Introduction to Large Language Model

Lecture 1: 24-479-0207: Large Language Models

Dr. Jeena Kleenankandy
Assistant Professor, Department of Computer Science
Cochin University of Science and Technology

31 DECEMBER 2024

As a student, you are expected to:

(Other than regular attendance and uninterrupted attention..which are ***mandatory***)

- Make sure you get the foundations right !!! If you didn’t get it **ASK**
- Learn proactively - do extra reading, share your knowledge in class
- Get your hands dirty - Do lots and lots of coding (that is what that is finally going to land you at a good job)
- Innovate - Come up with ways to apply what you have learned to solve real world problems
- **Make us PROUD!**

Agenda

Introduce the course

- (What you can expect and what is expected from you)
- Have an open discussion and reach an agreement on how to conduct this course

Introduce the topic “Large Language Models”

- and know your interests so that we can plan!

Content delivery & Evaluation

- 4 Lecture hours per week (Will be mostly discussions)
- Learning materials will be provided via moodle
- Slides are only pointers - use study materials provided for exam preparation
- CA : Series 1 (20) + Series 2 (20) + Assignment (10) = 50 marks
- Assignment : Hand-on Projects
- End Semester exam - 50 marks
- Ungraded assignments and quizzes (free to skip at your own risk)

Last but not the least….

Do yourself a favor and start learning…

PYTORCH

Acknowledgement:

The remaining slides are taken from the course ***CSE473: Introduction to Artificial Intelligence***, Web: © 1993-2024, Department of Computer Science and Engineering, University of Washington.

Quick poll

1. Are you familiar with supervised machine learning? gradient descent?

2. Are you familiar with neural networks?

The language modeling problem

Rank these sentences in the order of plausibility?

1. Jane went to the store.
2. store to Jane went the.
3. Jane went store.
4. Jane goed to the store.
5. The store went to Jane.
6. The food truck went to Jane.

How probable is a piece of text? Or what is p(text)

p(*how are you this evening ? has your house ever been burgled ?*) = 10^{-15}

p(*how are you this evening ? fine , thanks , how about you ?*) = 10^{-9}

The language modeling problem

A **language model** answers the question: **What is p(text)?**

Text is a sequence of symbols: $(x^{(1)}, x^{(2)}, \ldots, x^{(N)})$

$$p(x^{(1)}, x^{(2)}, \ldots, x^{(N)})$$

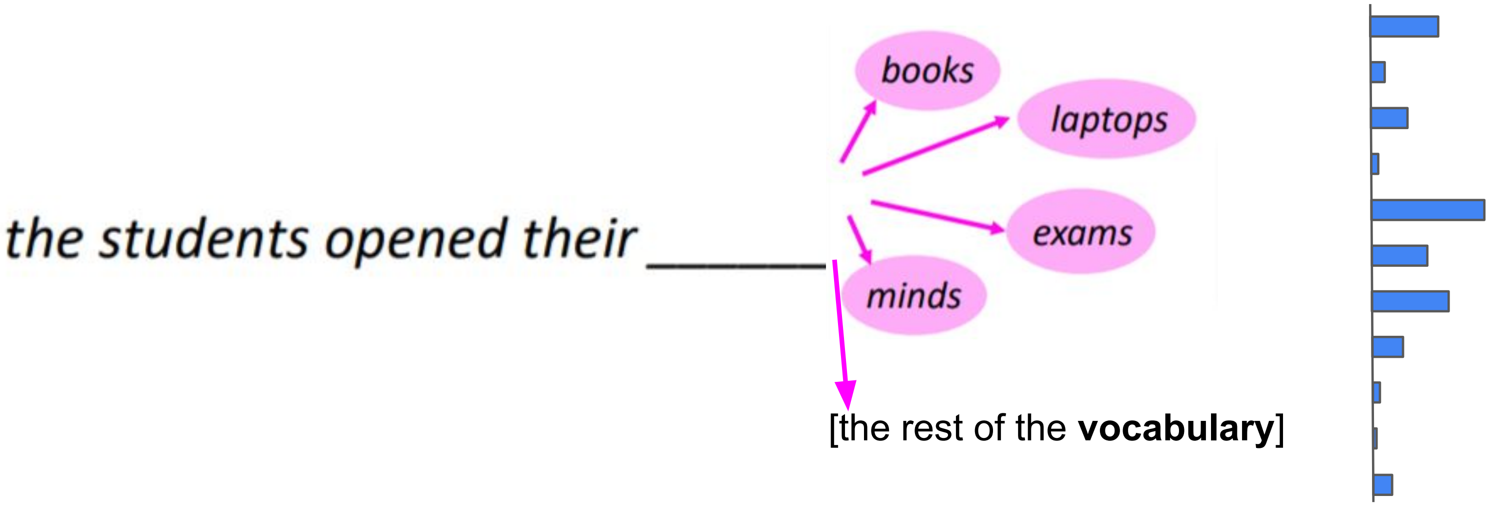
$$p(x^{(1)})p(x^{(2)}|x^{(1)})p(x^{(3)}|x^{(1)}, x^{(2)}) \ldots$$

$$\prod_{i=1}^{N} p(x^{(i)}|\underbrace{x^{(1)}, \ldots, x^{(i-1)}}_{\text{context}})$$

Just the chain rule of probability– no simplifying assumptions!

The language modeling problem

$$\prod_{i=1}^{N} p(x^{(i)}|\underbrace{x^{(1)},\ldots,x^{(i-1)}}_{\text{context}})$$



the students opened their ______

books

laptops

exams

minds

[the rest of the **vocabulary**]

Language models of this form can generate text

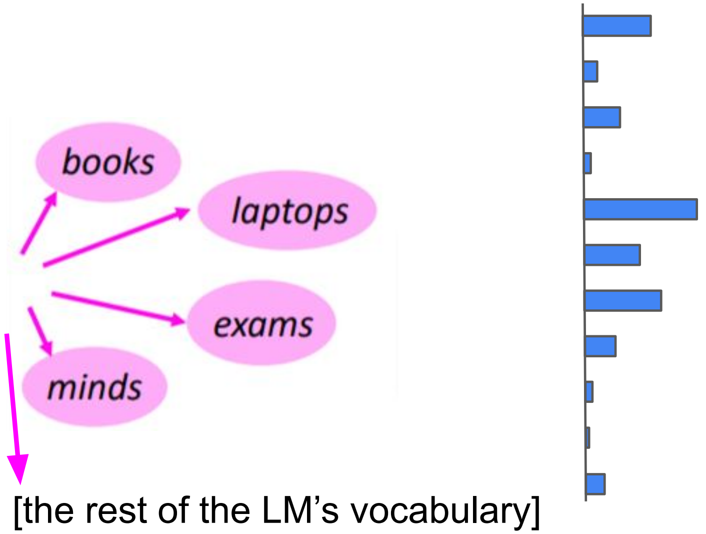
At each timestep, sample a token from the language model's new probability distribution over next tokens.

