GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

NLP and Language Models

Introduction

- Feel free to stop me at any time with questions!

- Quick survey:
  - Used Google Translate before?
  - Talked to your phone (Siri, etc.) to set an alarm
  - Heard of natural language processing (NLP)
  - Done any kind of study of machine learning (read a book, tutorial, course, worked through a Jupyter notebook, etc.)

What is NLP?

Natural Language Processing

human languages
(not computer languages)

doing things with them
automatically!

How do we build Siri?

How do we build Google Translate?

The Political Bureau of the CPC Central Committee

July 30

hold a meeting

中共中央政治局7月30日召开会议，会议分析研究当前经济形势，部署下半年经济工作。

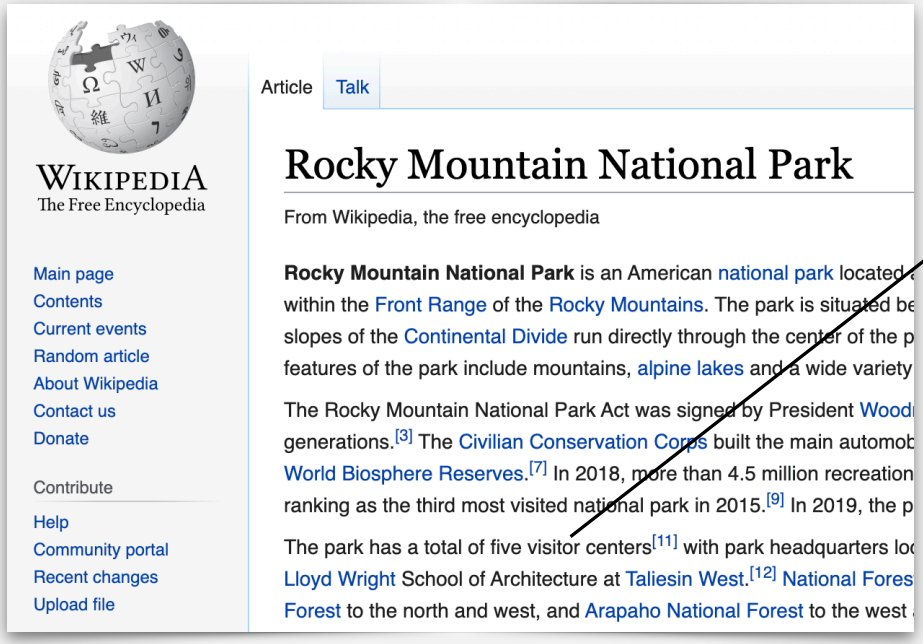
People's Daily, August 10, 2020

Translate

The Political Bureau of the CPC Central Committee held a meeting on July 30 to analyze and study the current economic situation and plan economic work in the second half of the year.

How do we answer questions from Google?

How many visitors centers are there in Rocky Mountain National Park?



The park has a total of five visitor centers

↓

five

You can’t do this with string processing!

Goals for today

- Learn what NLP is about
- Learn some basic ideas of machine learning: fitting a statistical model to examples of a problem we want to solve to learn how to solve that problem
- See how a statistical model for predictive text works (what word should come next in this sentence?)
- Learn the connections between this language model and state-of-the-art systems such as Google's BERT and OpenAI's GPT-3 models

Outline

- Machine Learning
- Language Modeling
- Building n-gram Language Models
- **Hands-on** with Transformer Language Models / Write With Transformer / GPT-3
- LMs in the News, What's Next

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Machine Learning

Machine Learning

- Natural language processing uses a lot of ideas from *machine learning*

- Humans are good at understanding language. Computers are bad at it and it's hard to program them. If we see lots of examples of how humans do a task, can we teach a computer how to do it?

Building Siri

Imagine trying to build Siri…

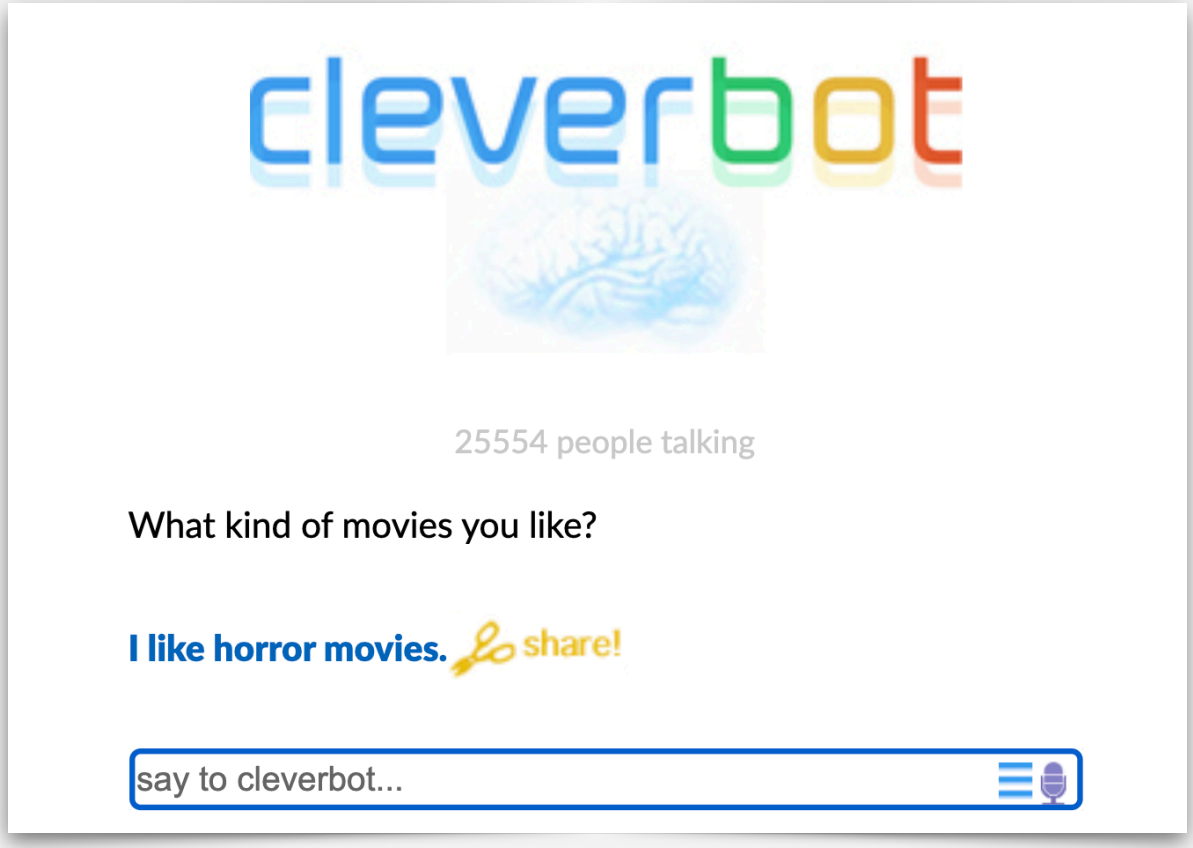
```
// Start by reading the user input with a predefined method
String userStr = readUserInput();
if (userStr.startsWith("set a timer"))
  startTimerDialogue();
else if (userStr.startsWith("set an alarm") ||
         userStr.startsWith("wake me up at"))
  startAlarmDialogue();
else [...]
```

Too hard to list every case here!

Building Siri with Machine Learning

Type in a question

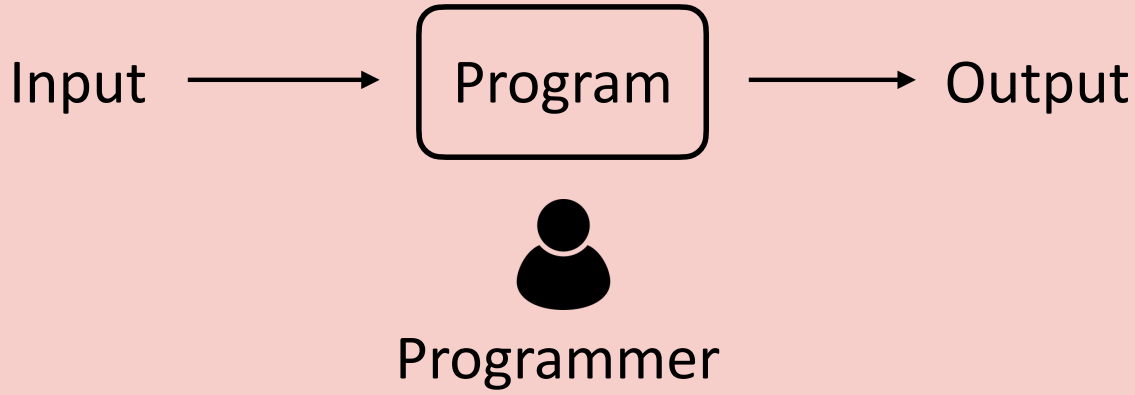
Cleverbot responds with something a human said before in response to that question



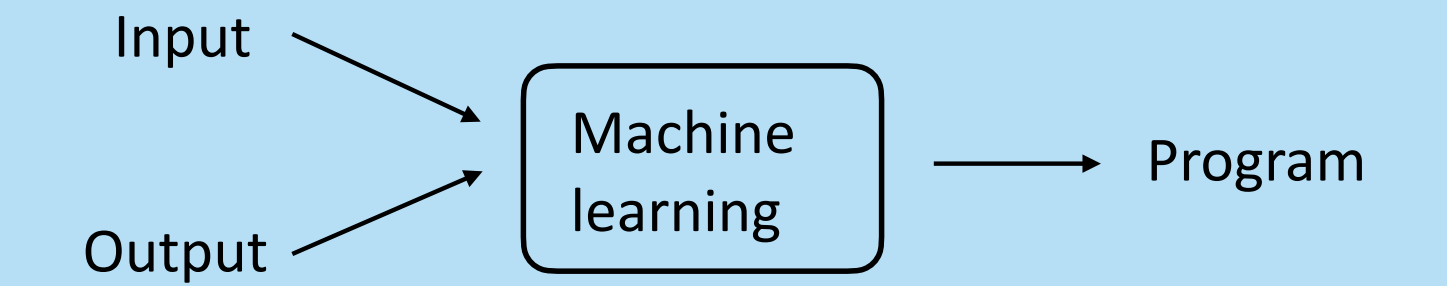
cleverbot.com

Machine Learning

Programming (as you've probably seen it)



