CS371N: Natural Language Processing

Lecture 1: Introduction

Greg Durrett
(he/him)

TEXAS
The University of Texas at Austin

Administrivia

- Lecture: Tuesdays and Thursdays 9:30am-10:45am in JGB 2.218
  - Recordings available afterwards on LecturesOnline
- Course website (including **syllabus**): http://www.cs.utexas.edu/~gdurrett/courses/fa2024/cs371n.shtml
- Ed Discussion board: link on Canvas
- Office hours: see course website and Canvas. Greg's are hybrid, some TA OHs are hybrid too. **Office hours start Thursday after class.**
- TAs: Juan Diego Rodriguez and Grace Kim.
- Office hours start today, and I will stay around after this class if you have questions

Course Requirements

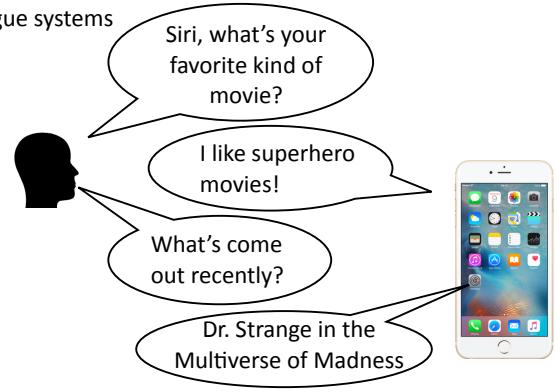
- CS 429
- Recommended: CS 331, familiarity with probability and linear algebra, programming experience in Python
- Helpful: Exposure to AI and machine learning (e.g., CS 342/343/363)
- Assignment 0 is out now (optional):
  - If this seems like it'll be challenging for you, come and talk to me (this is smaller-scale than the other assignments, which are smaller-scale than the final project)

Format and Accessibility

- Lectures will build in time for discussion, in-class exercises, and questions. Additional material is available as videos to watch either before or after lectures
  - Format: in-person to encourage discussion, but all materials are available asynchronously afterwards
- Equipment: useful to have a device for lecture to do Instapolls. For homework:
  - Lab machines available via SSH
  - A GPU is **not** required to complete the assignments! Having a GPU, GCP credits, or Google Colab access will be helpful for the final project though

What's the goal of NLP?

- Be able to solve problems that require deep understanding of text
- Example: dialogue systems



Machine Translation

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.

Question Answering

When was Abraham Lincoln born?

map to `Birthday` field

Name	Birthday
Lincoln, Abraham	2/12/1809
Washington, George	2/22/1732
Adams, John	10/30/1735

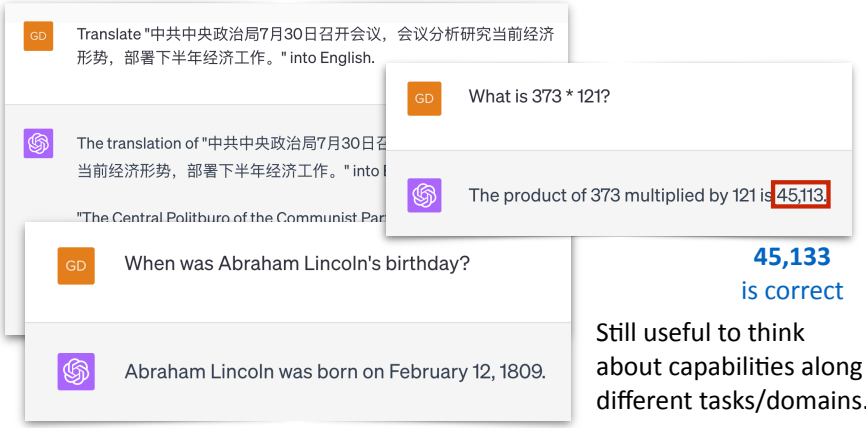
→ **February 12, 1809**

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

The park has a total of five visitor centers

five

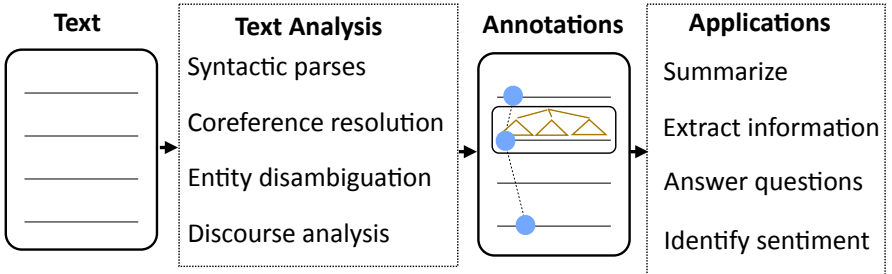
Generalist Systems?