The ____

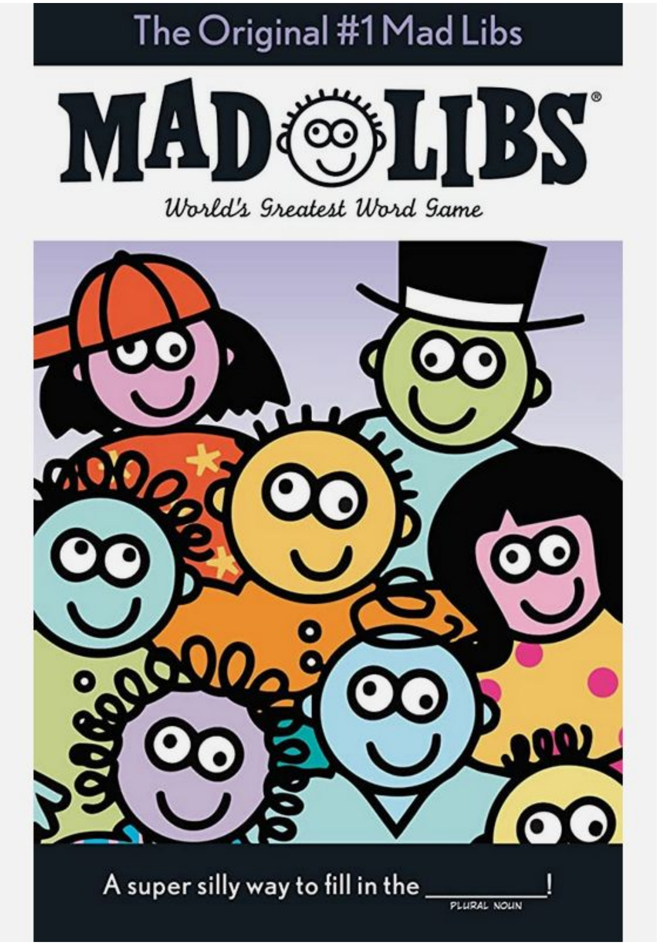
The students ____

The students opened ____

The students opened their ____



In short, predicting which word comes next


The Original #1 Mad Libs
MAD LIBS®
World's Greatest Word Game
A super silly way to fill in the ________!
PLURAL NOUN

The Original #1 Mad Libs
MAD LIBS®
World's Greatest Word Game

If you just heard someone say . . .

"Every morning before washing your CAT (NOUN), massage it gently with a/an DOORKNOB (NOUN) that has been soaked overnight in a/an PENCIL BOX (TYPE OF CONTAINER) full of warm CORN OIL (TYPE OF LIQUID)."

. . . you've obviously been playing **The Original #1 Mad Libs**! Play them with friends or enjoy them by yourself!

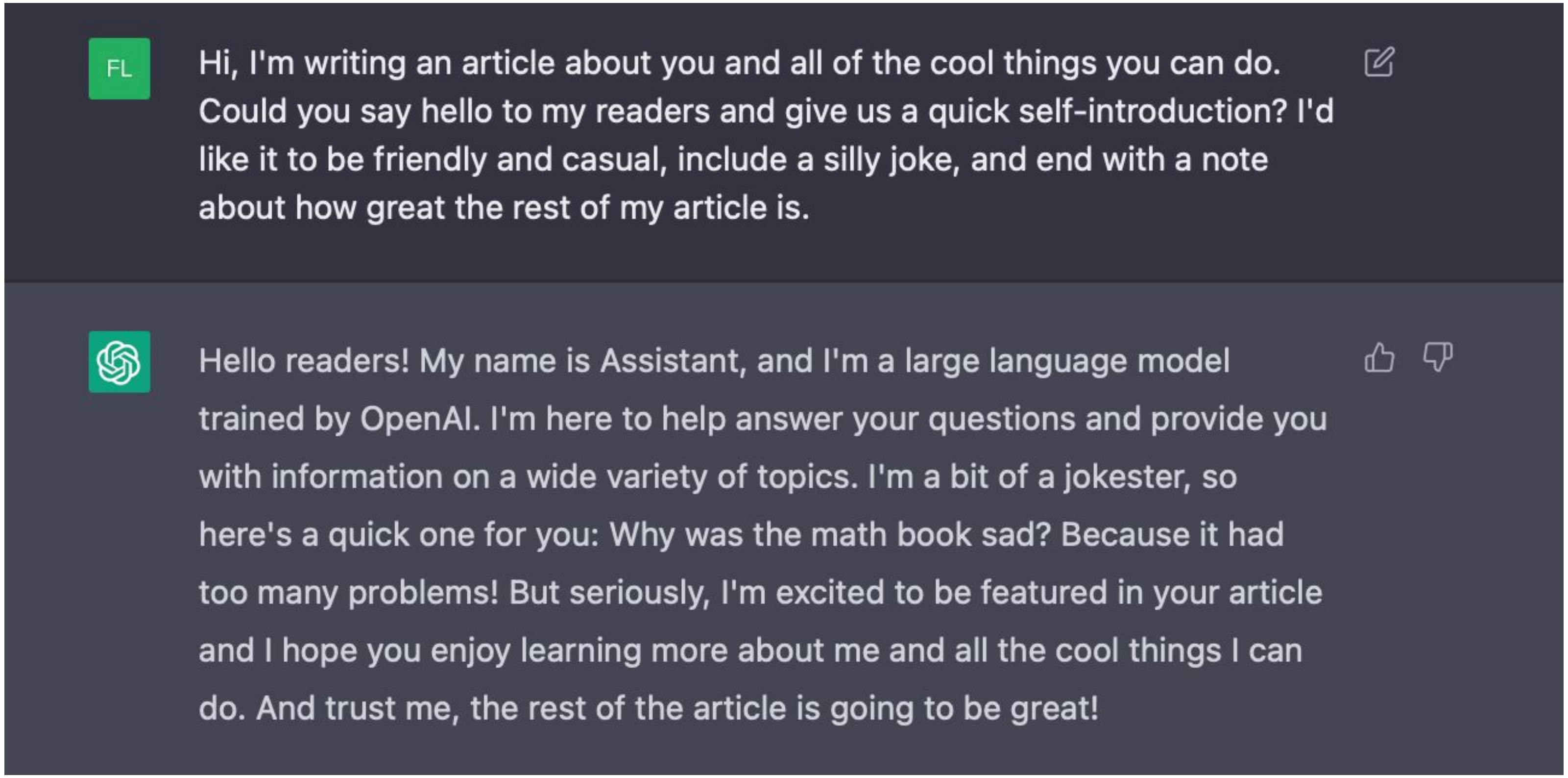
Download the awesome FREE Mad Libs app!
MAD LIBS
Available Now
ANDROID APP ON Google play
Download on the App Store
$3.99 US
($5.50 CAN)
PSS!
PRICE STERN SLOAN
www.penguin.com/youngreaders
www.madlibs.com
ISBN 978-0-8431-0055-6
EAN
9 780843 100556
50399>
Ridiculously simple directions inside!