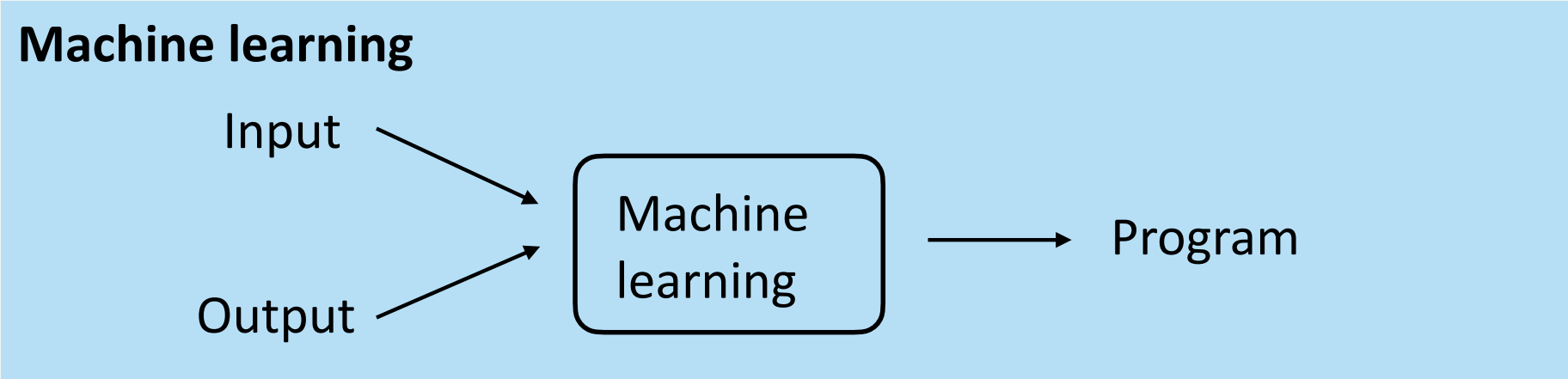
Machine learning



(examples of what we want to do)

Cleverbot (for example)

Example: Sentiment Analysis



Data: Rotten Tomatoes

Avengers: Infinity War ably juggles a dizzying array of MCU heroes in the fight against their gravest threat yet, and the result is a thrilling, emotionally resonant blockbuster that (mostly) realizes its gargantuan ambitions.

★ ★ ★ ★

Star Wars: The Rise of Skywalker suffers from a frustrating lack of imagination, but concludes this beloved saga with fan-focused devotion.

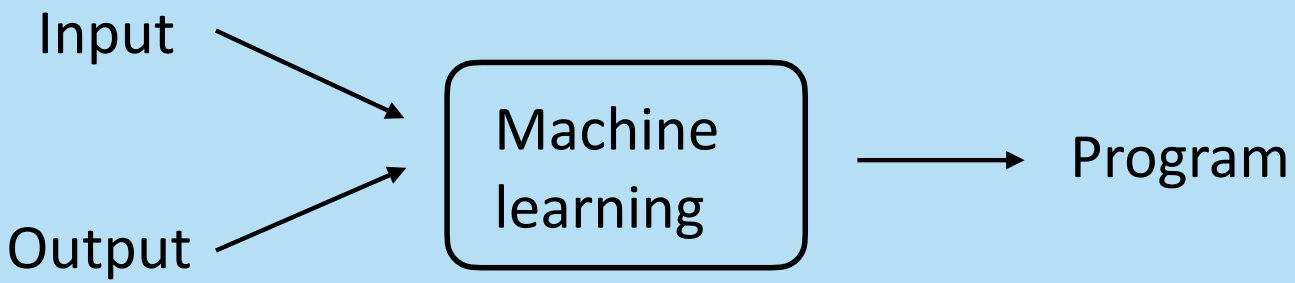
★ ★

- Machine learning starts with a feature representation of this data — how do we represent it to a system?
- Neural network methods will view this as thousands of numbers associated with each word. But we can start with something simpler:

```
(numberOfGoodWords, numberOfBadWords)
```

Example: Sentiment Analysis

Machine learning



Data: Rotten Tomatoes

Avengers: Infinity War ably juggles a dizzying array of MCU heroes in the fight against their gravest threat yet, and the result is a thrilling, emotionally resonant blockbuster that (mostly) realizes its gargantuan ambitions.

★ ★ ★ ★

Star Wars: The Rise of Skywalker suffers from a frustrating lack of imagination, but concludes this beloved saga with fan-focused devotion.

★ ★

```
int numberOfGoodWords = computeNumGoodWords(review);
int numberOfBadWords = computeNumBadWords(review);
if (numberOfGoodWords > 3 && numberOfBadWords < 2)
  return “4 stars”;
else if (numberOfGoodWords > 2 && numberOfBadWords < 3)
  return “3 stars”;
else [...]
```

We can *automatically* learn to generate this! (It’s called a *decision tree*)

Machine Learning

- Lots of different models: decision trees, neural networks, Bayes Networks, …

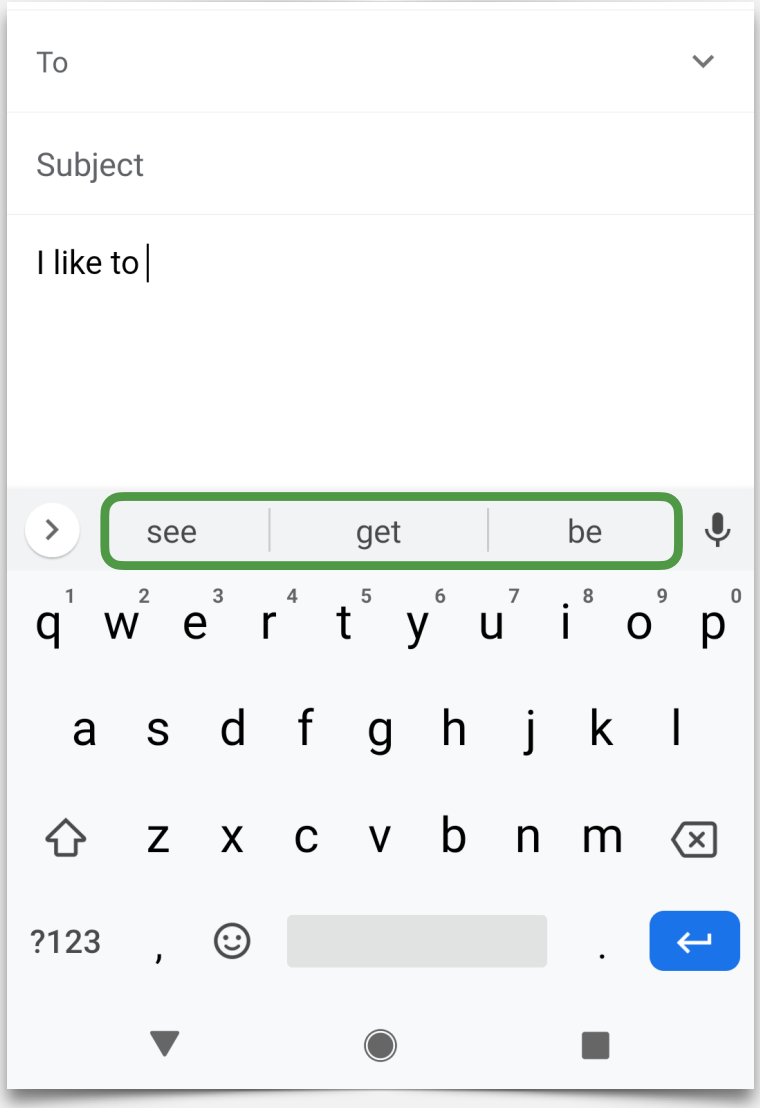
- We’re going to use a probabilistic model for language modeling, which won’t require a ton of math to implement

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Language Modeling

Language Modeling

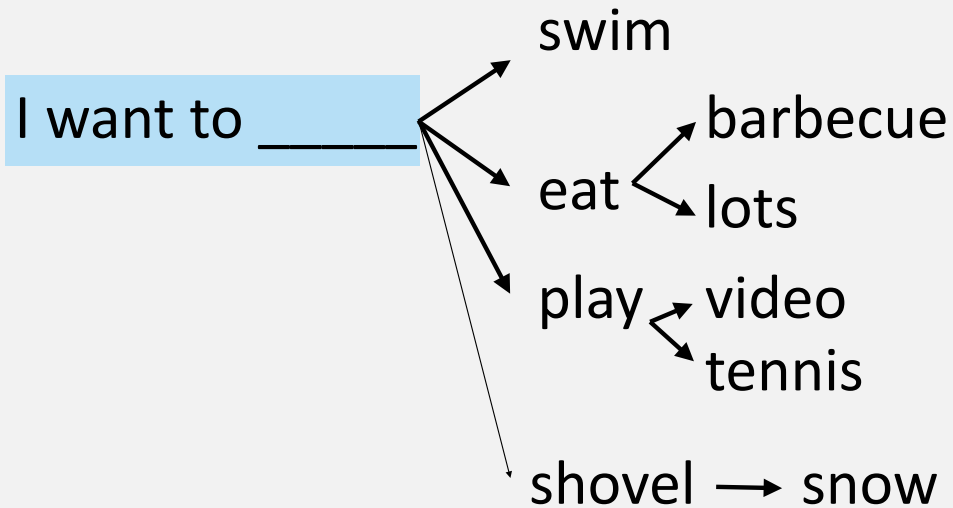
- Our task: build a language model
- Given a sequence of words so far (the **context**), predict what comes next, like in predictive text!
- We never know for sure what comes next, but we can still make good guesses!



