Introduction

Cornell CS 5740: Natural Language Processing
Yoav Artzi, Spring 2023

CS 5740

- Goal: learn the technical language of NLP
- Introduction-level class for contemporary NLP methods
- Emphasizing depth over breadth
- Technical hands-on experience
- Understand both potential and limitations
- Give you the tools to expand your knowledge and adapt

Overall Structure

Two(-ish) Main Parts

- Part 0: Quick warm up — to get us started with Assignment 1
- Part 1: Learning from raw data
- Part 2: Learning from annotated data

Overall Structure

Repeating Themes

- Data, evaluation
- Representation, learning, model design
- Modeling: insight vs. expressivity
    - Models where we can understand what is going on
    - Opaque (neural) models that are more expressive
- Engineering issues, scale
- Formulating tasks to operationalize problems

Today

- Go over technicalities
- Overall course structure
- Some history

Prerequisite Background

- Strong programming and ML experience
- All assignments are in Python, and background knowledge is assumed
- This class is not a tutorial on any specific software framework (e.g., PyTorch) — tons of online resources do it well
- PyTorch will be used extensively, and we assume you have experience with it
- **Strongly recommended: practical deep learning workshop**

Grading

- Assignments: 90pt
    - Four assignments
    - Not equally weighted: solo assignments weigh more
- Engagement: 10pt
- Points are not grade percentages

Assignments

- Four assignments, all mandatory
- Reporting requirements, use of libraries, and other details are specified in each assignment — if you have doubts, ask!
- All assignments are to be implemented in Python

Generative AI Expectations

- You may use generative products (e.g., ChatGPT) to learn about methods and using packages, similar to how you would use Stack Overflow and similar forums
- Do not simply use generated code
    - Don’t copy and modify, but instead learn from generated examples and write your own code
- All other uses of generative tools are not allowed
- Because it would just diminish what you get from this class

CS 5740

What does CS 5740 need from you?

- The right background
- A lot of hard work and time
- Deep and consistent engagement
- Getting things done on time!
- It will be a lot of fun, and you will learns tons, if you take it seriously :)

Tips

What to do?

- Attend the lectures — it really changes the course outcome
- Come on time! Late? Enter quietly and sit at the back
- Bring pen and paper 🗒🖋
- Discuss with others — this is what the forum is for!
- Again: Use the forum! Debugging, crazy ideas …. help each other (it counts as engagement, and that’s part of your grade)

Tips

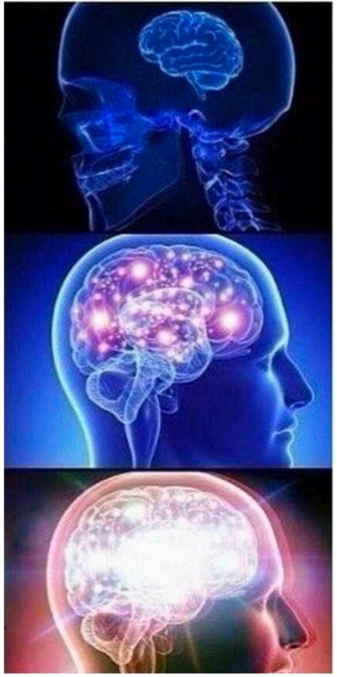
What not to do?

- Don’t take this class with an overloaded schedule (generally, just don’t overload your schedule)
- Don’t procrastinate on assignments — start them immediately when you get them
- Don’t cut corners on prerequisites
- Do not use electronics in class

Questions?

What is NLP?

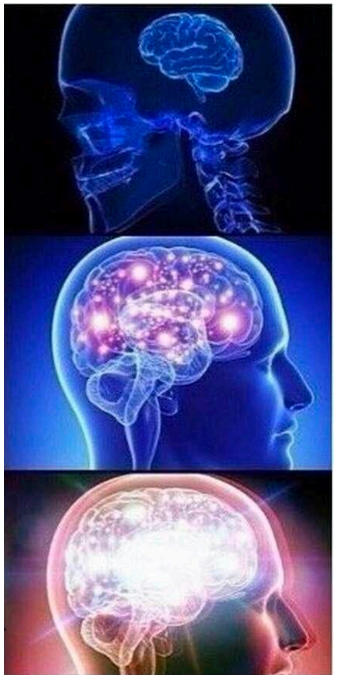
- Fundamental goal: deep understanding of broad language
  - Not just keyword matching, or just a narrow domain
- End systems can vary in complexity



- Simple:
- Complex:
- Unknown:

What is NLP?

- Fundamental goal: deep understanding of broad language
  - Not just keyword matching, or just a narrow domain
- End systems can vary in complexity



- Simple: spelling correction, text categorization
- Complex: speech recognition, machine translation, dialog, question answering, personal assistance
- Unknown: human-level comprehension (*is that just NLP?*)

NLP History in a Nutshell

Pre-statistics

(1) Colorless green ideas sleep furiously.

(2) Furiously sleep ideas green colorless

NLP History in a Nutshell

Pre-statistics

(1) Colorless green ideas sleep furiously.

