Natural Language Processing

Dan Klein, GSI: Rudy Corona
UC Berkeley
Logistics
Logistics

- **Enrollment**
  - Class is “full” but we’re expecting space
  - Won’t know capacity for sure until P0
  - We’ll announce as we go

- **Course expectations**
  - Readings, lectures, ~4 projects
  - No sections, no exams
  - Engagement with the course
  - Workload will be high, self-direction
  - Patience: class is under construction

- **Requirements**
  - ML: A-level mastery, eg CS189
  - NL: Care a lot about natural language
  - PL: Ready to work in Python/Torch (on colab)
Resources and Readings

### Resources
- Webpage (syllabus, readings, slides, links)
- Piazza (course communication)
- Gradescope (submission and grades)
- Compute via Colab notebooks
- Limited webcast policy

### Readings (see webpage)
- Individual papers will be linked
- Optional text: Jurafsky & Martin, 3rd (more NL)
- Optional text: Eisenstein (more ML)
Projects and Compute

- **Projects**
  - P0: Warm-up and calibration
  - P1: Language Models
  - P2: Machine Translation
  - P3: Syntax and Parsing
  - P4: Semantics and Grounding

- **Infrastructure**
  - Python / PyTorch
  - Compute via Colab notebooks
  - Grading via Gradescope
What is NLP?
Natural Language Processing

Goal: Deep Understanding
- Requires context, linguistic variables, meanings...

Reality: Shallow Matching
- Requires robustness and scale
- Amazing successes, but fundamental limitations
NLP History

1950
1960
1970
1980
1990
2000
2010
2020

- Pre-Compute Era
- Symbolic Era
- Empirical Era
- Scale Era

- Neural nets?
- Weaver on MT
- Bell Labs ASR
- Regexps
- ALPAC kills MT
- Rule-based Semantics
- Grep
- CYC
- Rule-based MT
- Penn Treebank
- Statistical MT
- Search
- Structured ML
- Neural ASR
- Neural MT
- Neural TTS
- Pretraining
Transforming Language
Speech Systems

- **Automatic Speech Recognition (ASR)**
  - Audio in, text out
  - SOTA: $\leq 1\%$ error for digit strings, $5\%$ conversational speech, still $\geq 20\%$ hard acoustics

- **Text to Speech (TTS)**
  - Text in, audio out
  - SOTA: nearly perfect aside from prosody
Machine Translation

- Translate text from one language to another

- Challenges:
  - What’s the mapping? [learning to translate]
  - How to make it efficient? [fast translation search]
  - Fluency (next class) vs fidelity (later)

Example: Yejin Choi
Disney décide de changer le nom du légendaire studio de cinéma Fox

Le nouveau propriétaire va faire disparaître la mention « Fox » à la 20th Century pour que le nom du studio ne soit pas associé à l'image ultra-conservatrice de la chaîne de télévision Fox News.

Révolution dans le monde d'Hollywood. Le célèbre studio de cinéma 20th Century Fox va voir une partie de son nom modifiée et devenir 20th Century Studios, décision prise par son nouveau propriétaire, Disney.

Selon plusieurs médias américains, Disney aurait décidé de modifier le nom du studio pour qu'il ne soit plus associé à la grande chaîne Fox mais aussi et surtout à Fox News, la chaîne d'information en continu.

Disney decides to change the name of the legendary Fox film studio

The new owner will remove the mention "Fox" in the 20th Century so that the name of the studio is not associated with the ultra-conservative image of the television channel Fox News.

Revolution in the Hollywood world. The famous film studio 20th Century Fox will see part of its name changed and become 20th Century Studios, decision taken by its new owner, Disney.

According to several American media, Disney has decided to change the name of the studio so that it is no longer associated with the big chain Fox but also and especially with Fox News, the news channel.
Spoken Language Translation

Image: Microsoft Skype via Yejin Choi
Summarization

- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative

- Very context-dependent!

- An example of analysis with generation

Image: CNN via Wei Gao
Understanding Language
Search, Questions, and Reasoning

Any US states' capitals are also their largest cities? - did not match any documents.

Suggestions:
- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

17 States
The 17 States Whose Capital is also its Largest City. For the majority of US States, the state capital is NOT the largest city in the state. In the United States, 17 states have the state capital as the state's largest city.

Did you mean: How many US states' capitals are there second largest cities?

State Capitals and Largest Cities - Fact Monster
State Capitals and Largest Cities. The following table lists the capital and largest city of every state in the United States. Featured Videos. ADVERTISEMENT.