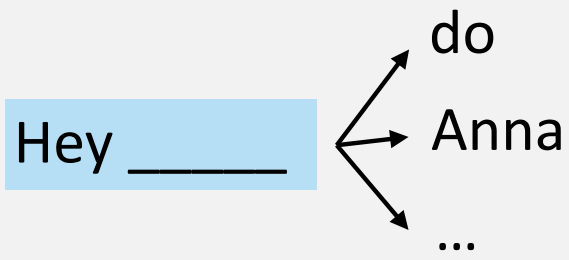
Exercises: How to predict the next word?

1. Suppose we have the context "I want to ____". Lots of words can come next and form sensible sentences. Think about a few words that can come next; what do these have in common?

Language Modeling



grammatical
but less likely

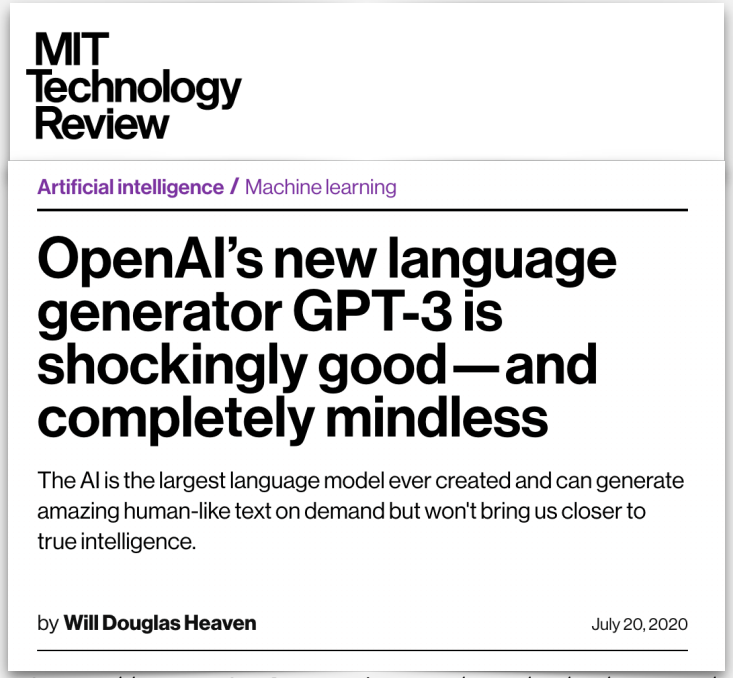


The capital of Nebraska is _____ Lincoln

one right answer but
may be hard to predict!

Why Language Modeling?

It turns out these systems can do a lot! We'll talk more about this later.



MIT Technology Review

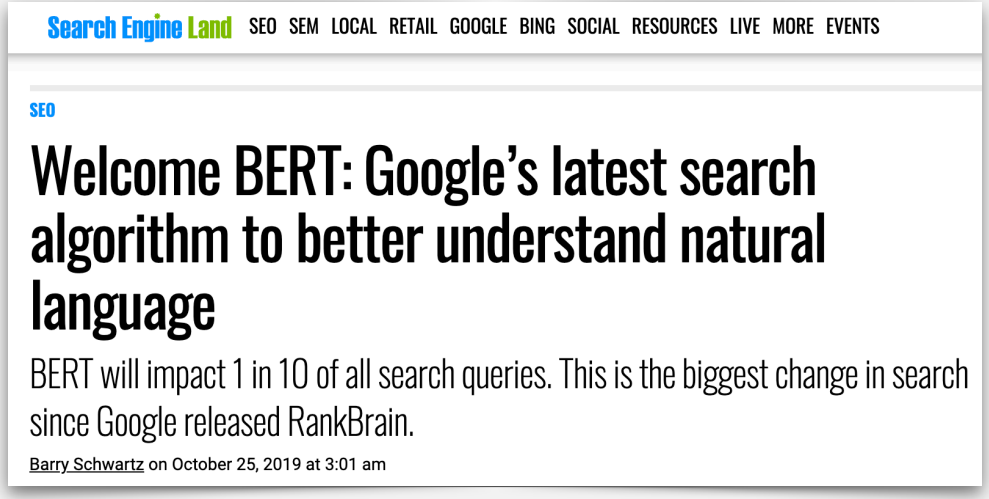
Artificial intelligence / Machine learning

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

by **Will Douglas Heaven** July 20, 2020

https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/



Search Engine Land SEO SEM LOCAL RETAIL GOOGLE BING SOCIAL RESOURCES LIVE MORE EVENTS

SEO

Welcome BERT: Google's latest search algorithm to better understand natural language

BERT will impact 1 in 10 of all search queries. This is the biggest change in search since Google released RankBrain.

Barry Schwartz on October 25, 2019 at 3:01 am

https://searchengineland.com/welcome-bert-google-artificial-intelligence-for-understanding-search-queries-323976

N-gram Language Modeling

- Our focus: build a model that predicts the next word based on the previous one or two words

- *n*-gram: a sequence of *n* words

  I like to = 3-gram

  I really want to go = 5-gram

- *n*-gram language model: predict the next word based on the previous *n*-1 words

  I want to _____ $\xrightarrow[\text{2-gram}]{}$ to _____

2. Suppose we have the context "to ____". How does this change the predictions of the next word compared to "I want to ____"?

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Building n-gram LMs

2-gram Language Model

- We are going to learn a 2-gram (bigram) language model

- This is **a conditional probability distribution:**
  P(next word = *y* | previous word = *x*)

  "the probability of the next word is *y given that* the previous word is *x*"

P(next word = *Austin* | previous word = *to*) = 0.2

"if we see *to* I think there's a 20% chance the next word is *Austin*"

P(next word = *Europe* | previous word = *to*) = 0.1
P(next word = *Mexico* | previous word = *to*) = 0.1
P(next word = *eat* | previous word = *to*) = 0.1

…

P(next word = *was* | previous word = *to*) = 0.0

These have to add up to 1 over the *vocabulary* (every possible word *y* could be)