45,133 is correct

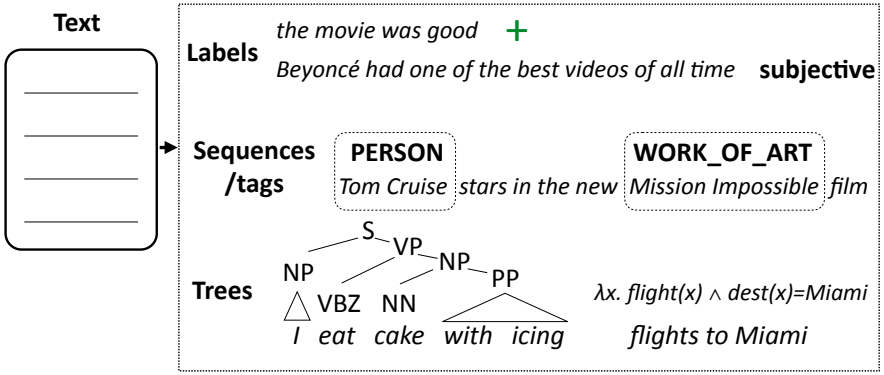
Still useful to think about capabilities along different tasks/domains.

Classical NLP Analysis Pipeline



- NLP is about building these pieces! (largely using statistical approaches)
- Lots of this is done end-to-end with neural nets. But analysis is still useful…

How do we represent language?



- Question: What ambiguities do these representations need to help us resolve?

Why is language hard?
(and how can we handle that?)

Language is Ambiguous!

- Hector Levesque (2011): "Winograd schema challenge" (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they advocated violence

The city council refused the demonstrators a permit because they feared violence

The city council refused the demonstrators a permit because they _____ violence

- >5 datasets in the last few years examining this problem and commonsense reasoning
- Referential ambiguity

Language is Ambiguous!

Teacher Strikes Idle Kids

Ban on Nude Dancing on Governor's Desk

Iraqi Head Seeks Arms

- Syntactic and semantic ambiguities: parsing needed to resolve these, but need context to figure out which parse is correct

example credit: Dan Klein

Language is **Really** Ambiguous!

- There aren't just one or two possibilities which are resolved pragmatically

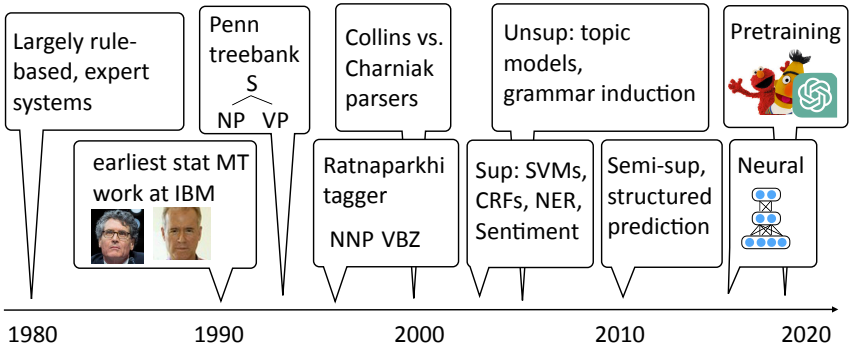
il fait vraiment beau →
- It is really nice out
- It's really nice
- The weather is beautiful
- It is really beautiful outside
- He makes truly beautiful
- It fact actually handsome

- Combinatorially many possibilities, many you won't even register as ambiguities, but systems still have to resolve them

What techniques do we use?

(to combine data, knowledge, linguistics, etc.)

A brief history of (modern) NLP



Pretraining

- Language modeling: predict the next word in a text $P(w_i|w_1, \ldots, w_{i-1})$

P(*w* | I want to go to) = 0.01 Hawai'i
0.005 LA
0.0001 class

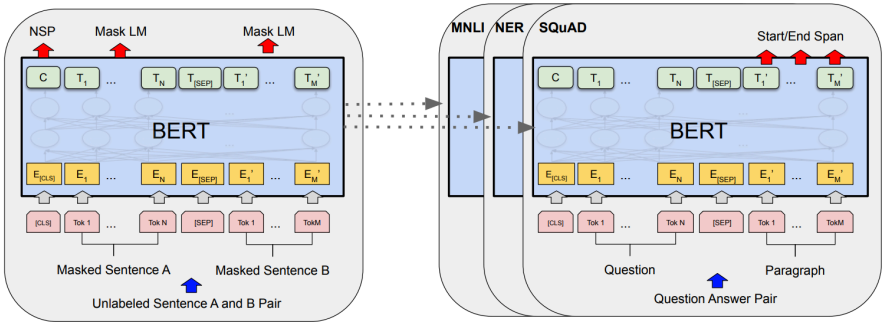
: use this model for other purposes

P(*w* | the acting was horrible, I think the movie was) = 0.1 bad
0.001 good

- Model understands some sentiment?
- Train a neural network to do language modeling on massive unlabeled text, fine-tune it to do {tagging, sentiment, question answering, …}

Peters et al. (2018), Devlin et al. (2019)

BERT



Pre-training Fine-Tuning

- Key parts which we will study: (1) Transformer architecture; (2) what data is used (both for pre-training and fine-tuning)

Devlin et al. (2019)

GPT and In-Context Learning

- Even more "extreme" setting: no gradient updates to model, instead large language models "learn" from examples in their context
- Many papers studying why this works. We will read some!