(2) Furiously sleep ideas green colorless

It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not.” (Chomsky 1957)

NLP History in a Nutshell

Pre-statistics

- 70s and 80s: more linguistic focus
  - Emphasis on deeper models, syntax and semantics
  - Manually engineered systems (i.e., rule based)
  - Toy domains
  - Weak empirical evaluation

NLP History in a Nutshell

1990s Empirical Revolution

“Whenever I fire a linguist our system performance improves.”
– Jelinek, 1988

- Corpus-based methods produce the first widely used tools
- Deep linguistic analysis often traded for robust approximations
- Empirical evaluation is essential

“Of course, we must not go overboard and mistakenly conclude that the successes of statistical NLP render linguistics irrelevant (rash statements to this effect have been made in the past, e.g., the notorious remark, “Every time I fire a linguist, my performance goes up”). The information and insight that linguists, psychologists, and others have gathered about language is invaluable in creating high-performance broad-domain language understanding systems …”

- Lillian Lee (2001) http://www.cs.cornell.edu/home/llee/papers/cstb/index.html

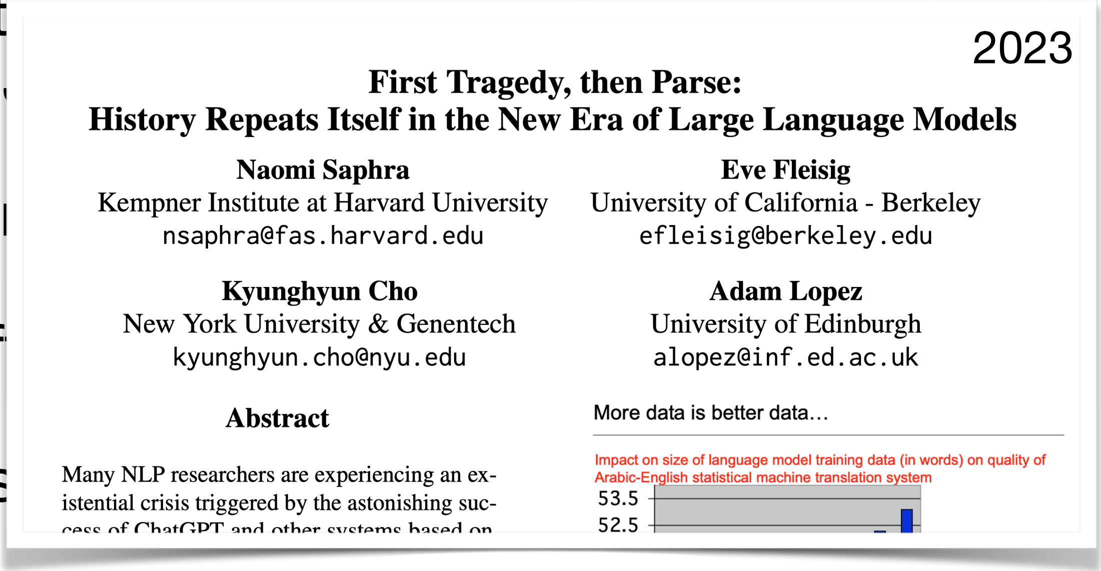
NLP History in a Nutshell

1990s Empirical Revolution

“Whenever I fire a linguist

–

- Corpus-based methods p
- Deep linguistic analysis of
- Empirical evaluation is ess



2023

First Tragedy, then Parse:
History Repeats Itself in the New Era of Large Language Models

Naomi Saphra
Kempner Institute at Harvard University
nsaphra@fas.harvard.edu

Eve Fleisig
University of California - Berkeley
efleisig@berkeley.edu

Kyunghyun Cho
New York University & Genentech
kyunghyun.cho@nyu.edu

Adam Lopez
University of Edinburgh
alopez@inf.ed.ac.uk

Abstract

Many NLP researchers are experiencing an existential crisis triggered by the astonishing success of ChatGPT and other systems based on

More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system

53.5

52.5

> *“Of course, we must not go overboard and mistakenly conclude that the successes of statistical NLP render linguistics irrelevant (rash statements to this effect have been made in the past, e.g., the notorious remark, “Every time I fire a linguist, my performance goes up”). The information and insight that linguists, psychologists, and others have gathered about language is invaluable in creating high-performance broad-domain language understanding systems …”*

- Lillian Lee (2001) http://www.cs.cornell.edu/home/llee/papers/cstb/index.html

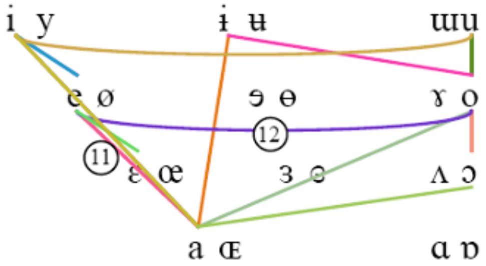
NLP History in a Nutshell

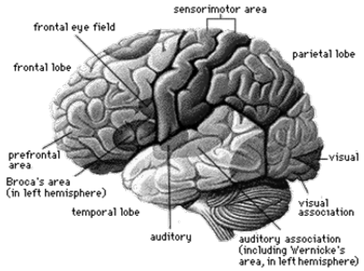
1990s Empirical Revolution

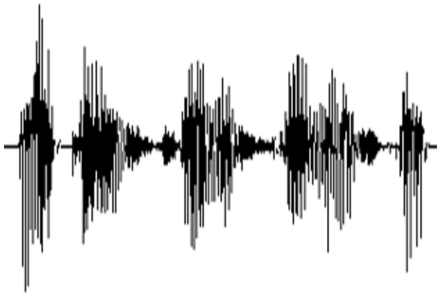
- 1990s Empirical Revolution
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!
  - Machine translations starts to work well. Yay!
- 2010s: NLP+X, excitement about neural networks (again), pre-trained representations
  - Robust speech recognition and personal assistants show up
- 2020s: large language models, large-scale data and hardware …