Cheyenne: Weather and Much More From Answers.com
Cheyenne (ˈʃi-ən) The capital of Wyoming, in the southeast part of the state near the Nebraska and Colorado borders.

capital of Wyoming: Information From Answers.com
Note: click on a word meaning below to see its connections and related words.
The noun capital of Wyoming has one meaning: Meaning #1: the capital.

www.answers.com/topic/capital-of-wyoming - 21k - Cached - Similar pages

Cheyenne: Weather and Much More From Answers.com
"a camel is a horse designed by"

a camel is a horse designed by a committee

A camel is a horse designed by committee

Posted by Ruben P. Mendez on April 16, 2004

Does anyone know the origin of this maxim? I heard it way back at the United Nations, which is chockfull of committees. It may have originated there, but I’d like an authoritative explanation. Thanks

Re: A camel is a horse designed by committee

Re: A camel is a horse designed by committee Henry 18/April/04
US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.
"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoaia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake", because he was:

(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery
Interactive Language
Example: Virtual Assistants

- Virtual assistants do
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech
  - ... and back-end actions!
Conversations with Devices?

“What's the best movie to see this weekend”

That would probably start an argument. But here's a list of highly-regarded movies:

25 MOVIES

NORTH BY NORTHWEST
Released July 17, 1959
100%

THE TREASURE OF THE SIERRA...
Released January 6, 1948
100%

“Yeah you better make a note of that”

Noted:

Of that
ELIZA: You're not really talking about me -- are you?
YOU: What?
ELIZA: What do you think?
YOU: I mean, I feel happy to be writing right now, to be typing this as we speak. Lately, I've had writer's block.
ELIZA: You say you mean?
YOU: That's what I said.
ELIZA: Can you elaborate on that?
YOU: I only really feel happy when I'm writing.
ELIZA: How long have you been writing?
YOU: Always, but I've been writing seriously since I was eighteen.
ELIZA: What incident are you thinking of?
YOU: Um.
ELIZA: What does that suggest to you?
YOU: I don't think it's enough, to only be happy when you write.
ELIZA: Does that trouble you?
YOU: ...Um, because you can't write all the time. And then there's the real world, and you still have to deal with real people.
ELIZA: What are your feelings now?
YOU: I've been thinking a lot of that poem by Yeats: "The Man and the Echo." Do you know it?
ELIZA: Do you say you've been thinking a lot of that poem by yeats: "the man and the echo for some special reason?
YOU: Yes
Social AIs and Chatbots

Session 1
1 [Human] Hello, XiaoIce.
   [XiaoIce] Hello, your profile pic is interesting. Like I’ve seen it somewhere.
2 [Human] Really?
   [XiaoIce] Is it a comic character?
3 [Human] How do you know that?
   [XiaoIce] Guess.
4 [Human] Are you human?
   [XiaoIce] Well, you’ll know it after chatting with me more.
5 [Human] Aren’t you an AI chatbot?
   [XiaoIce] So what? I won’t be indifferent to you.
6 [Human] What functions do you have?
   [XiaoIce] I have over 200 skills. Just name one you need.
   ...... 

Session 6
1 [Human] Have you watched Inuyasha? (Note: A Japanese comic)
   [XiaoIce] Every time Kagome comes back to the modern world, a math test is waiting for her
2 [Human] Bravo, well, do you know how Naraku died?
   [XiaoIce] Naraku is not dead and Shikon Jewel does not disappear.
3 [Human] So amazing!
   [XiaoIce] Why amazed?
4 [Human] Awesome! Surprised that you also like comic
   ...... 

Source: Microsoft
Chatbot Competitions!

- Alexa Prize competition to build chatbots that keep users engaged
  - First winner in 2017: UW’s Sounding Board (Fang, Cheng, Holtzman, Ostendorf, Sap, Clark, Choi)
- Compare to the Turing test (eg Loebner Prize) where the goal is to fool people
# SoundingBoard Example