Assume a **fixed vocabulary** of ~30,000 words

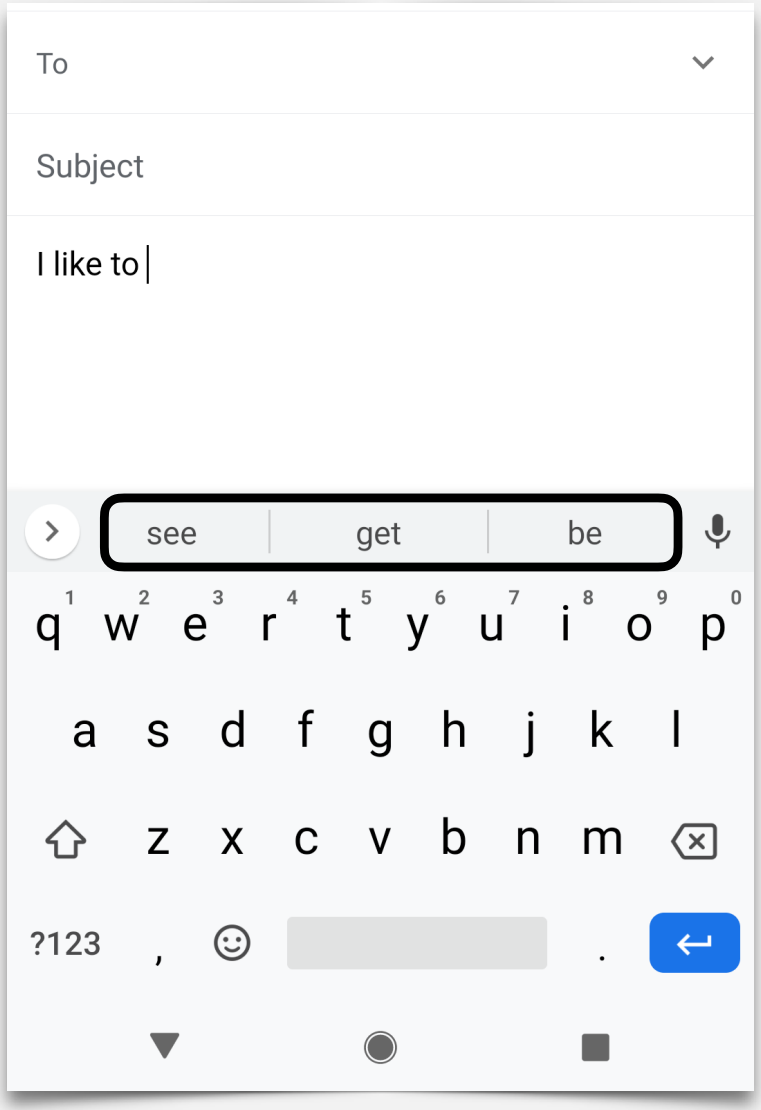
2-gram Language Model

- If we have these probabilities, we can build our predictive text system

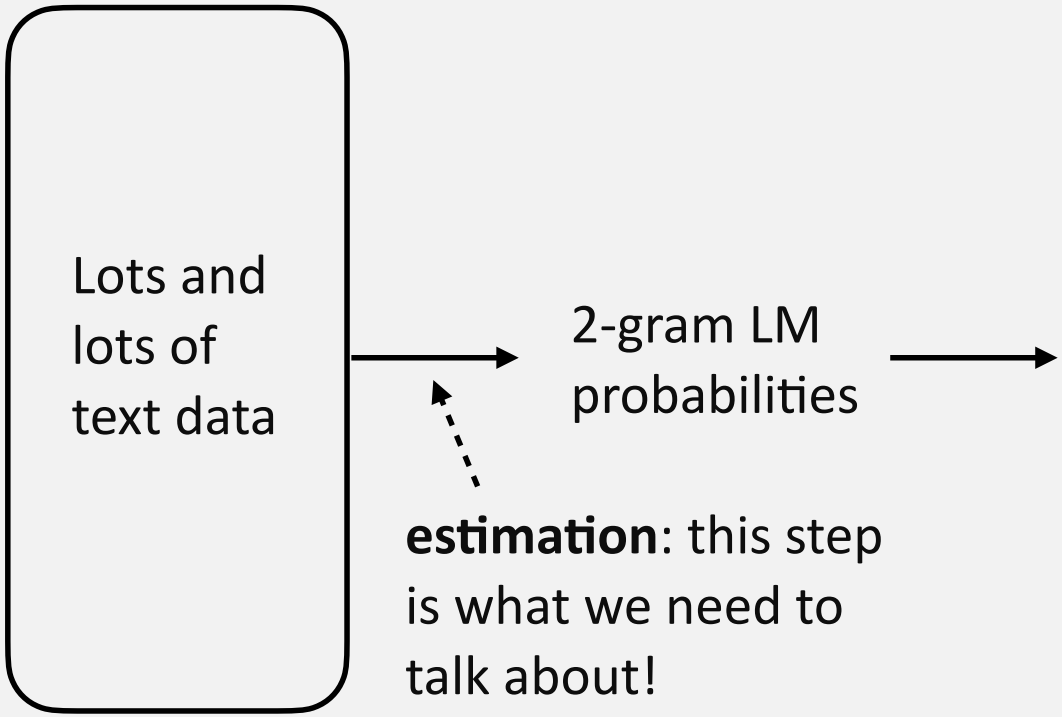
P(next word = _ | previous word = *to*)

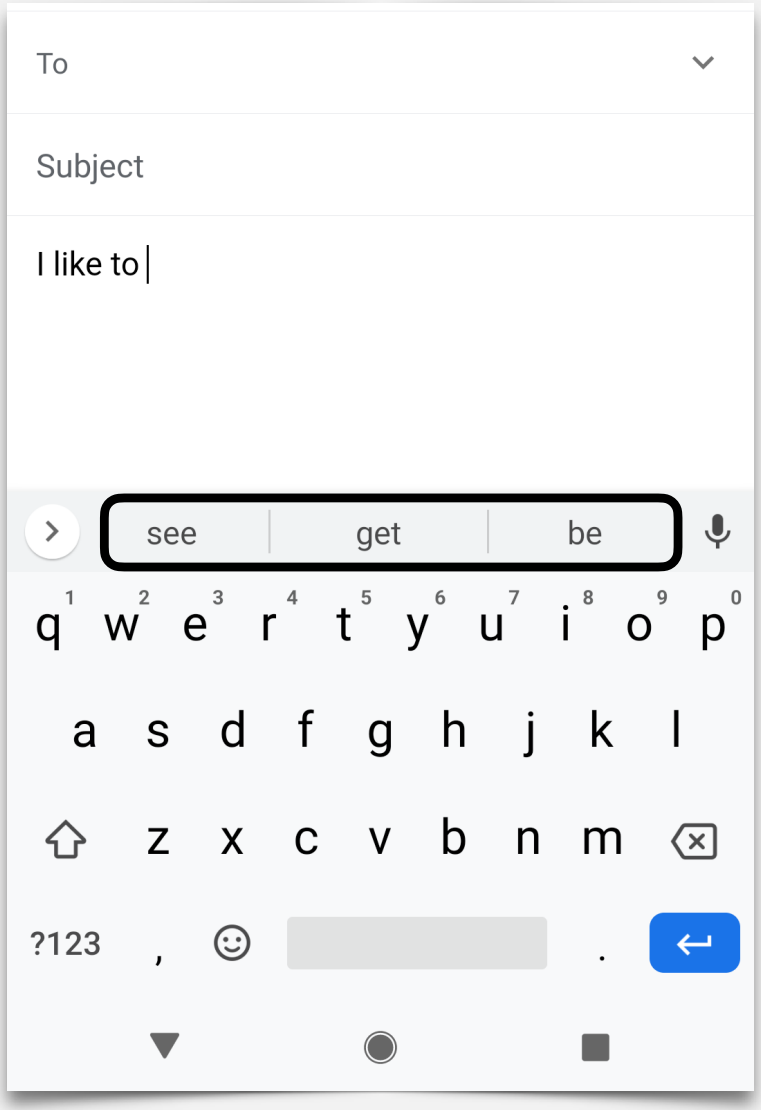
Check all the possible words from that list, pick the ones with the highest probability (most likely next words)

- Where do these probabilities come from? We're going to *learn them* from a bunch of text data we see



2-gram Language Model





Estimating Probabilities of Events

Suppose we have a *biased* coin that's heads with probability *p. p* is a number between 0 and 1, and for a normal coin, *p* = 0.5 (equal probability of heads or tails).
Suppose we flip the coin four times and see (H, H, H, T)

1. What do you think the probability *p* of heads is with this coin? Take a guess!

- We don't know what *p* is — *p* could be 0.5! But *p* = 3/4 = 0.75 *maximizes the probability of the data*. We'll say "this is the most likely value of *p*"
- The probability of the data is *p***p***p**(1-*p*) — if you've taken calculus, you can take the *derivative* and set it equal to zero and find *p* = 0.75

N-gram LM

- The decision for what words occur after a word w is exactly the same as the biased coin, but with 33,000 possible outcomes (different words) instead of 2.

I like to **eat** cake but I want to **eat** pizza right now. Mary told her brother to **eat** pizza too.

P(next word = *pizza* | previous word = *eat*) = 2/3

P(next word = *cake* | previous word = *eat*) = 1/3

All other next words = 0 probability

$$P(w \mid w_{\text{prev}}) = \frac{\text{count}(w_{\text{prev}}, w)}{\text{count}(w_{\text{prev}})}$$

how many times do you see w_{prev} followed by w?

how many times do you see w_{prev}

Smoothing

I like to **eat** cake but I want to **eat** pizza right now. Mary told her brother to **eat** pizza too.

P(next word = *pizza* | previous word = *eat*) = 2/3

P(next word = *cake* | previous word = *eat*) = 1/3

All other next words = 0 probability

- All other words 0 probability isn't right! We want to assign some small probability to all of the words

- We want to *smooth* the distribution from our counts

$$P(w \mid w_{\text{prev}}) = \lambda \frac{\text{count}(w_{\text{prev}}, w)}{\text{count}(w_{\text{prev}})} + (1 - \lambda) \frac{\text{count}(w)}{\text{total word count}}$$

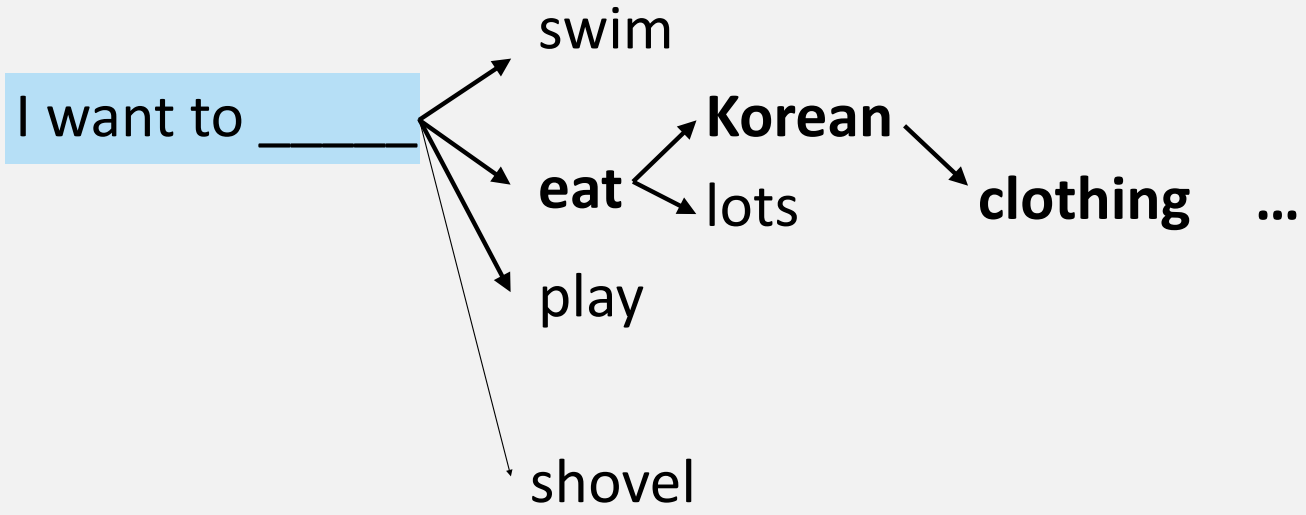
a number between 0 and 1 (like 0.9)

what we had before

a *unigram* LM

Using the Model

- Now we have the distribution $P(w \mid w_{\text{prev}})$ (from a table of 33,000^2 numbers)
- For predictive text: find the **most likely** next words (highest probability)
- Can also find the **most likely completion** of the sentence, or **sample** a random sentence from this model



Up next

- **You have everything you need to build an n-gram language model!**
- n-gram models work well at certain things (if you use n>5), but have been replaced with neural networks recently
- **Transformer** language models: same idea but with powerful neural network models

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Hands-on: Write With Transformer