Related Fields

- **Computational Linguistics**
  - Using computational methods to learn more about how language works
  - We end up doing this and using it
- **Cognitive Science**
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!
- **Speech**
  - Mapping audio signals to text
  - Great out separately from NLP from signal processing (in ECE)
  - Techniques have converged significantly (but communities remain largely separate)







Core Challenges

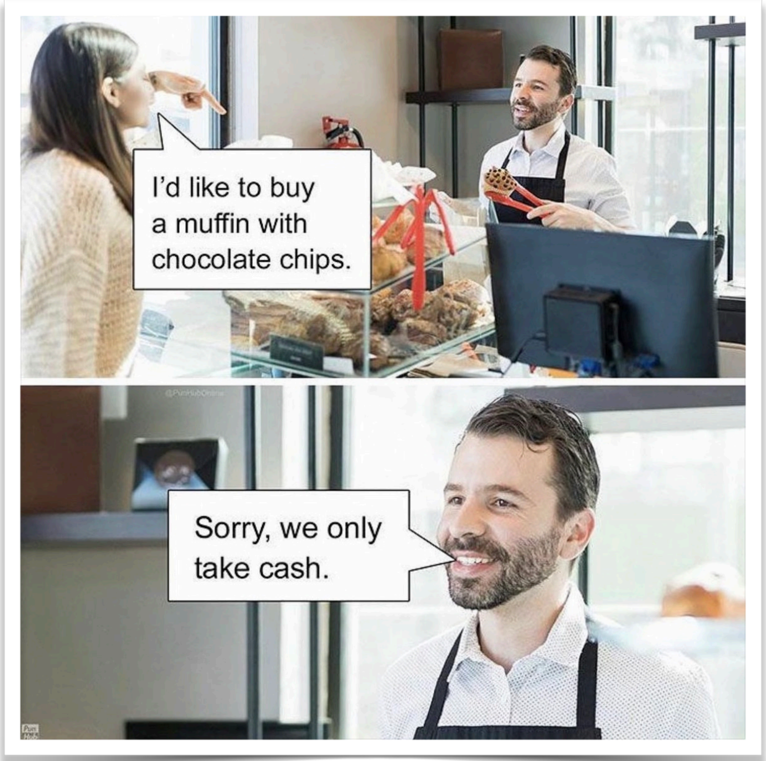
We can understand programming language.
Why is natural language different?

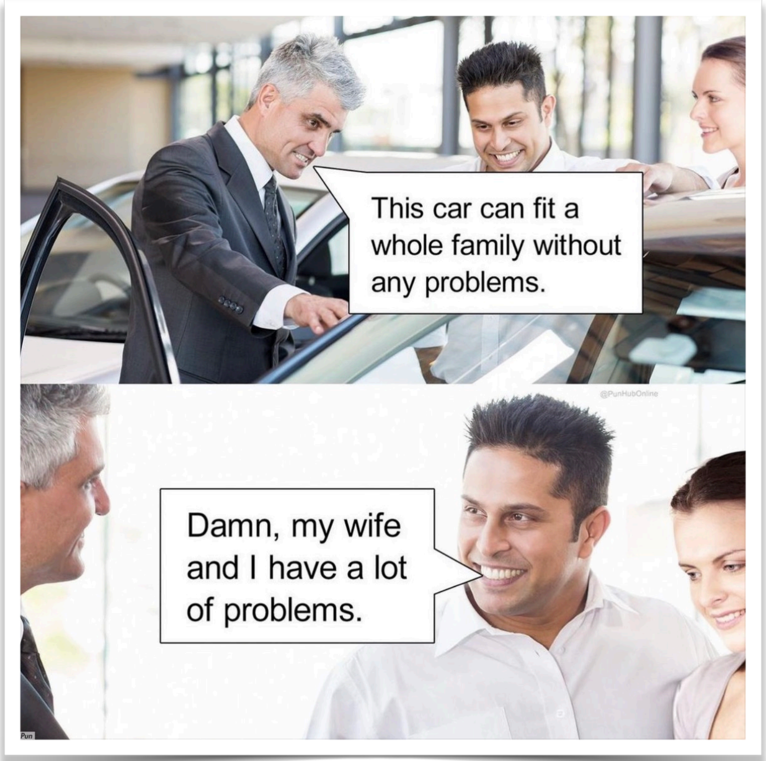
Core Challenges

We can understand programming language.
Why is natural language different?