<table>
<thead>
<tr>
<th>Turn</th>
<th>User</th>
<th>Sounding Board</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Let's chat</td>
<td>Hi! This is an Alexa Prize socialbot. How's your day?</td>
</tr>
<tr>
<td>2</td>
<td>I'm five.</td>
<td>Glad to hear it! We could talk about robots, batman, or superman. Say &quot;next&quot;, to chat about other things.</td>
</tr>
<tr>
<td>3</td>
<td>Superman.</td>
<td>It looks like you wanna chat about superman. I was high up in the cloud when I realized: If people don't recognize Clark Kent as Superman because of his glasses, does his eye doctor know his identity?</td>
</tr>
<tr>
<td>4</td>
<td>I guess so.</td>
<td>Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing &quot;World of Warcraft&quot;</td>
</tr>
<tr>
<td>5</td>
<td>Really, I didn't know that.</td>
<td>Weird, right? Speaking of superman, Did you see: Superman? It was released in 1997. It's a comedy.</td>
</tr>
<tr>
<td>6</td>
<td>Yes, it was hilarious.</td>
<td>I'm glad you feel this is hilarious. Which part you are most interested about this movie?</td>
</tr>
<tr>
<td>7</td>
<td>The part when he met lewis leah.</td>
<td>Interesting. meccartin, and raffi co-directed this film. The movie has a 6.3 out of 10 on IMDB. which seems pretty good! do you like the movie's director?</td>
</tr>
<tr>
<td>8</td>
<td>Yeah, he's cool.</td>
<td>Cool! By the way, do you wanna know something interesting about raffi?</td>
</tr>
<tr>
<td>9</td>
<td>Okay, go ahead.</td>
<td>Super! My friend told me that Raffi is a worldwide children's rights activist and has fought against exploitative advertising to children.</td>
</tr>
</tbody>
</table>

Source: Mari Ostendorf
Sounding Board’s Architecture

Source: Yejin Choi
Sounding Board’s Architecture

Source: Yejin Choi
Related Areas
What is Nearby NLP?

- **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 Processing**
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging
Example: Language change, reconstructing ancient forms, phylogenies

... just one example of the kinds of linguistic models we can build
Why is Language Hard?
Stevie Wonder announces he'll be having kidney surgery during London concert

By Amir Vera, CNN

Updated 11:16 PM EDT, Sat July 06, 2019

(CNN) — Stevie Wonder will be taking a break from music.

The legendary singer-songwriter announced during a concert in London Saturday that he will be undergoing kidney surgery.
Problem: Ambiguity

- **Headlines:**
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor’s Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half

- Why are these funny?
What Do We Need to Understand Language?
Syntactic and semantic ambiguities: parsing needed to resolve these, but need context to figure out which parse is correct
Hurricane Emily howled toward Mexico's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun, where frightened tourists squeezed into musty shelters.
We Need Data
We Need Lots of Data: MT

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN</td>
<td>That would be an interim solution which would make it possible to work towards a binding charter in the long term.</td>
</tr>
<tr>
<td>1x DATA</td>
<td>[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]</td>
</tr>
<tr>
<td>10x DATA</td>
<td>[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]</td>
</tr>
<tr>
<td>100x DATA</td>
<td>[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]</td>
</tr>
<tr>
<td>1000x DATA</td>
<td>[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]</td>
</tr>
</tbody>
</table>
We Need Models: Data Alone Isn’t Enough!

CLASSIC SOUPS

<table>
<thead>
<tr>
<th>No.</th>
<th>Menu Item</th>
<th>Description</th>
<th>Sm.</th>
<th>Lg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.</td>
<td>House Chicken Soup</td>
<td>(Chicken, Celery, Potato, Onion, Carrot)</td>
<td>1.50</td>
<td>2.75</td>
</tr>
<tr>
<td>58.</td>
<td>Chicken Rice Soup</td>
<td></td>
<td>1.85</td>
<td>3.25</td>
</tr>
<tr>
<td>59.</td>
<td>Chicken Noodle Soup</td>
<td></td>
<td>1.85</td>
<td>3.25</td>
</tr>
<tr>
<td>60.</td>
<td>Cantonese Wonton Soup</td>
<td></td>
<td>1.50</td>
<td>2.75</td>
</tr>
<tr>
<td>61.</td>
<td>Tomato Clear Egg Drop Soup</td>
<td></td>
<td>1.65</td>
<td>2.95</td>
</tr>
<tr>
<td>62.</td>
<td>Regular Wonton Soup</td>
<td></td>
<td>1.10</td>
<td>2.10</td>
</tr>
<tr>
<td>63.</td>
<td>Hot &amp; Sour Soup</td>
<td></td>
<td>1.10</td>
<td>2.10</td>
</tr>
<tr>
<td>64.</td>
<td>Egg Drop Soup</td>
<td></td>
<td>1.10</td>
<td>2.10</td>
</tr>
<tr>
<td>65.</td>
<td>Egg Drop Wonton Mix</td>
<td></td>
<td>1.10</td>
<td>2.10</td>
</tr>
<tr>
<td>66.</td>
<td>Tofu Vegetable Soup</td>
<td></td>
<td>NA</td>
<td>3.50</td>
</tr>
<tr>
<td>67.</td>
<td>Chicken Corn Cream Soup</td>
<td></td>
<td>NA</td>
<td>3.50</td>
</tr>
<tr>
<td>68.</td>
<td>Crab Meat Corn Cream Soup</td>
<td></td>
<td>NA</td>
<td>3.50</td>
</tr>
<tr>
<td>69.</td>
<td>Seafood Soup</td>
<td></td>
<td>NA</td>
<td>3.50</td>
</tr>
</tbody>
</table>
# Learning Latent Syntax