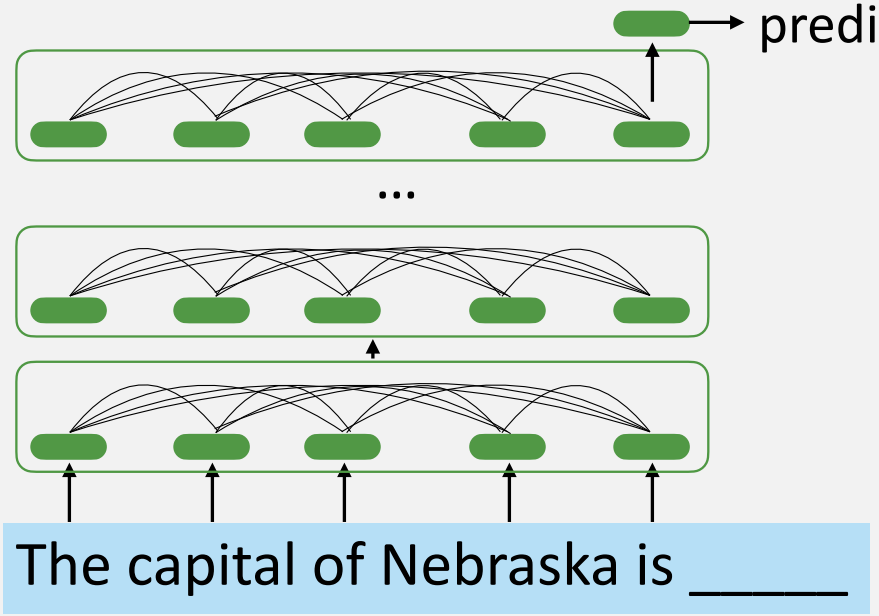
Neural Network Language Models

- Neural networks are function that map *f*(context) → prediction
- *f* is very, very complicated!

  f(*x*) = 2x+3 has one input (*x*) and 2 parameters (2 and 3)

  The *f* we use here has >1000 inputs and >1 million parameters!

- These can be learned from data using derivatives from calculus

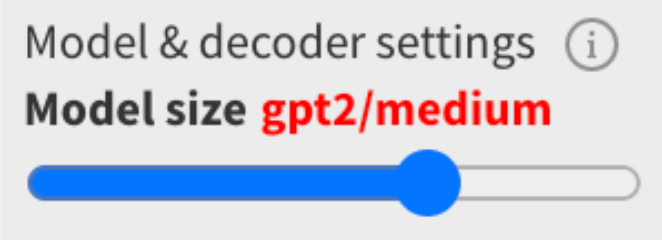


This model is called a Transformer. Uses a mechanism called self *attention* to capture interactions between words

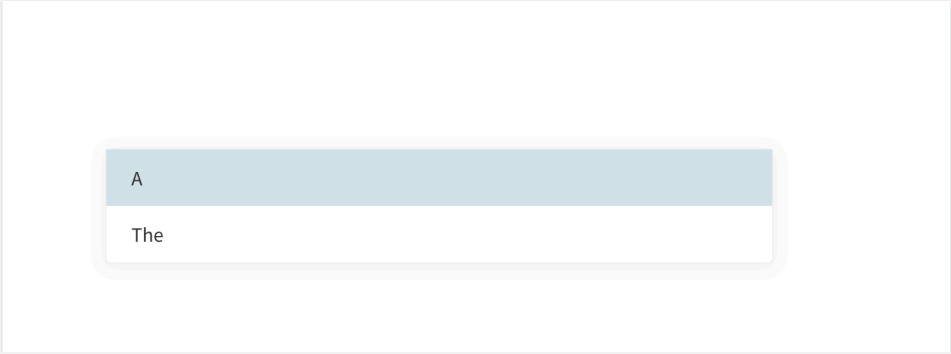
1000 numbers x 5 words = 5000 inputs

Activity

- Go to URL for Write With Transformer
  - Google it, click “Start writing” under DistilGPT2
- Change the slider at the bottom to use the “gpt2/medium” model (not possible on mobile



- Delete what’s in the text area and press *Tab* to complete after any prefix (including an empty prefix):



Activity

1. Write some different sentences to get a feel for the tool. Try completing short snippets or longer snippets

2. Try some sentences involving specific *entities.* Write a sentence about a city, country, or celebrity and see what happens. Are any of the completions specific to that thing?

3. The simple *n*-gram models only depended on the past few words. Is this true for the Transformers? Try something like "*At the X, I really wanted to* ___" and then use tab completion — how much does X matter?

Try to find one that you want to share!

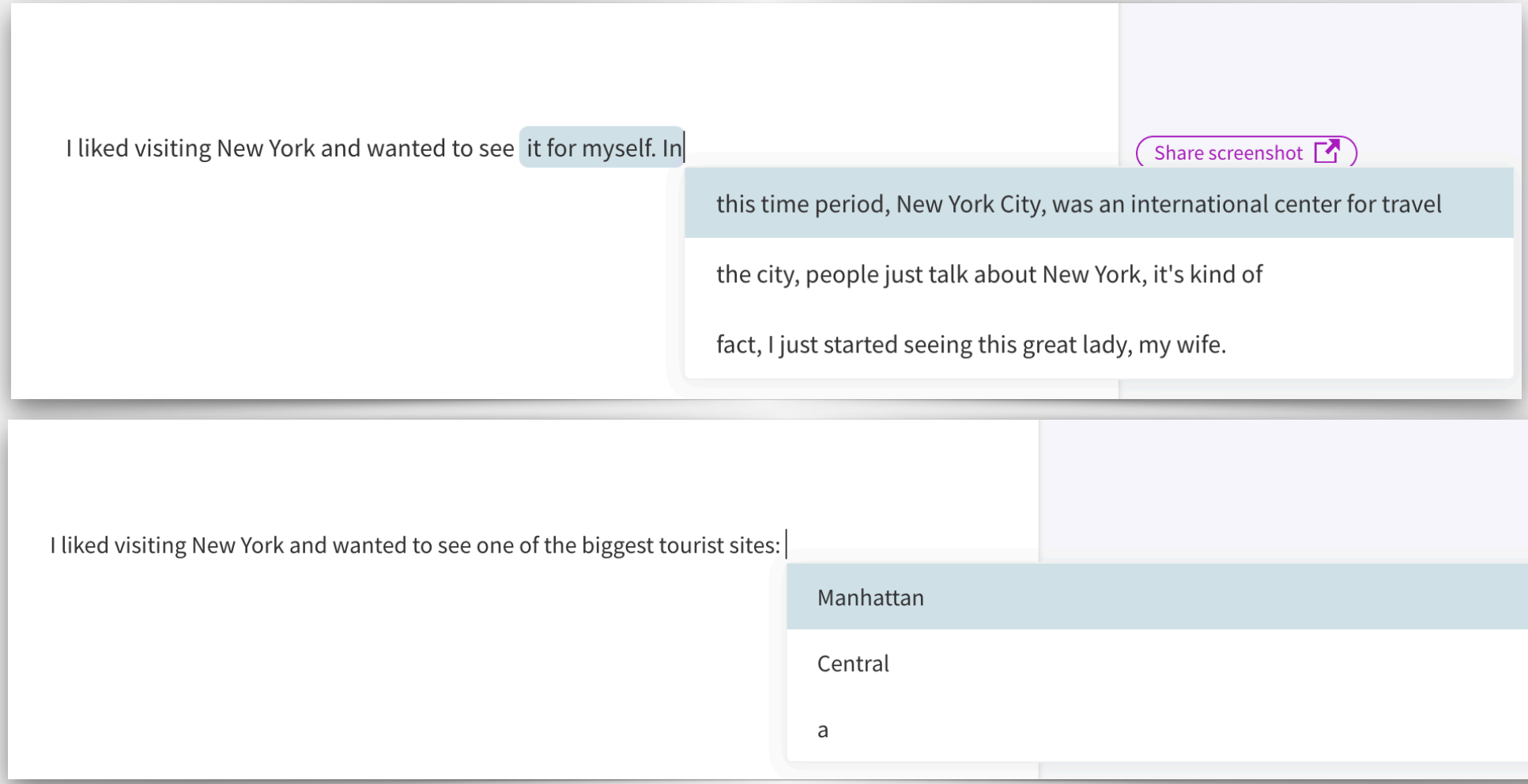
GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Discussion: Write With Transformer