- Ambiguity
- Scale
- Sparsity

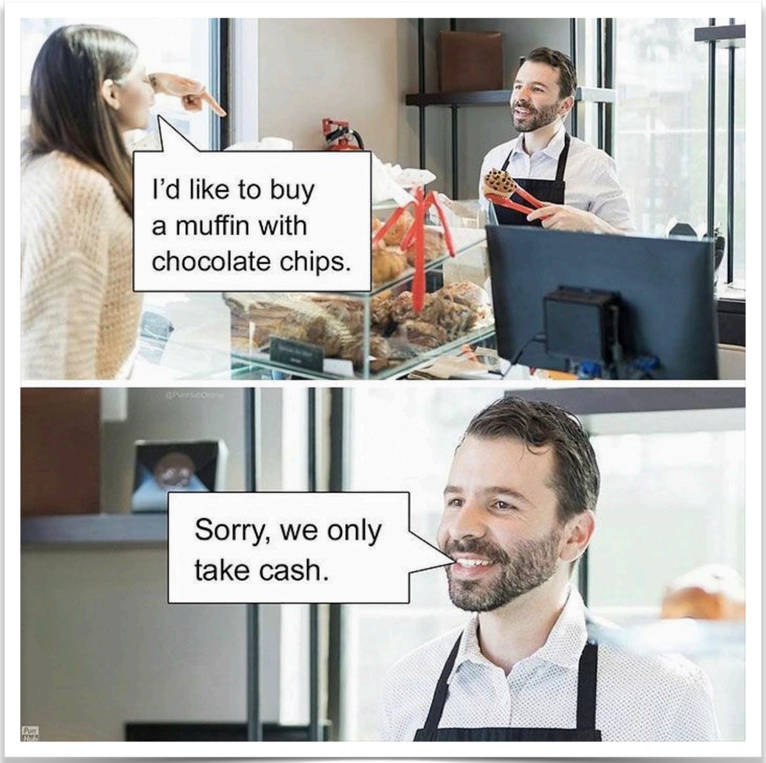
Ambiguity

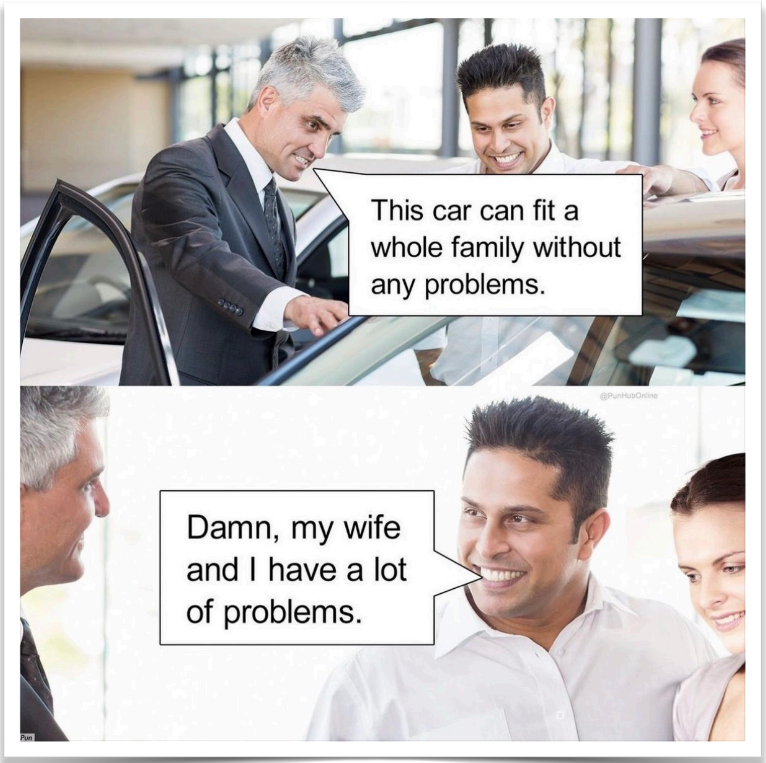

I'd like to buy a muffin with chocolate chips.
Sorry, we only take cash.
@PunHubOnline

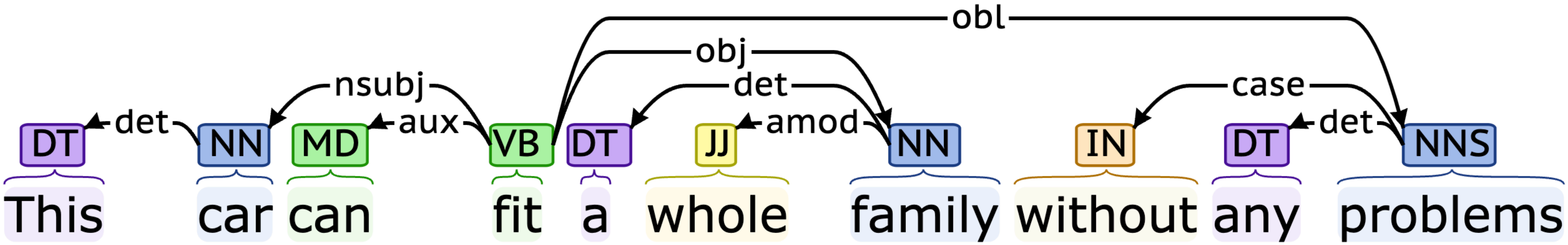

This car can fit a whole family without any problems.
Damn, my wife and I have a lot of problems.
@PunHubOnline

Ambiguity

Syntactic Ambiguity


I'd like to buy a muffin with chocolate chips.
Sorry, we only take cash.