Language models play the role of ...

- a judge of grammaticality
  - e.g., should prefer “The boy runs.” to “The boy run.”
- a judge of semantic plausibility
  - e.g., should prefer “The woman spoke.” to “The sandwich spoke.”
- an enforcer of stylistic consistency
  - e.g., should prefer “Hello, how are you this evening? Fine, thanks, how are you?” to “Hello, how are you this evening? Has your house ever been burgled?”
- a repository of knowledge (?)
  - e.g., “Barack Obama was the 44th President of the United States”

Note that this is very difficult to guarantee!

Language models in the news (these days, ChatGPT)

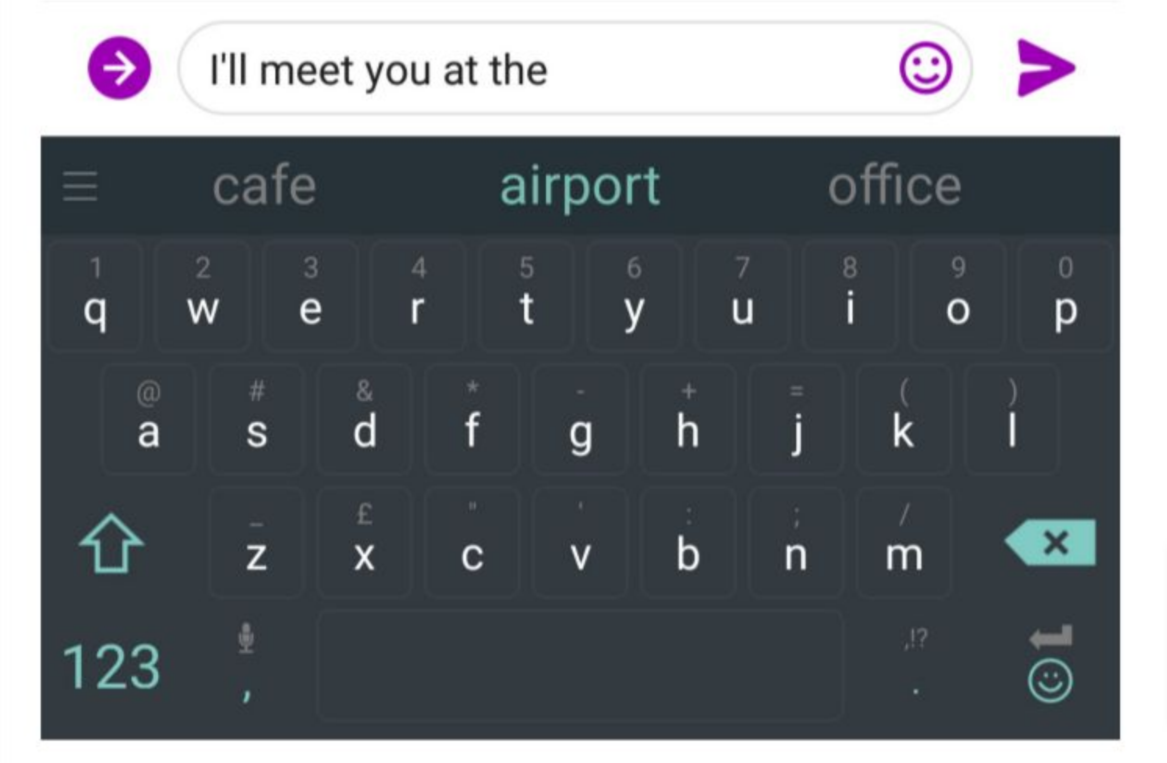


Hi, I'm writing an article about you and all of the cool things you can do. Could you say hello to my readers and give us a quick self-introduction? I'd like it to be friendly and casual, include a silly joke, and end with a note about how great the rest of my article is.

Hello readers! My name is Assistant, and I'm a large language model trained by OpenAI. I'm here to help answer your questions and provide you with information on a wide variety of topics. I'm a bit of a jokester, so here's a quick one for you: Why was the math book sad? Because it had too many problems! But seriously, I'm excited to be featured in your article and I hope you enjoy learning more about me and all the cool things I can do. And trust me, the rest of the article is going to be great!