DISCUSSION: WRITE WITH TRANSFORMER

Assessing the Model

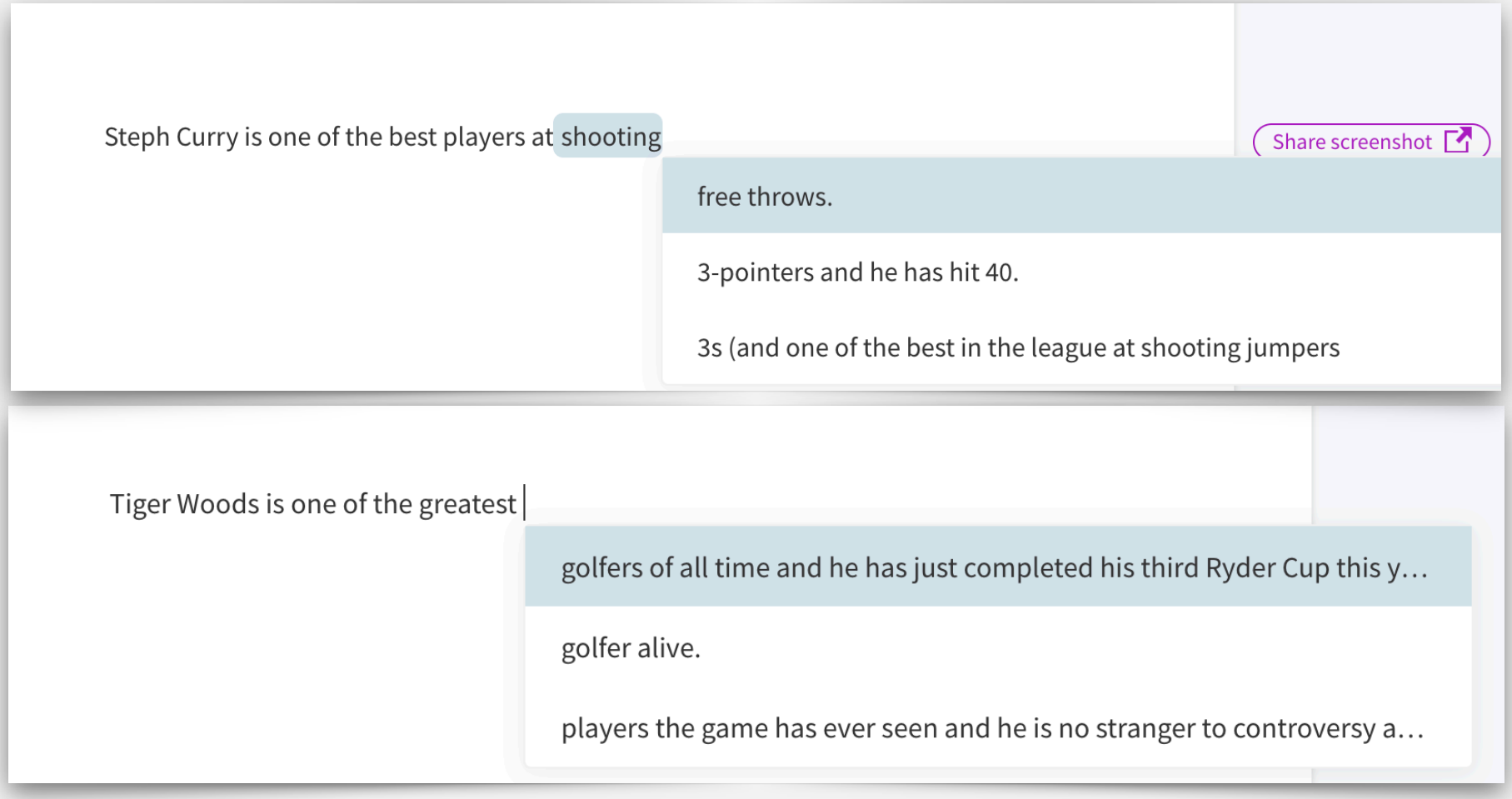
- What does the model say about New York?



DISCUSSION: WRITE WITH TRANSFORMER

Assessing the Model

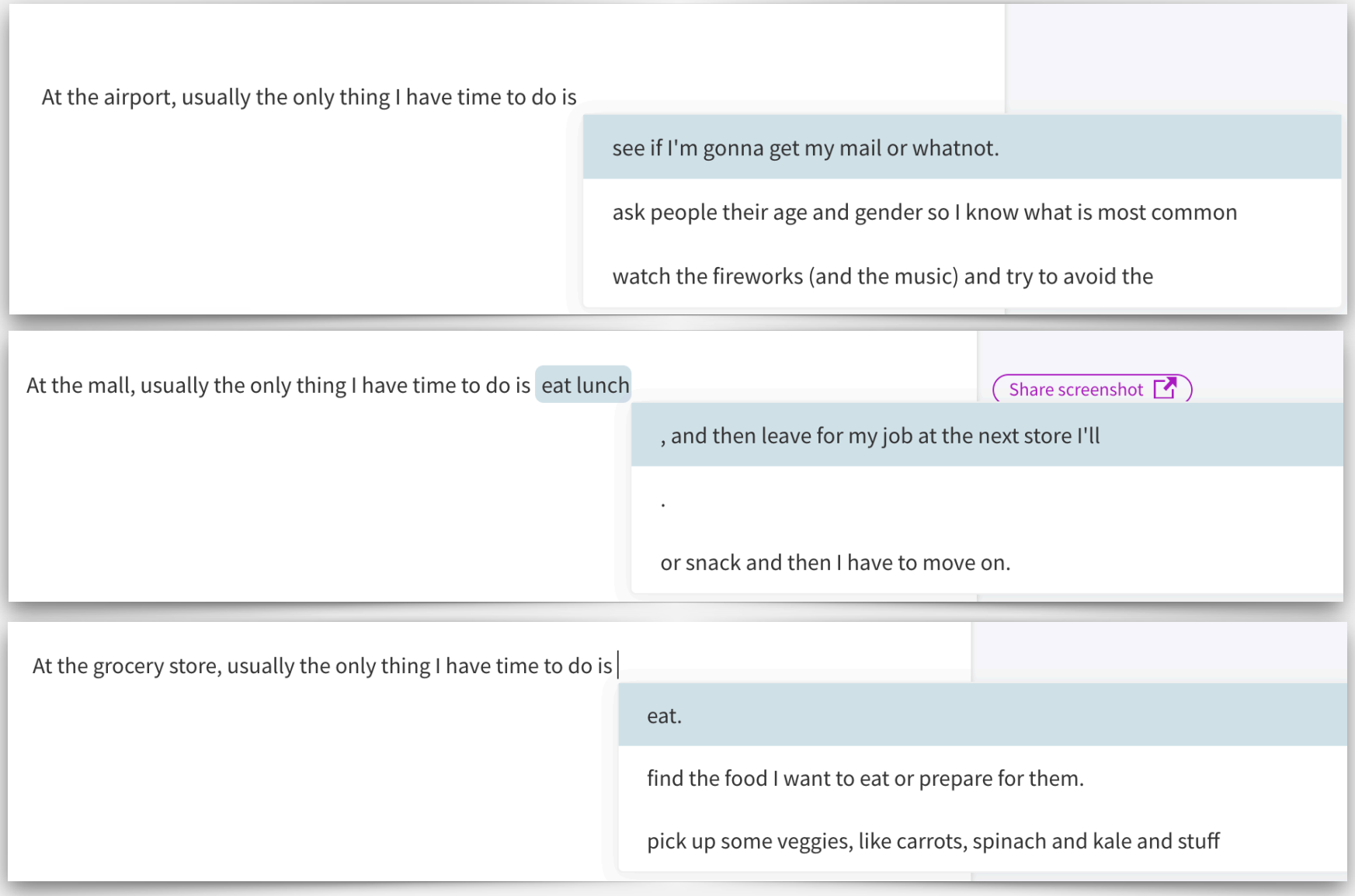
- Sports celebrities:



DISCUSSION: WRITE WITH TRANSFORMER

Assessing the Model

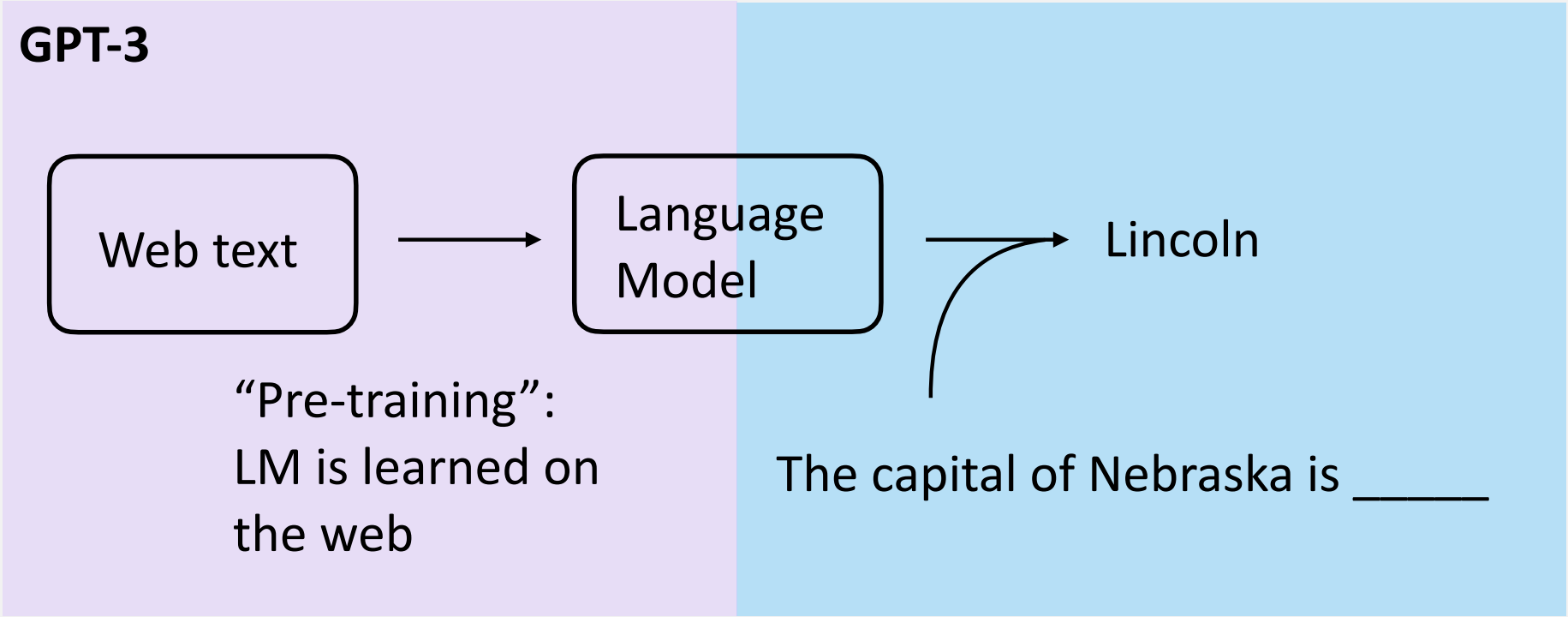
- Sometimes it uses the context well sometimes not!



GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

GPT-3

Using Large Language Models



- These models are trained over a ton of data (a curated scrape of the web). So they will have seen information about Nebraska and Lincoln.
- A big enough model can answer questions **even without being trained to do so**. What else can we get these models to do?

Let’s try it out…

Why does this work?