This car can fit a whole family without any problems.
Damn, my wife and I have a lot of problems.



obl
obj
nsubj det case
det aux amod det

DT NN MD VB DT JJ NN IN DT NNS

This car can fit a whole family without any problems

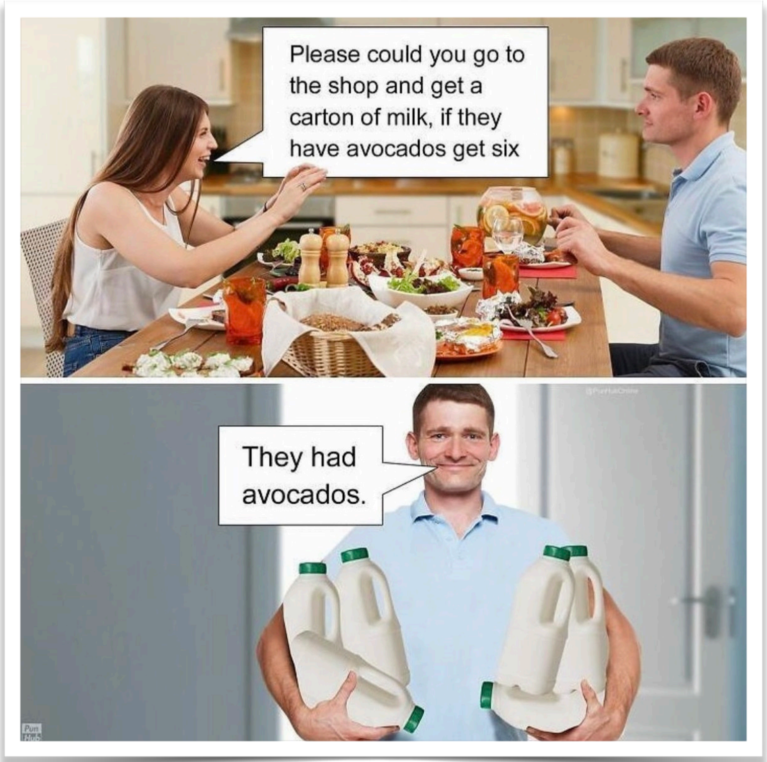
Ambiguity

Syntactic Ambiguity

ellipsis

noun (plural **ellipses**)

the omission from speech or writing of a word or words that are superfluous or able to be understood from contextual clues: *it is very rare for an ellipsis to occur without a linguistic antecedent*


Please could you go to the shop and get a carton of milk, if they have avocados get six
They had avocados.

Ambiguity

Semantic Ambiguity

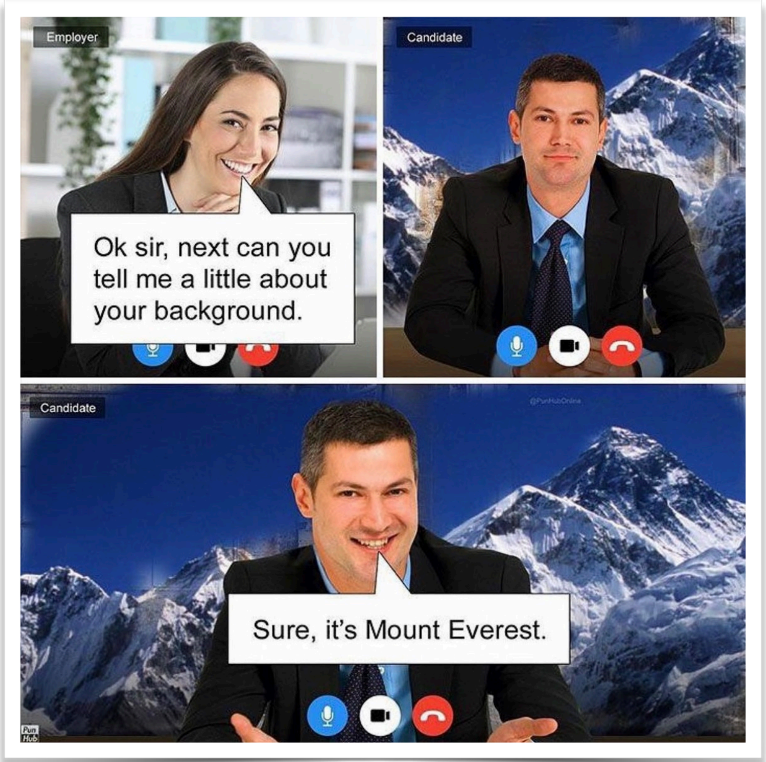
background | ˈbakˌground |
noun

1 *[in singular]* the area or scenery behind the main object of contemplation, especially when perceived as a framework for it: *the house stands against a background of sheltering trees*.

- the part of a picture or design that serves as a setting to the main figures or objects, or that appears furthest from the viewer: *the background shows a landscape of domes and minarets* | *the word is written in white on a red background*.
- a position or function that is not prominent or conspicuous: *after that evening, Athens remained* ***in the background***.
- *Computing* used to describe tasks or processes running on a computer that do not need input from the user: *programs can be left running* ***in the background***.
- *Physics* low-intensity radiation from radioisotopes present in the natural environment.
- unwanted signals, such as noise in the reception or recording of sound.