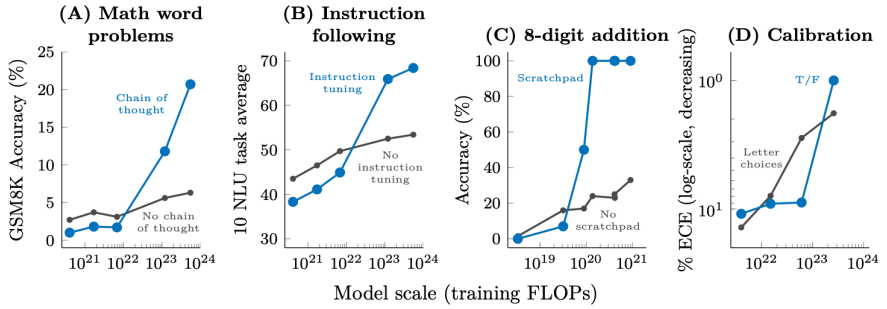
Few-shot

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

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

Brown et al. (2020)

Scaling Laws



- Many of the methods that work in LLMs today only make sense and only work because the models are so big!

Kaplan et al. (2020), Jason Wei et al. (2022)

Where are we?

- We have very powerful neural models that can fit lots of datasets
- Data: we need data that is not just correctly labeled, but reflects what we actually want to be able to do
- Users: systems are not useful unless they do something we want
- Language/outreach: who are we building this for? What languages/dialects do they speak?

Social Impact

- NLP systems are increasingly used in the world

…and increasingly we have to reckon with their impact

- This lecture: let's warm up by thinking about these issues a bit

Social Impact

- Rate your awareness of the social impact of NLP, AI, and machine learning from 1 to 5, where 1 is little awareness and 5 is strong awareness (5 = you feel like you could write a blog post about a current issue).
- Describe one scenario where you think deployment of an NLP system might pose ethical challenges *due to the application* itself (i.e., using NLP to do "bad stuff")
- Describe one scenario where you think deployment of an NLP system might pose ethical challenges due to *unintended* consequences (e.g., unfairness, indirectly causing bad things to happen, etc.).

Outline of the Course

- Classification: linear and neural, word representations (3.5 weeks)
- Language modeling, Transformers, pre-training (2.5 weeks)
- Tagging, parsing, and linguistic structure (2 weeks, ending in midterm)
- Modern pre-trained models, ChatGPT, etc. (2.5 weeks)
- Applications, modern topics, and ethics (2.5 weeks)
- Goals:
  - Cover fundamental techniques used in NLP
  - Understand how to look at language data and approach linguistic phenomena
  - Cover modern NLP problems encountered in the literature: what are the active research topics in 2023?

Coursework

- Five assignments, worth 40% of grade
  - Mix of writing and implementation;
  - Assignment 0 is out now, optional diagnostic
  - ~2 weeks per assignment except for A4
  - 5 “slip days” throughout the semester to turn in assignments 24 hours late
  - Submission on Gradescope

These assignments require understanding the concepts, writing performant code, and thinking about how to debug complex systems. **They are challenging; start early!**

Office hours: please come! However, **the course staff are not here to debug your code!** We **will** help you understand the concepts and come up with debugging strategies!

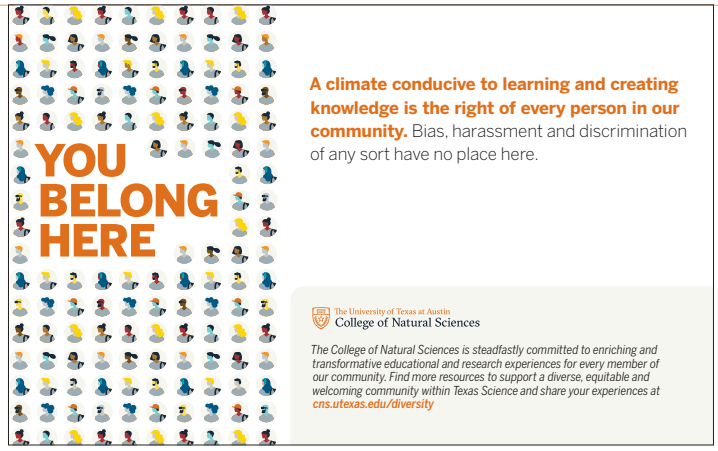
Coursework

- Midterm (25% of grade), take-home
  - Similar to written homework problems
- Final project (25% of grade)
  - Groups of 1 or 2
  - Standard project: understanding dataset biases
  - Independent projects are possible: these must be proposed earlier (to get you thinking early) and will be held to a high standard!
- Social Impact Responses, UT Instapoll (10% of the grade)
  - These will be done online and can be done during or after class

Academic Honesty

- You may work in groups, but your final writeup and code **must be your own**
- Don’t share code with others!

Conduct



Survey

- See Instapoll (you can answer later as well)