucial.**
- BERT is mildly expensive to train, but it's nothing compared to the tens of M of USD required by GPT3: research on large language models is increasingly hard to do and impossible to replicate / validate outside few labs.

“An odd number of apples”



Open questions and food for thoughts

- How large language models compare to formal semantics? If multi-modal **grounding** is at least a first attempt at “referential” knowledge, **compositionality** is still a very active area of research .
- Further areas to explore are **pragmatics** and **language acquisition**: babies learn language in a very different way (note however, that babies are also *not* doing Montague grammar)! Which generally points to the question we have been avoiding all along: is there **one** notion of “*understanding a language*”?
- If we consider “referential semantics” as the abstract ideal as far as entailments / truth conditions etc. go (in the same sense as logic is ideal reasoning implemented in faulty systems), is it conceivable that meaning gets embodied differently in different systems (a brain vs a neural network)?

If so, will we achieve understanding multiple times in evolution and through multiple paths, as it happened with the bat vs the eagle wing?

See you, space cowboys