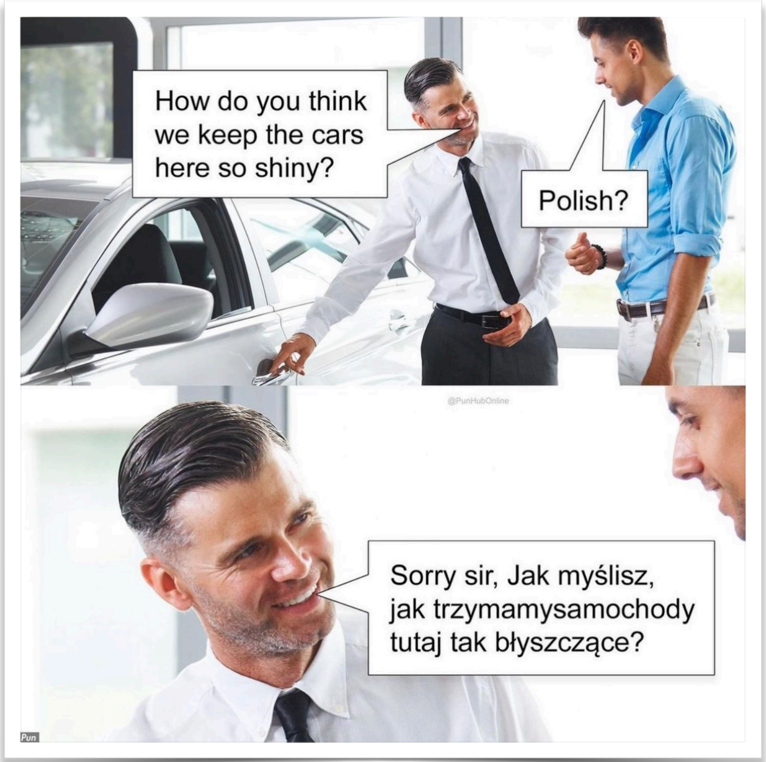
2 the circumstances or situation prevailing at a particular time or underlying a particular event: *the political and economic background* | *[as modifier] : background information*.

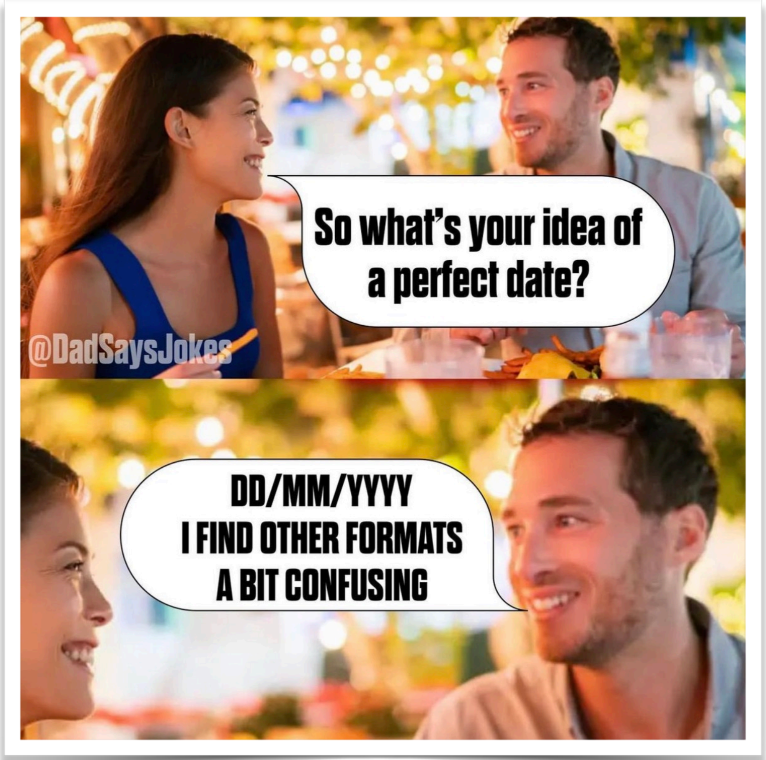
- a person's education, experience, and social circumstances: *she has a background in nursing* | *a mix of students from many different backgrounds*.


Employer
Candidate
Ok sir, next can you tell me a little about your background.
Candidate
@PunHubOnline
Sure, it's Mount Everest.
Pun Hub

Ambiguity

Semantic Ambiguity


How do you think we keep the cars here so shiny?
Polish?
@PunHubOnline
Sorry sir, Jak myślisz, jak trzymamysamochody tutaj tak błyszczące?
Pun


So what's your idea of a perfect date?
@DadSaysJokes
DD/MM/YYYY I FIND OTHER FORMATS A BIT CONFUSING

Ambiguity

Semantic Ambiguity

At last, a computer that understands you like your mother.

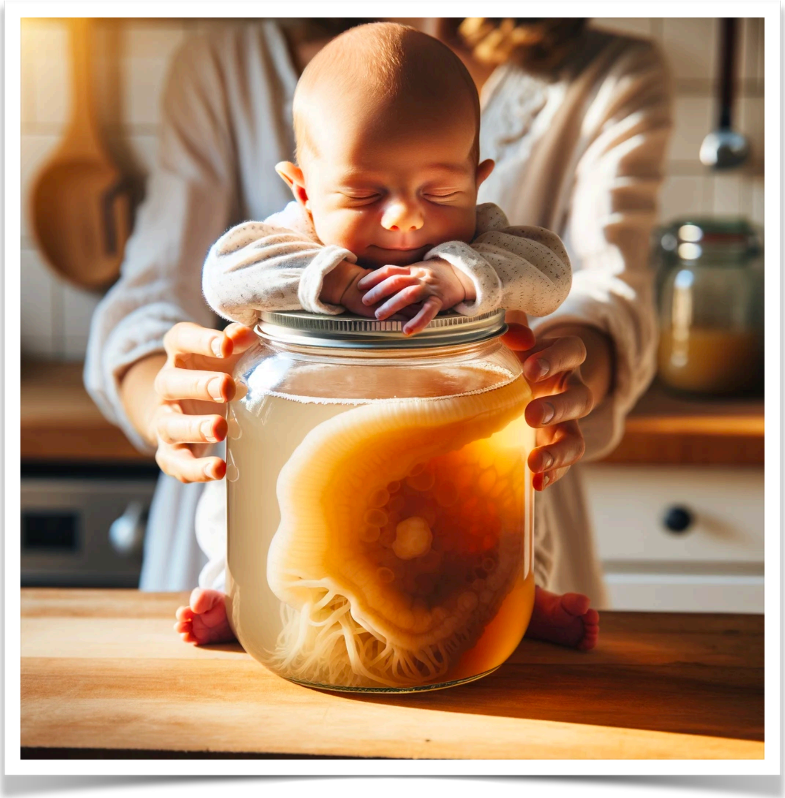
[Example from Lillian Lee]

