

CS378: Natural Language Processing

Lecture 1: Introduction

Greg Durrett



Credit: Stephen Roller



Administrivia

- ▶ Lecture: Tuesdays and Thursdays 9:30am - 10:45am
- ▶ Course website (including **syllabus**):
<http://www.cs.utexas.edu/~gdurrett/courses/fa2021/cs378.shtml>
- ▶ edstem board (~Piazza): link on the course website
- ▶ Office hours: see course website. Mix of Zoom (instructor and TAs) and on-campus (TAs). Please come!
- ▶ TAs: Kaj Bostrom and Abhilash Potluri (say hi!)
- ▶ All office hours start next week, but I will stay around after this class (including on Zoom) 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
 - ▶ We'll do plenty of discussion groups in class. Piazza is also available to find teammates
- ▶ 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 or GCP credits could be helpful **if** you want to pursue an independent project



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
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?

Name	Birthday	map to Birthday field
Lincoln, Abraham	2/12/1809	→ February 12, 1809
Washington, George	2/22/1732	
Adams, John	10/30/1735	

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

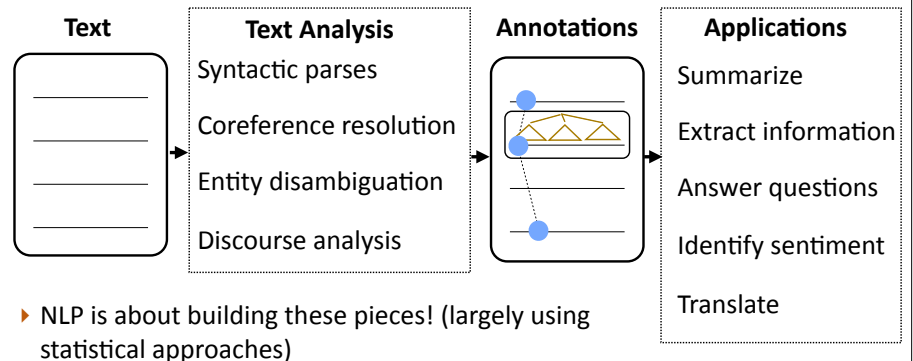


The park has a total of five
visitor centers

five



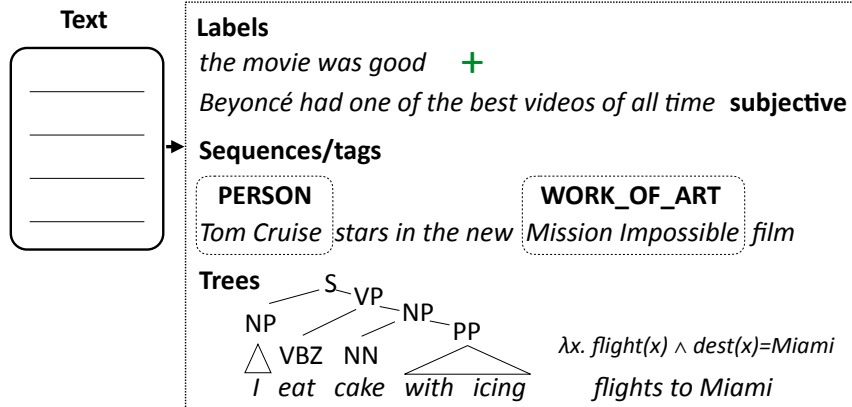
NLP Analysis Pipeline



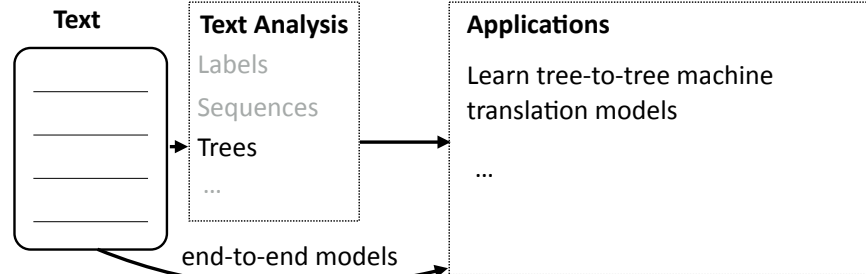
- ▶ NLP is about building these pieces! (largely using statistical approaches)



How do we represent language?