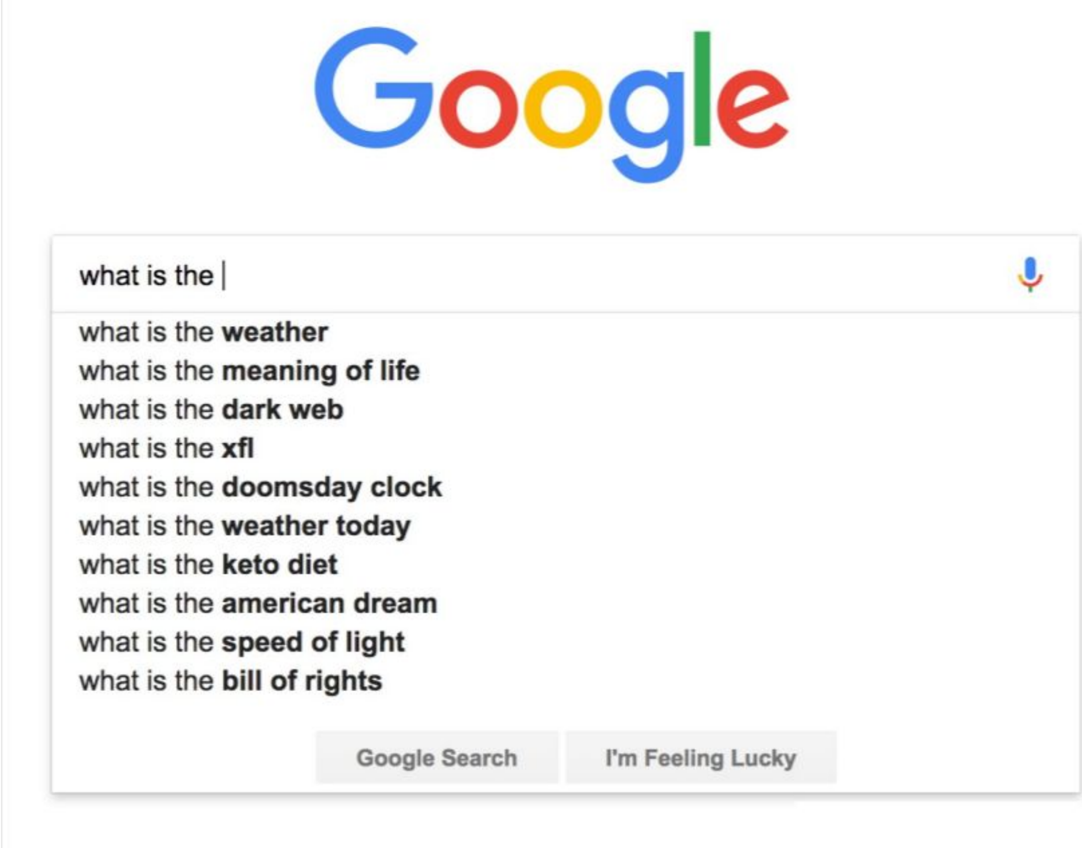
Image taken from Springboard

We use language models every day



I'll meet you at the

We use language models every day



Google

what is the |

what is the **weather**
what is the **meaning of life**
what is the **dark web**
what is the **xfl**
what is the **doomsday clock**
what is the **weather today**
what is the **keto diet**
what is the **american dream**
what is the **speed of light**
what is the **bill of rights**

Why language modeling?

- Machine translation
  - p(*strong winds*) > p(*large winds*)

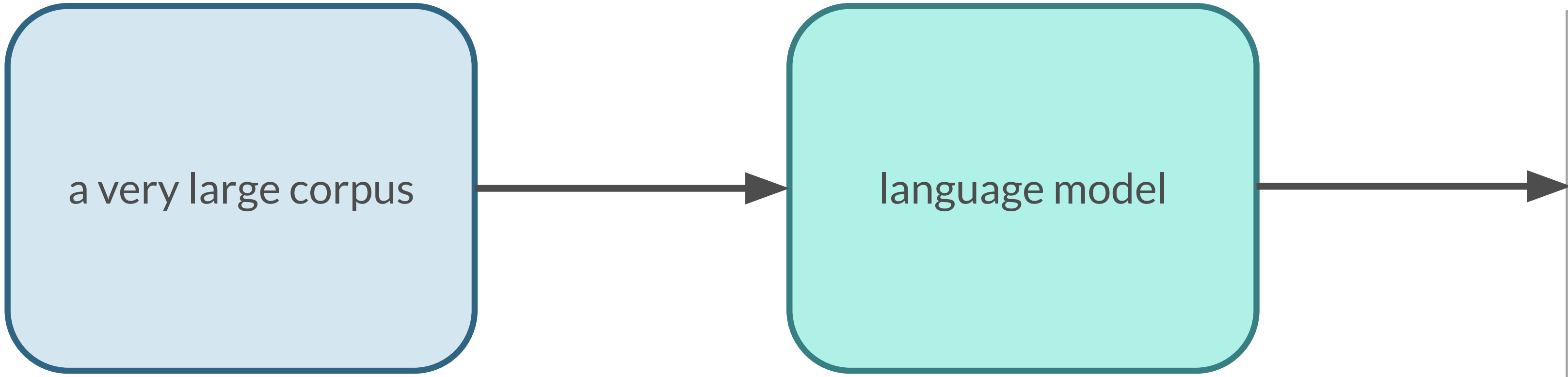
- Spelling correction
  - The office is about fifteen minuets from my house
  - p(*about fifteen minutes from*) > p(*about fifteen minuets from*)

- Speech recognition
  - p(*I saw a van*) >> p(*eyes awe of an*)

- Summarization, question-answering, handwriting recognition, OCR, etc.

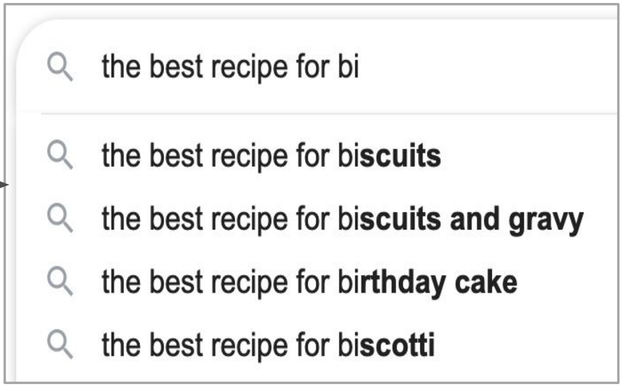
How we learn a language model

Language modeling



a very large corpus

language model



the best recipe for bi

the best recipe for bi**scuits**

the best recipe for bi**scuits and gravy**

the best recipe for bi**rthday cake**

the best recipe for bi**scotti**

How do we learn a language model?

Estimate probabilities using text data

- Collect a textual corpus
- Find a distribution that maximizes the probability of the corpus – maximum likelihood estimation

A naive solution: count and divide

- Assume we have N training sentences
- Let $x_1, x_2, \ldots, x_n$ be a sentence, and $c(x_1, x_2, \ldots, x_n)$ be the number of times it appeared in the training data.
- Define a language model:

$$p(x_1, \ldots, x_n) = \frac{c(x_1, \ldots, x_n)}{N}$$

No generalization!