Ambiguity

Semantic Ambiguity

At last, a computer that understands you like your mother.

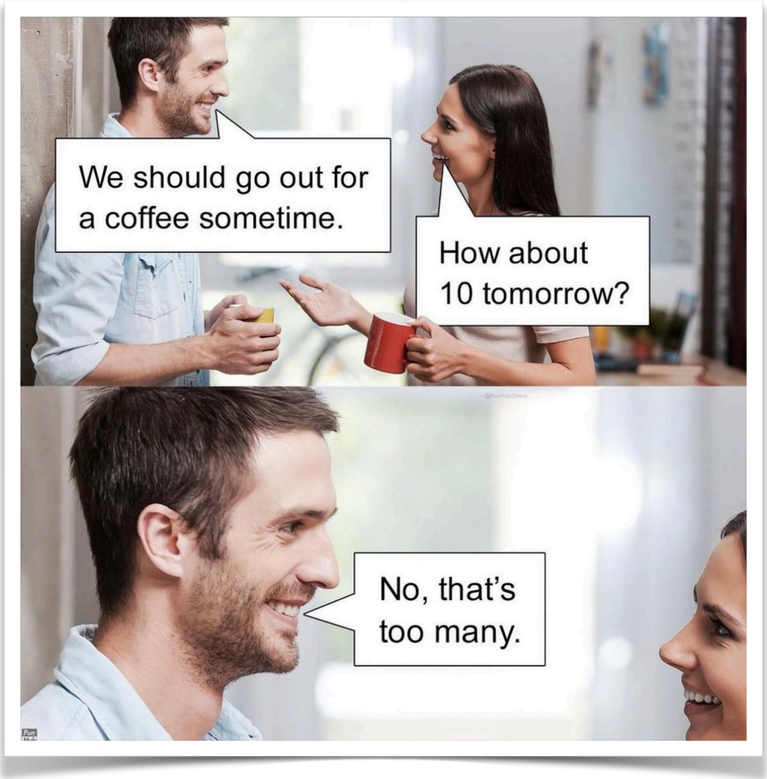
- Direct meanings:
  - It understands you like your mother (does) [presumably well]
  - It understands (that) you like your mother
- “mother” could mean:
  - a woman who has given birth to a child
  - a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar
- Context matters, e.g. if previous sentence was: “Wow, Amazon predicted that you would need to order a big batch of new vinegar brewing ingredients.😍”

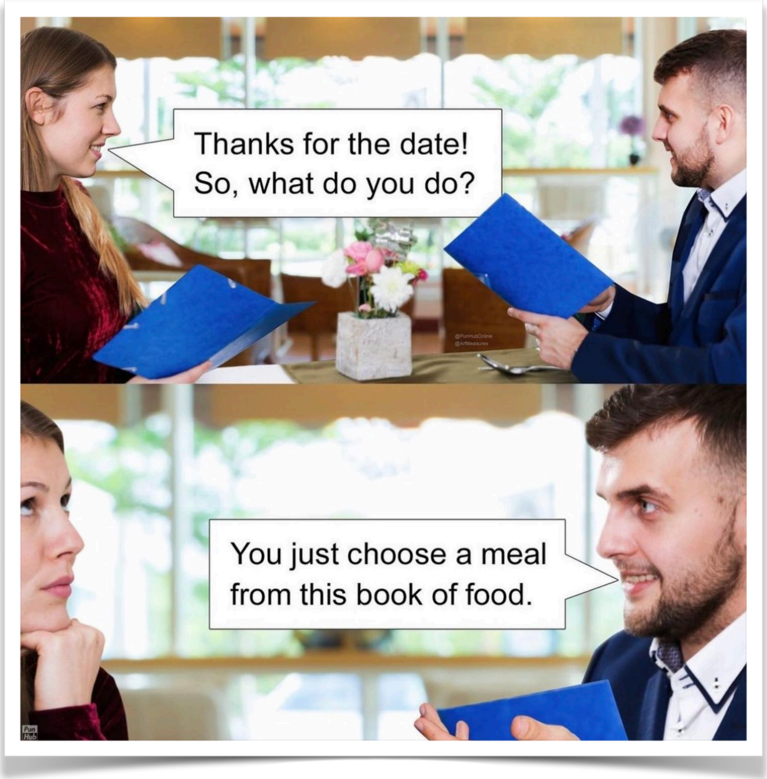


[Example from Lillian Lee]