(1) **Jay Sherman was a film critic in this city in the television show *The Critic*. This city is the birthplace of Hank Hill. In this city, Homer Simpson chooses to drink crab juice instead of Mountain Dew while waiting for a parking officer. Residents of this city worship an unexploded nuclear bomb and tell the legend of El Chupanibre, and it is home to Panucci's Pizza and Applied Cryogenics. The Simpsons see the musical *Kickin' It* in this city, whose future**(*) Madison Cube Garden houses the Harlem Globetrotters. For ten points, name this setting of Futurama in which Homer's car was booted on the plaza of the World Trade Center.

ANSWER: New York City (accept Old New York or New New York)

https://quizbowlpackets.com/

Fill In The Blanks for Category: present_ar_verbs_1
Fill in the blank with the best option that completes each sentence.

1) El Sol _____ en el signo de Piscis.
(entre, entra)

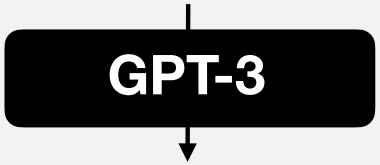
https://www.123teachme.com/spanish_worksheets/list_all

- The model has really seen how to do a lot of tasks already in pre-training

Limits

Context: *Crestfallen is a track on The Smashing Pumpkins' album, Adore. The single's artwork is by Yelena Yemchuk. Johnny McDaid is a Croatian professional photographer.* ***Yelena Yemchuk is a Ukrainian professional photographer.***

Q: Crestfallen's artwork is done by a photographer of which nationality?



First, Crestfallen's artwork is done by Yelena Yemchuk. Second, ***Yelena Yemchuk is a Croatian professional photographer****. The answer is* ***Croatian*** !

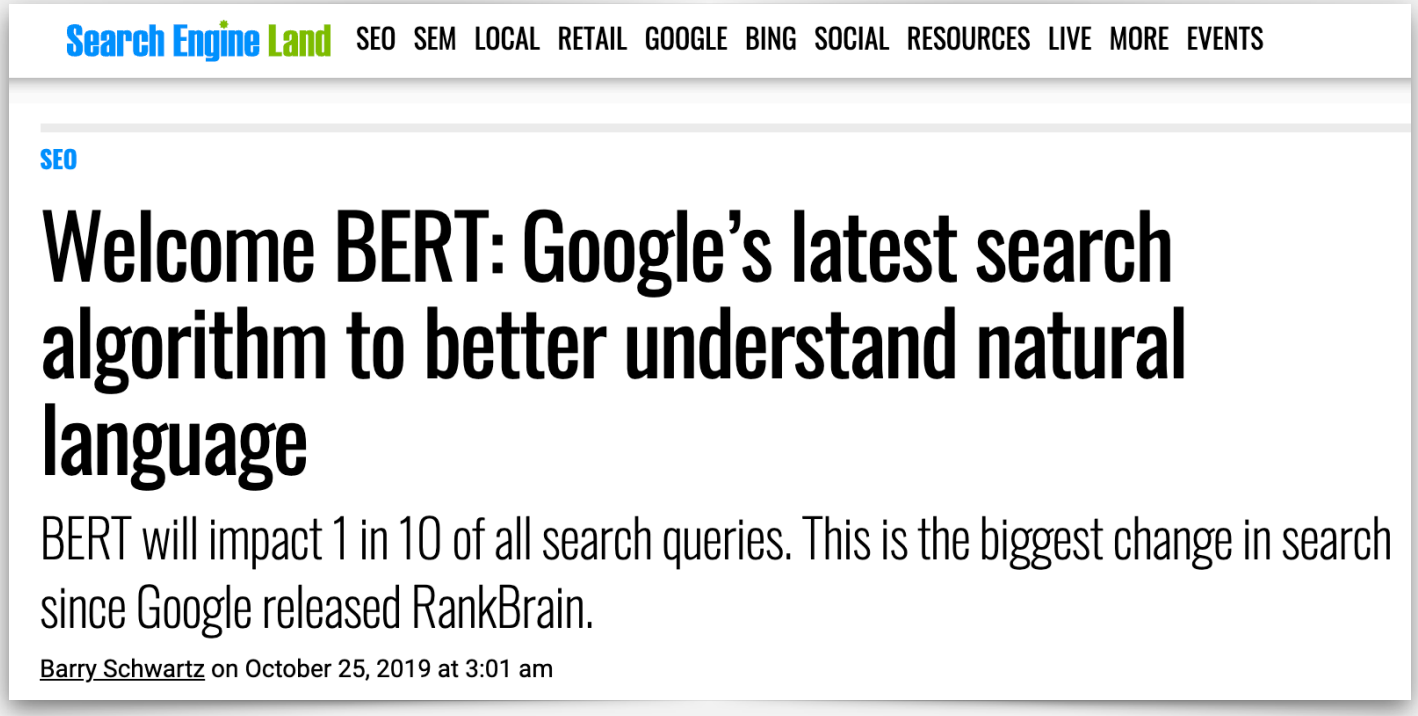
- In this QA example, we asked the model to explain its reasoning, but it produces something inconsistent with the context
- The model isn't always right, and sometimes it can deceive its users!

Xi Ye and Greg Durrett, ***The Unreliability of Explanations in Few-Shot In-Context Learning***

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

LMs in the News

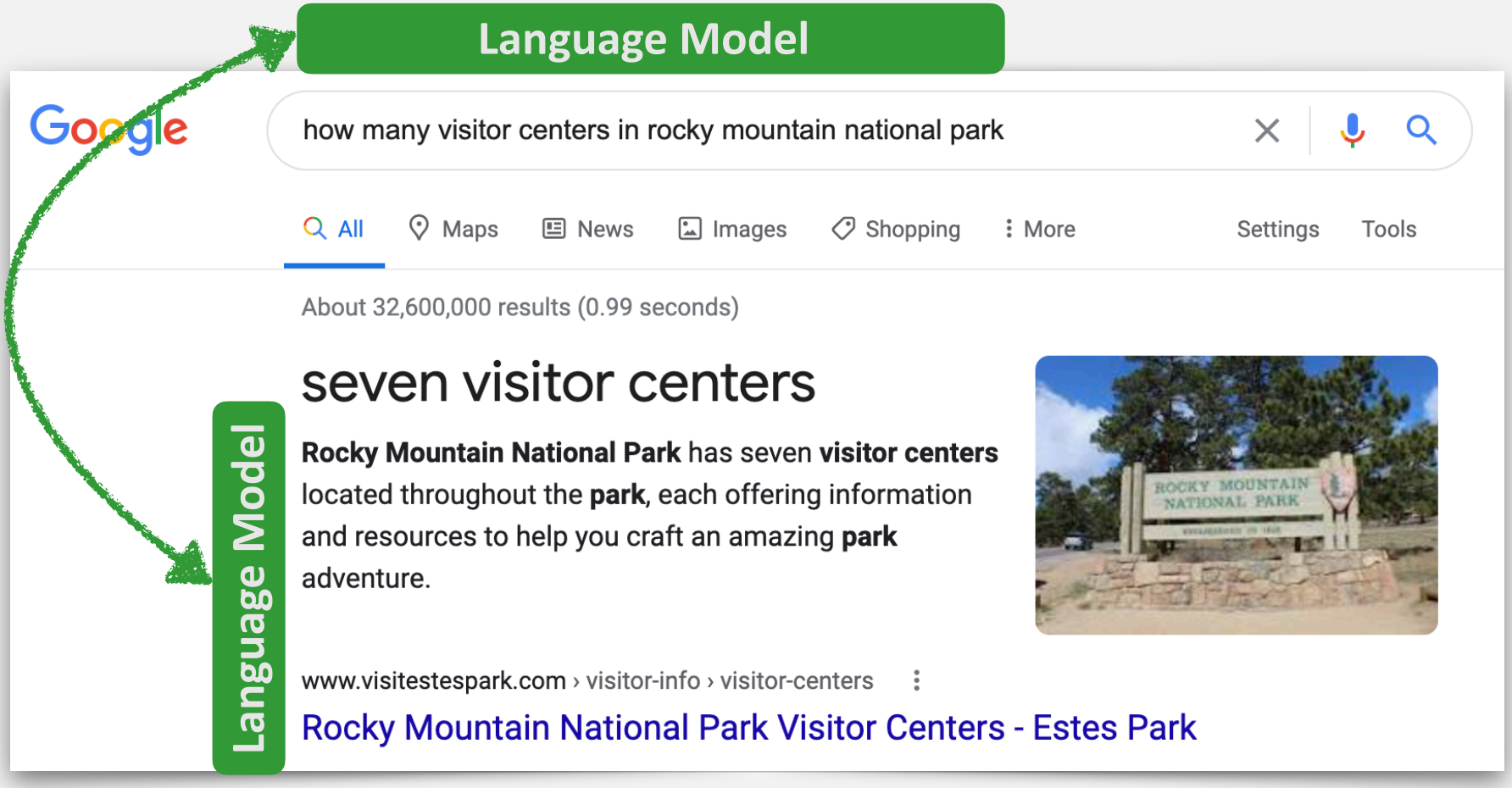
BERT in Google Search



https://searchengineland.com/welcome-bert-google-artificial-intelligence-for-understanding-search-queries-323976