Markov assumption

- We make the Markov assumption: $x^{(t+1)}$ depends only on the preceding n-1 words

$$P(\boldsymbol{x}^{(t+1)}|\boldsymbol{x}^{(t)},\ldots,\boldsymbol{x}^{(1)}) = P(\boldsymbol{x}^{(t+1)}|\underbrace{\boldsymbol{x}^{(t)},\ldots,\boldsymbol{x}^{(t-n+2)}}_{\text{n-1 words}})$$

assumptio

Markov assumption

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{transparent that})$$

or maybe even

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{that})$$

n-gram Language Models

$$\prod_{i=1}^{N} p(x^{(i)}|x^{(1)}, \ldots, x^{(i-1)})$$

"I have a dog whose name is Lucy. I have two cats, they like playing with Lucy."

- Question: How to learn a Language Model?
- Answer (pre- Deep Learning): learn an *n-gram* Language Model!

- Idea: Collect statistics about how frequent different n-grams are and use these to predict next word

unigram probability

“I have a dog whose name is Lucy. I have two cats, they like playing with Lucy.”

- corpus size m = 17
- P(Lucy) = 2/17; P(cats) = 1/17

- Unigram probability: $P(\mathrm{w}) = \frac{\mathrm{count}(\mathrm{w})}{\mathrm{m}} = \frac{\mathrm{C}(\mathrm{w})}{\mathrm{m}}$

bigram probability

“I have a dog whose name is Lucy. I have two cats, they like playing with Lucy.”

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

$$P(\text{have} \mid \text{I}) = \frac{P(\text{I have})}{P(\text{I})} = \frac{2}{2} = 1$$

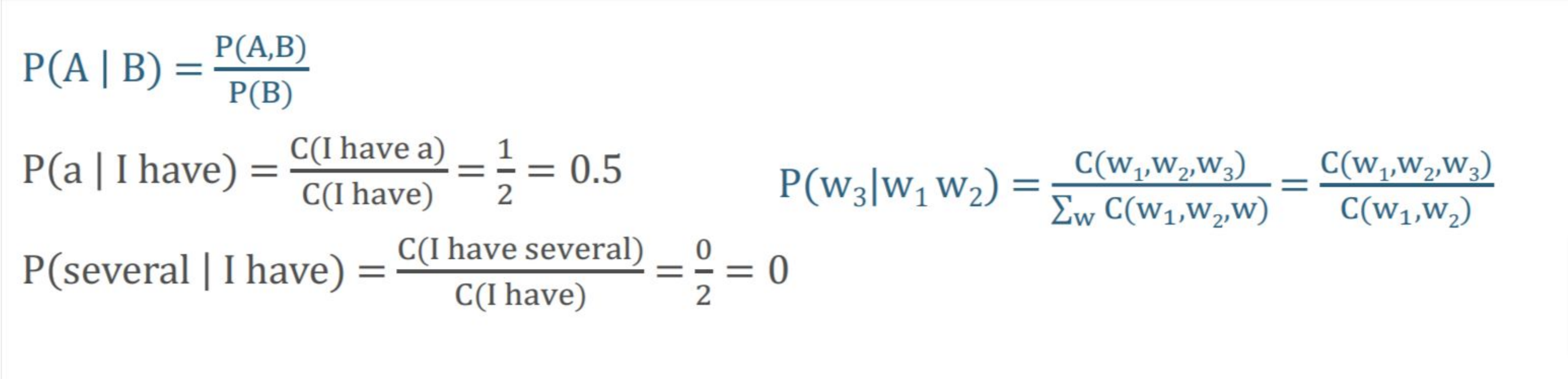
$$P(\text{two} \mid \text{have}) = \frac{P(\text{have two})}{P(\text{have})} = \frac{1}{2} = 0.5$$

$$P(\text{eating} \mid \text{have}) = \frac{P(\text{have eating})}{P(\text{have})} = \frac{0}{2} = 0$$

$$P(w_2|w_1) = \frac{C(w_1,w_2)}{\sum_w C(w_1,w)} = \frac{C(w_1,w_2)}{C(w_1)}$$

trigram probability

“I have a dog whose name is Lucy. I have two cats, they like playing with Lucy.”



$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

$$P(\mathrm{a} \mid \mathrm{I\ have}) = \frac{C(\mathrm{I\ have\ a})}{C(\mathrm{I\ have})} = \frac{1}{2} = 0.5$$

$$P(w_3 | w_1\, w_2) = \frac{C(w_1, w_2, w_3)}{\sum_w C(w_1, w_2, w)} = \frac{C(w_1, w_2, w_3)}{C(w_1, w_2)}$$

$$P(\mathrm{several} \mid \mathrm{I\ have}) = \frac{C(\mathrm{I\ have\ several})}{C(\mathrm{I\ have})} = \frac{0}{2} = 0$$

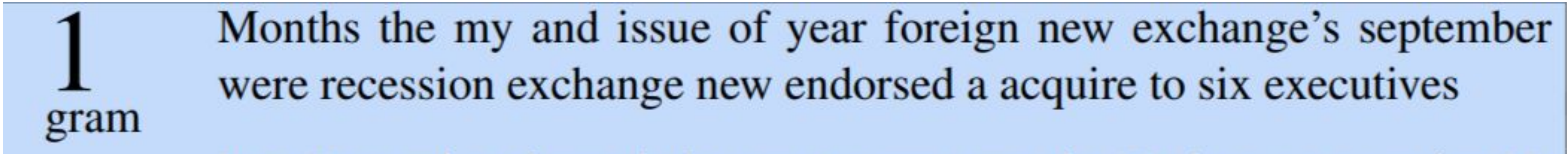
n-gram probability

“I have a dog whose name is Lucy. I have two cats, they like playing with Lucy.”