Ambiguity

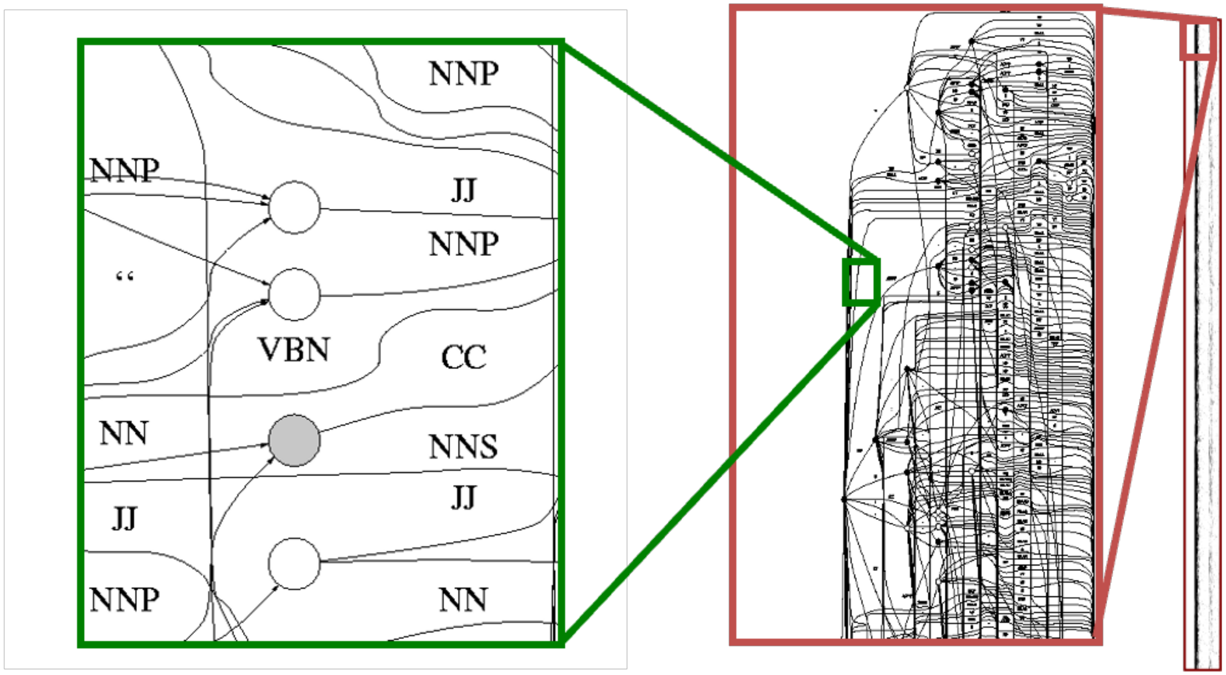
Pragmatic Ambiguity




Thanks for the date! So, what do you do?
You just choose a meal from this book of food.

Scale

- People did know that language was ambiguous!
  - …but they hoped that all interpretations would be “good” ones
  - …they didn’t realize how bad it would be



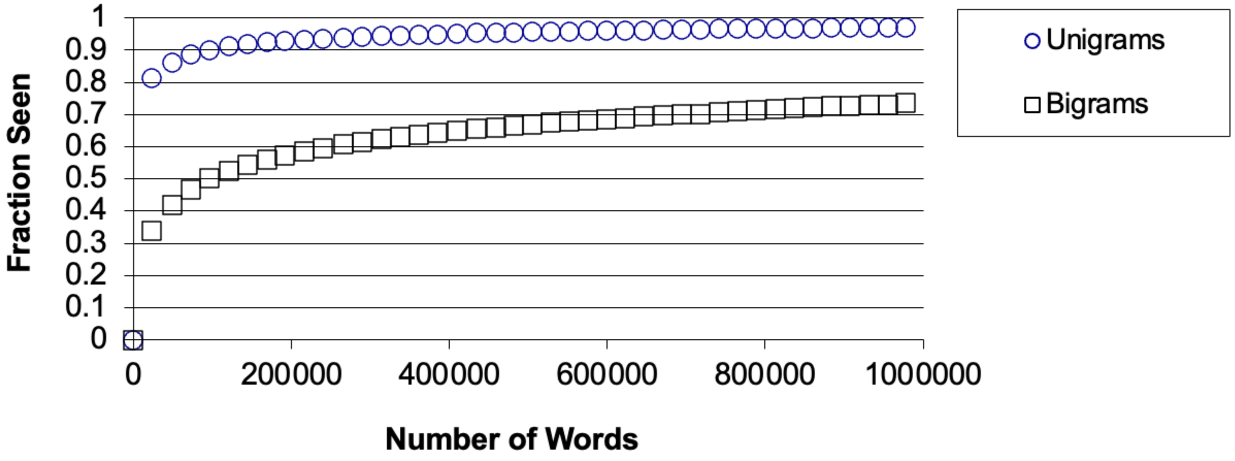
Sparsity

- A corpus is a collection of text
  - Annotated vs. raw text
  - Balanced vs. uniform corpora
- Examples:
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - RedPajama: 30T raw tokens
  - The entire Web?

Sparsity

- Raw corpora became extremely large
- But, sparsity is always a problem

New words/word pairs



Acknowledgements

We thank the following sources for presentation materials:

- University of Washington CSE 517 by Luke Zettlemoyer
- UT Austin CS 388 by Greg Durrett
- Berkeley CS 288 by Alane Suhr and Dan Klein (and older versions of the class by Dan Klein)