How do we use these representations?



- Main question: What representations do we need for language? What do we want to know about it? What ambiguities do we need to 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 two 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

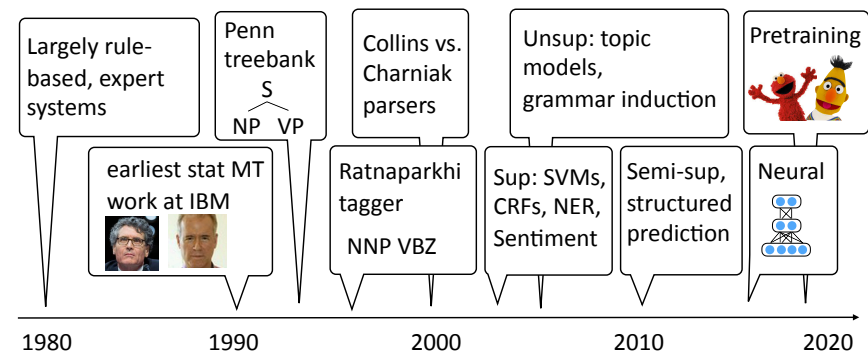
il fait vraiment beau → It is really nice out
It's really nice
The weather is beautiful
It is really beautiful outside
He makes truly beautiful
It fact actually handsome

- 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 $P(w_i | w_1, \dots, w_{i-1})$

$P(w \mid \text{I want to go to}) =$ 0.01 Hawai'i
0.005 LA
0.0001 class



: use this model for other purposes

$P(w \mid \text{the acting was horrible, I think the movie was}) =$ 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)



Interpretability

- When we have complex models, how do we understand their decisions?

The movie is mediocre, maybe even bad. **Negative** 99.8%

The movie is mediocre, maybe even bad. **Negative** 98.0%

The movie is mediocre, maybe even bad. **Negative** 98.7%

The movie is mediocre, maybe even bad. **Positive** 63.4%

The movie is mediocre, maybe even bad. **Positive** 74.5%

The movie is mediocre, maybe even bad. **Negative** 97.9%

The movie is mediocre, maybe even bad. Wallace, Gardner, Singh
Interpretability Tutorial at EMNLP 2020



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)
- ▶ Text analysis: tagging and parsing (3 weeks) <= takes us to the midterm
- ▶ Generation, applications: language modeling, machine translation (3 weeks)
- ▶ Question answering, pre-training (2 weeks)
- ▶ Applications and miscellaneous (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 2020?



Coursework

- ▶ Five assignments, worth 45% of grade (A1-4: 10%, A5: 5%)
 - ▶ Mix of writing and implementation;
 - ▶ Assignment 0 is out now, optional diagnostic
 - ▶ ~2 weeks per assignment except for A5
 - ▶ 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 (20% of grade), take-home
 - ▶ Similar to written homework problems
- ▶ Final project (25% of grade)
 - ▶ Groups of 1 or 2
 - ▶ Standard project: neural network models for question answering
 - ▶ 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



A climate conducive to learning and creating knowledge is the right of every person in our community. Bias, harassment and discrimination of any sort have no place here.

 The University of Texas at Austin
College of Natural Sciences

The College of Natural Sciences is steadfastly committed to enriching and transformative educational and research experiences for every member of our community. Find more resources to support a diverse, equitable and welcoming community within Texas Science and share your experiences at cns.utexas.edu/diversity



Survey

- ▶ See Instapoll (you can answer later as well)