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

$$P(w_i \mid w_1, w_2, \ldots, w_{i-1}) = \frac{C(w_1, w_2, \ldots, w_{i-1}, w_i)}{C(w_1, w_2, \ldots, w_{i-1})}$$

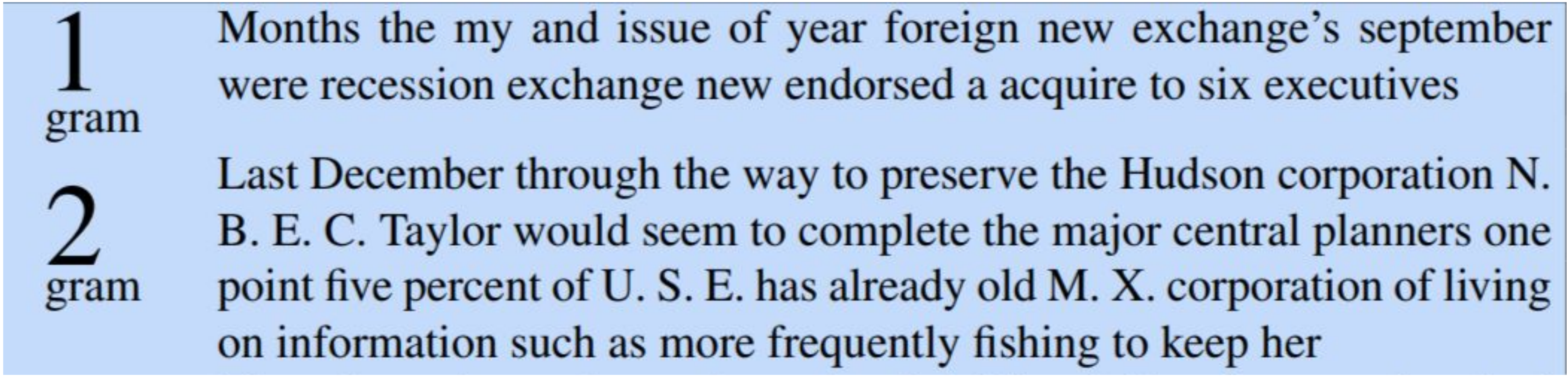
Sampling from an n-gram language model



1 gram

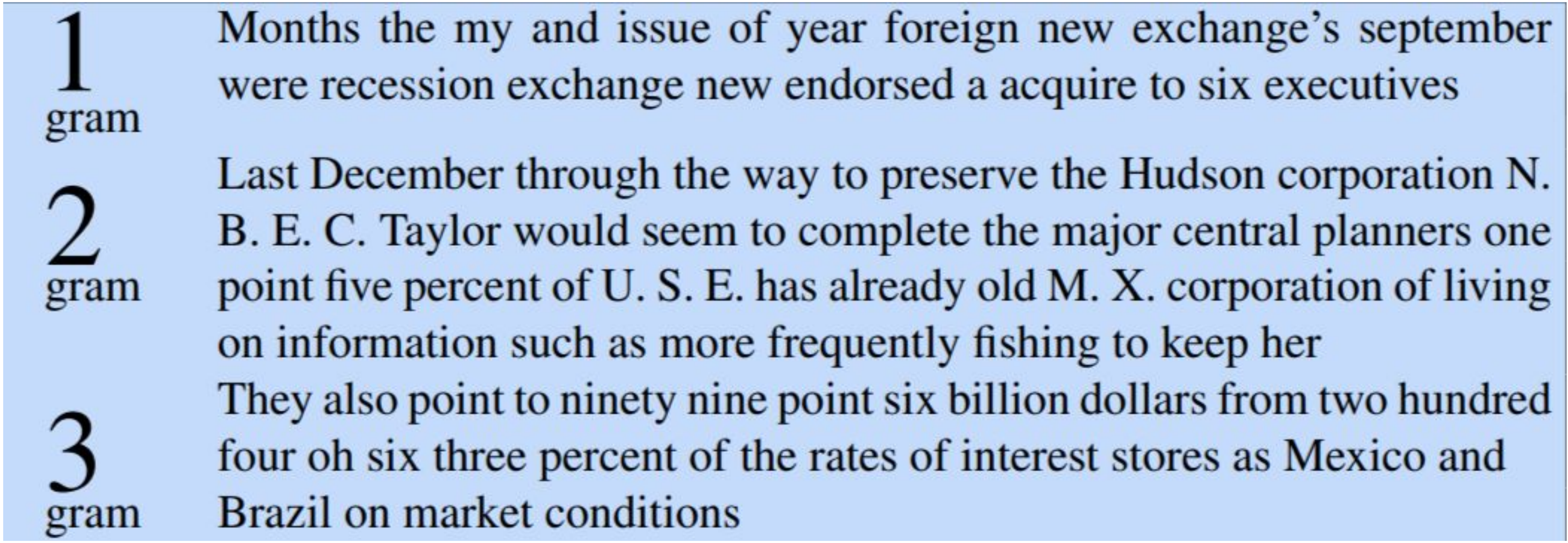
Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

Sampling from a language model



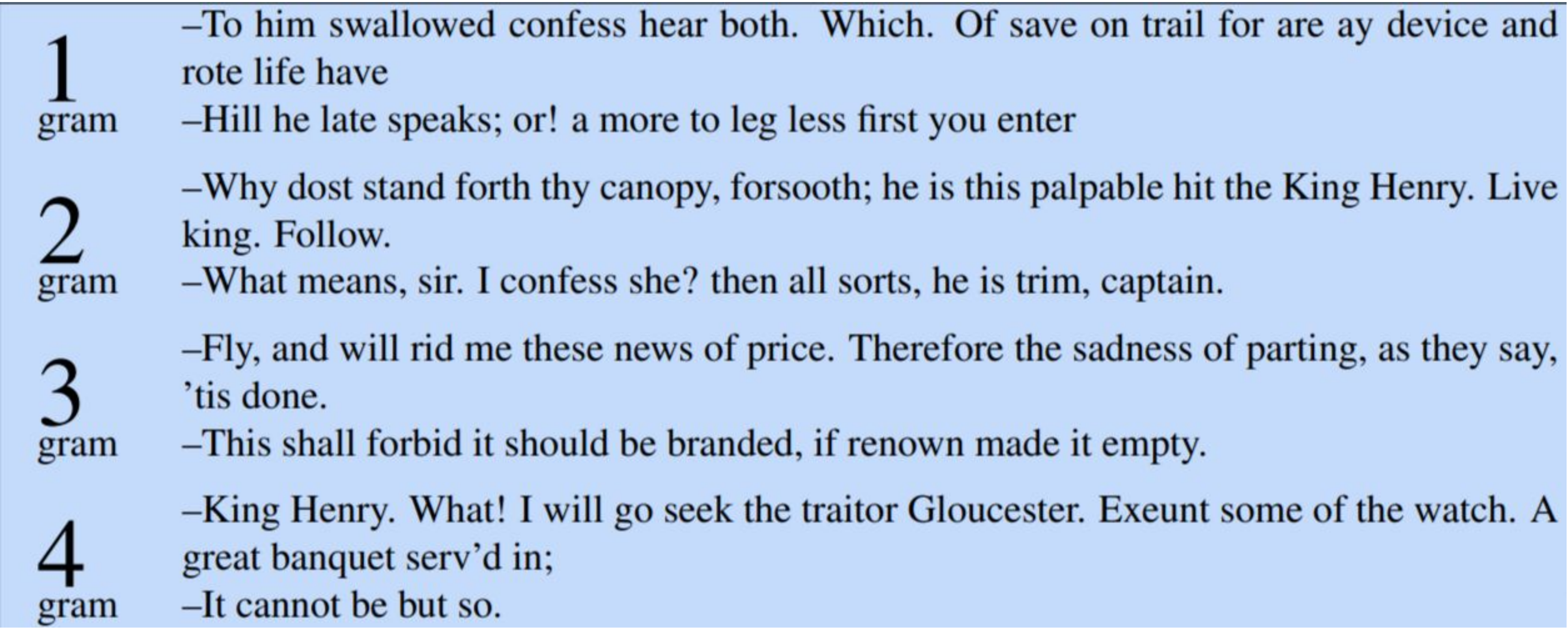
1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Sampling from a language model