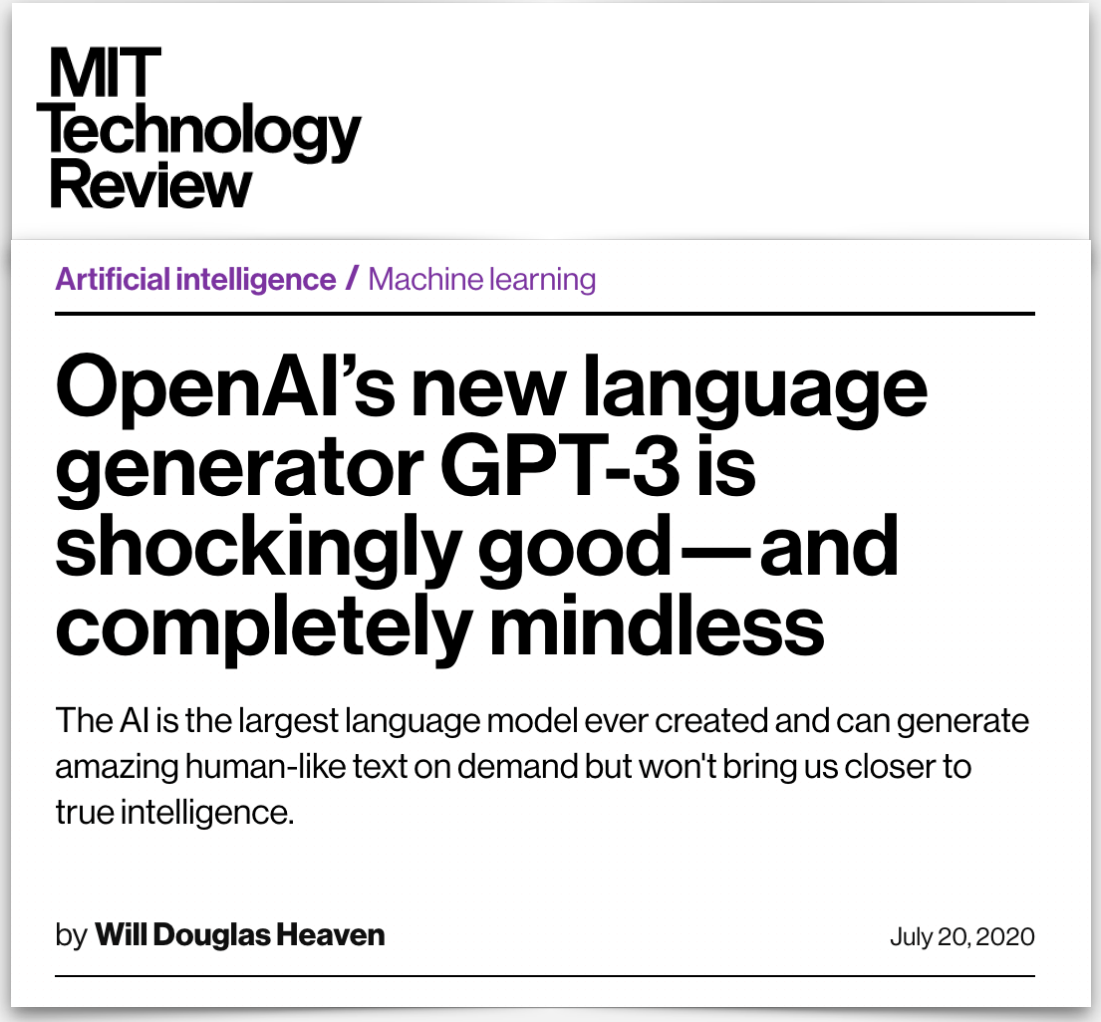
- BERT is a language model from Google very similar to what we've seen so far

BERT in Google Search



- Google compares LM predictions for these to recognize that this answers the question (other tools used to find the exact answer)

GPT-3



MIT Technology Review

Artificial intelligence / Machine learning

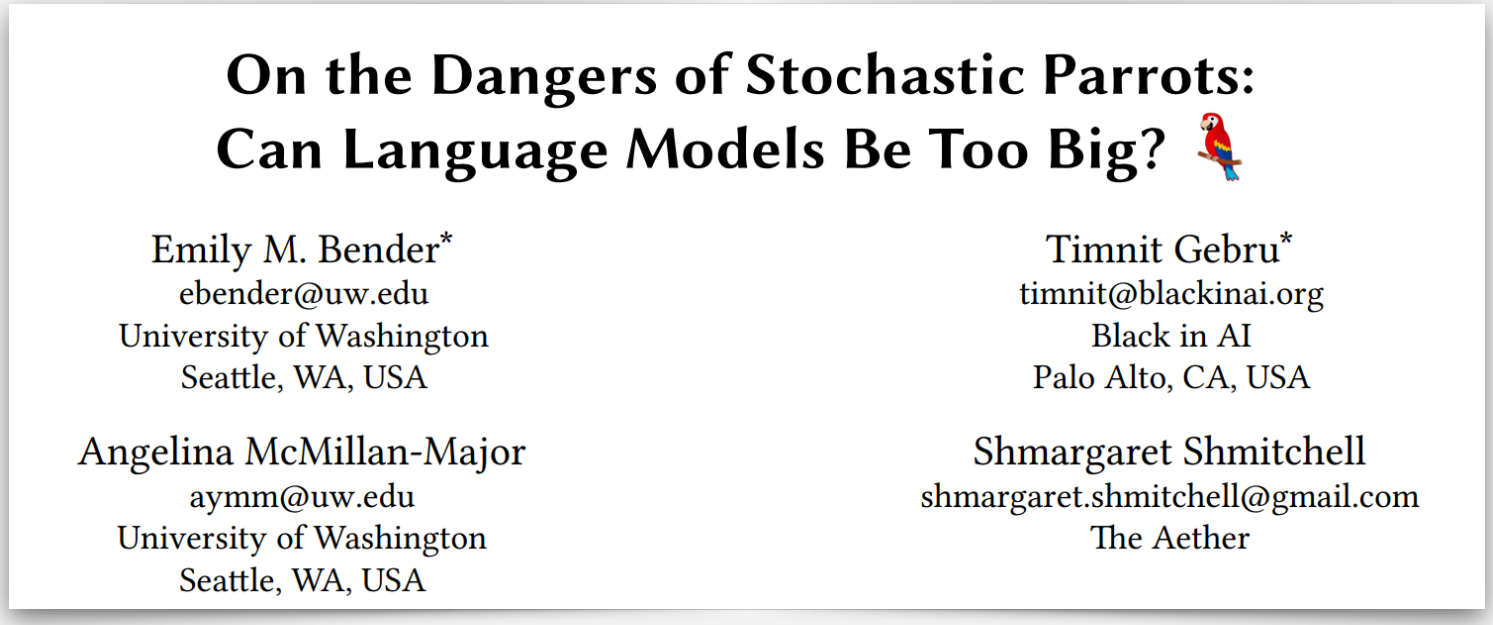
OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

by **Will Douglas Heaven** July 20, 2020

- GPT-3: much bigger than what's on Write With Transformer
- People want to do all sorts of things like medical question answering… but can these really work?

Stochastic Parrots



On the Dangers of Stochastic Parrots:
Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

- These models take lots of computation and have high energy consumption
- By memorizing text data, these models can **encode biases**. See what happens if you prompt the model with a sentence about white people vs. black people, or men vs. women, and how those answers differ.
- These models can be seen as “parrots”, memorizing patterns in data but not understanding deeply (but what is understanding?)

Timnit Gebru

Technology

Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.

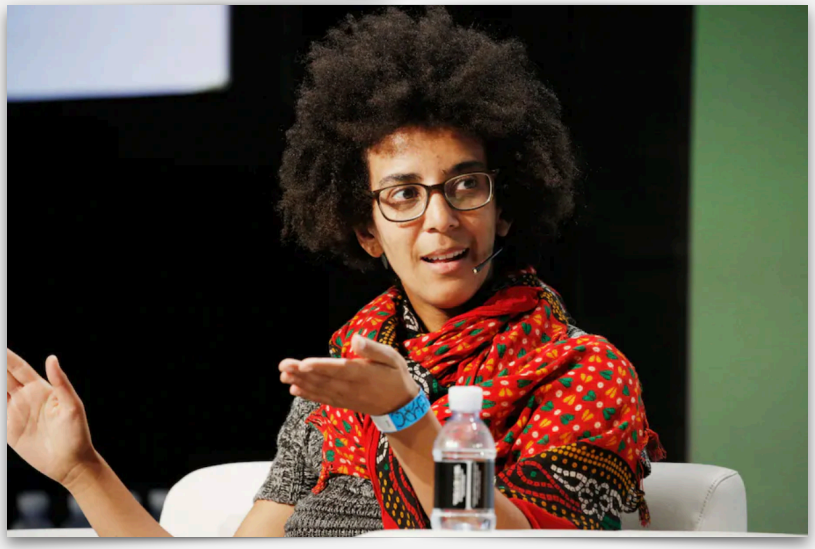
Gebru is one of the most high-profile Black women in her field and a powerful voice in the new field of ethical AI, which seeks to identify issues around bias, fairness, and responsibility.

https://www.washingtonpost.com/technology/2020/12/23/google-timnit-gebru-ai-ethics/

Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.

Timnit Gebru, one of the few Black women in her field, had voiced exasperation over the company's response to efforts to increase minority hiring.

https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html



GOOGLE

TIMNIT GEBRU WAS FIRED FROM GOOGLE — THEN THE HARASSERS ARRIVED

Even three months after Gebru's controversial termination from the AI Ethics team, the sustained campaign of aggressive tweets and emails keeps coming

By Zoe Schiffer | @ZoeSchiffer | Mar 5, 2021, 11:30am EST

https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Where to go next

How to study this more

Courses to take:

- More programming or software engineering can help but isn't critical
- Machine learning or data science
- Math to learn: probability, calculus (but not essential!)

Online courses

- Sentiment Analysis tutorial: `https://realpython.com/sentiment-analysis-python/`
- Andrew Ng's Coursera course: `https://www.coursera.org/learn/machine-learning`

Further Reading

- Understanding more about neural networks: Chris Olah, Jay Alammar

  `https://colah.github.io/`

  `https://jalammar.github.io/`

- Latest big language models:

  `https://openai.com/blog/better-language-models/`

  `https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html`