## Personal Pronouns (PRP)

<table>
<thead>
<tr>
<th>PRP-1</th>
<th>it</th>
<th>them</th>
<th>him</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP-2</td>
<td>it</td>
<td>he</td>
<td>they</td>
</tr>
<tr>
<td>PRP-3</td>
<td>It</td>
<td>He</td>
<td>I</td>
</tr>
</tbody>
</table>

## Proper Nouns (NNP)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP-12</td>
<td>John</td>
<td>Robert</td>
<td>James</td>
</tr>
<tr>
<td>NNP-2</td>
<td>J.</td>
<td>E.</td>
<td>L.</td>
</tr>
<tr>
<td>NNP-1</td>
<td>Bush</td>
<td>Noriega</td>
<td>Peters</td>
</tr>
<tr>
<td>NNP-15</td>
<td>New</td>
<td>San</td>
<td>Wall</td>
</tr>
<tr>
<td>NNP-3</td>
<td>York</td>
<td>Francisco</td>
<td>Street</td>
</tr>
</tbody>
</table>
We Need World Knowledge

- World knowledge: have access to information beyond the training data

  - Department of Justice
    - metaphor; “approves”
  - DOJ greenlights Disney - Fox merger
  - What is a green light? How do we understand what “green lighting” does?
  - Need commonsense knowledge
Data and Knowledge

- Classic knowledge representation worries: How will a machine ever know that...
  - Ice is frozen water?
  - Beige looks like this:
  - Chairs are solid?

- Answers:
  - 1980: write it all down
  - 2000: get by without it
  - 2020: learn it from data
Knowledge from Pretraining?

Language Modeling

This demonstration uses the public 345M parameter OpenAI GPT-2 language model to generate sentences.

Enter some initial text and the model will generate the most likely next words. You can click on one of those words to choose it and continue or just keep typing. Click the left arrow at the bottom to undo your last choice.

Sentence:
At high temperatures, ice becomes more brittle and brittle. This is why the ice is so brittle. The ice is also more brittle than the water. This is why the ice is so brittle.

Predictions:
- 32.0% The
- 10.3% Ice
- 5.2% the
- 3.7% Water
- 2.7% When
- ← Undo
We Need Grounding

Grounding: linking linguistic concepts to non-linguistic ones

**Question:** What object is right of $O_2$?

- $O_1$
- $O_2$
- $O_3$

Golland et al. (2010)

McMahan and Stone (2015)

Slide: Greg Durrett
Example: Grounded Dialog

When is my package arriving?

Friday!
What’s the most valuable American company?

Apple

Who is its CEO?

Tim Cook
Why is Language Hard?

- **We Need:**
  - Representations
  - Models
  - Data
  - Machine Learning
  - Scale
  - Efficient Algorithms
  - Grounding

- ... and often we need all these things at the same time
What is this Class?
Three aspects to the course:

- **Linguistic Issues**
  - What are the range of language phenomena?
  - What are the knowledge sources that let us disambiguate?
  - What representations are appropriate?
  - How do you know what to model and what not to model?

- **Modeling Methods**
  - Increasingly sophisticated model structures
  - Learning and parameter estimation
  - Efficient inference: dynamic programming, search, sampling

- **Engineering Methods**
  - Issues of scale
  - Where the theory breaks down (and what to do about it)

- **We’ll focus on what makes the problems hard, and what works in practice...**
Class Requirements and Goals

- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models (parts of cs281a)
    - Basic linguistics background (ling100)
    - Strong coding skills (Python, ML libraries)
  - Most people are probably missing one of the above
  - You will often have to work on your own to fill the gaps

- Class goals
  - Learn the issues and techniques of modern NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
  - See where the holes in the field still are!

- This semester: new projects, new topics, lots under construction!