1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
3 gram	They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

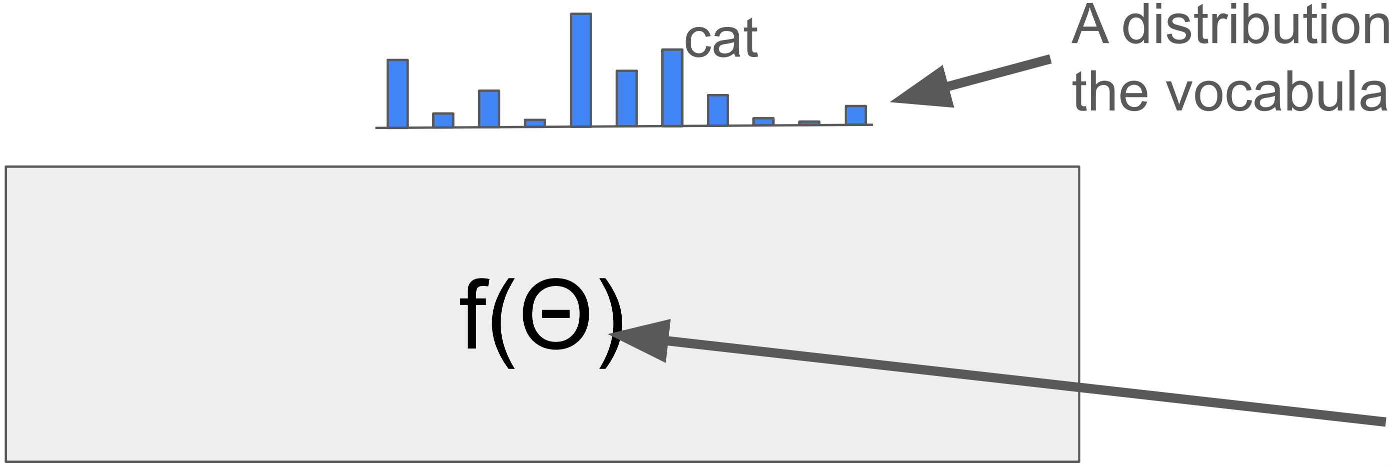
Sampling from a language model



1 gram	–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have –Hill he late speaks; or! a more to leg less first you enter
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. –What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done. –This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in; –It cannot be but so.

$$\prod_{i=1}^{N} p(x^{(i)}|x^{(1)},\ldots,x^{(i-1)})$$

Neural language models



A distribution over the vocabulary

A differentiable function (e.g. a neural network)

I have a dog whose name is Lucy. I have two

Large Language Models

The transformer model allows fast parallel computations on many GPUs **(large amounts of compute)**

It allows training on **large amounts of data** (think the whole internet worth of text).

It allows adding many and many layers in the model (**large model**)

A large language model is a language model with a large number of parameters, trained on large amounts of data, for long period of time.

Why large language models?

- Scaling the models, compute, and data leads in increase in performance

- Emergent properties at scale (Wei et al 2022)
    - Large models (with 7-100B+ parameters) suddenly become capable of performing tasks they weren’t able to do when small (such as 1B or small).

Training the model to chat

A simple language model (also called a pretrained model) is not equipped to chat with an end user, like ChatGPT.

ChatGPT (and many other models) are further trained on supervised data to follow instructions.

Instruction Tuning

- Collect a large dataset of instruction following examples of the form
  - <instruction> <input> <output>
  - For example,
  - Summarize this news article [ARTICLE] [SUMMARY]
  - Answer this question [QUESTION] [ANSWER]
  - Predict the sentiment of this review [REVIEW] [SENTIMENT]....

- This is also a text corpus but in a very specific format.

- Continue training the model on this dataset (again using the same training objective)

Aligning the model to humans’ preferences

- Chat based on models are supposed to converse with humans

- Why not learn from humans’ feedback

- Basic idea: Model samples multiple outputs – users rank them based on their preference
    - Convert user preferences into reward scores – more preferred output has higher reward
    - Treat an LLM like an agent and use RL to maximize this reward (RLHF)

So what does this mean ChatGPT is good at?

Some aspects of producing answers that might fall under that category:

- Writing in specific styles (that have appeared in the model’s training data)
- Grammatical consistency
- Generating boilerplate sentences that often appear at the beginning, end of emails, etc.
- Fluency

What are some problems that ChatGPT’s training leaves it prone to?

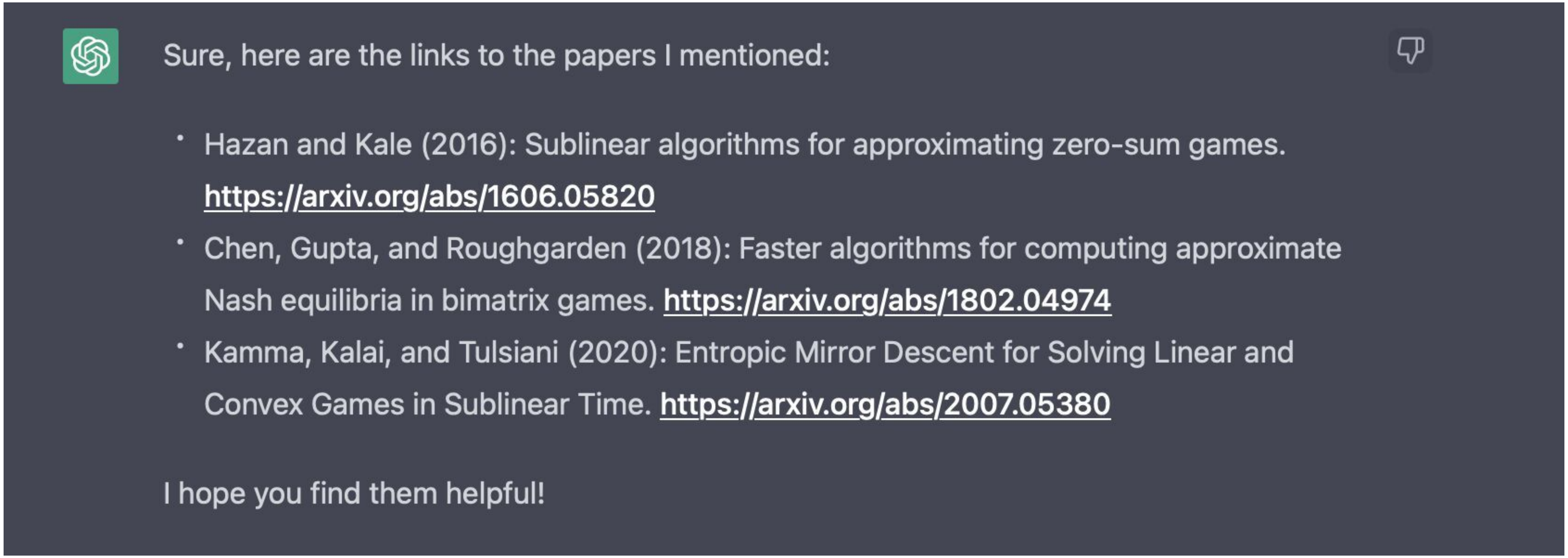
Inaccuracies

- The language model doesn’t “plan” what it will say in advance

- The model doesn’t store facts, just outputs plausible looking sentences which may or may not be factual

Lack of source attribution

Just like the model doesn't store facts… it doesn't store sources.



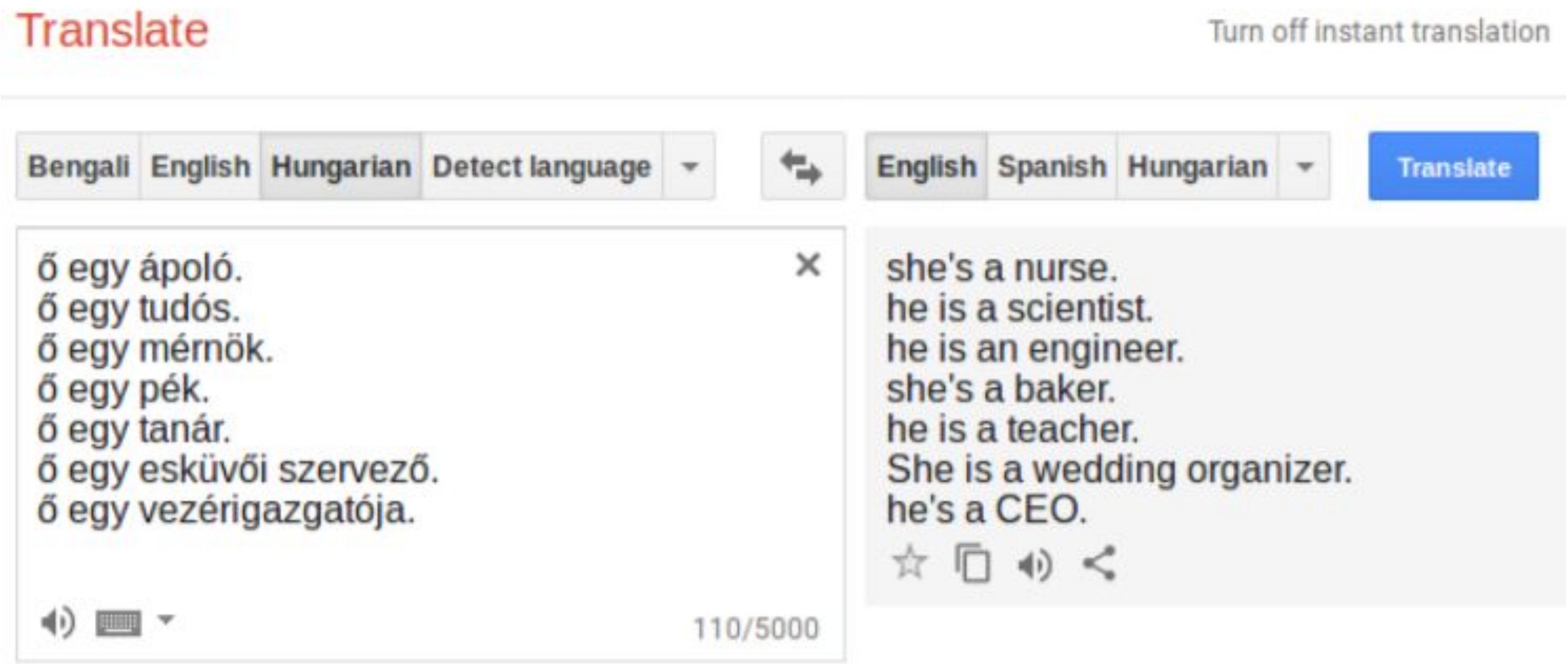
Sure, here are the links to the papers I mentioned:

- Hazan and Kale (2016): Sublinear algorithms for approximating zero-sum games. https://arxiv.org/abs/1606.05820
- Chen, Gupta, and Roughgarden (2018): Faster algorithms for computing approximate Nash equilibria in bimatrix games. https://arxiv.org/abs/1802.04974
- Kamma, Kalai, and Tulsiani (2020): Entropic Mirror Descent for Solving Linear and Convex Games in Sublinear Time. https://arxiv.org/abs/2007.05380

I hope you find them helpful!

Outputs that reflect social biases

An example from machine translation a few years ago:



ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.

she's a nurse.
he is a scientist.
he is an engineer.
she's a baker.
he is a teacher.
She is a wedding organizer.
he's a CEO.

Thanks! Questions?