Natural Language Processing

Dan Klein, GSI: Rudy Corona
UC Berkeley

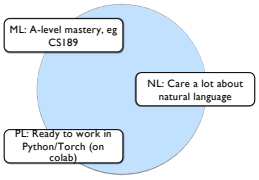
1

Logistics

2

Logistics

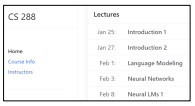
- Enrollment
  - Class is "full" but we're expecting space
  - Won't know capacity for sure until P0
  - We'll announce as we go
- Course expectations
  - Readings, lectures, ~4 projects
  - No sections, no exams
  - Engagement with the course
  - Workload will be high, self-direction
  - Patience: class is under construction
- Requirements

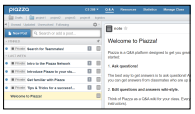


3

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)





4

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

5

What is NLP?

7

Natural Language Processing

Goal: Deep Understanding

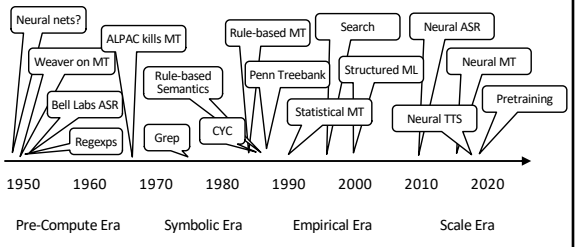
- Requires context, linguistic variables, meanings…

Reality: Shallow Matching

- Requires robustness and scale
- Amazing successes, but fundamental limitations

8

NLP History



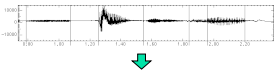
9

Transforming Language

10

Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: <<1% error for digit strings, 5% conversational speech, still >>20% hard acoustics



"Speech Lab"

Microsoft reaches 'human parity' with new speech recognition system

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

Google launches more realistic text-to-speech service powered by DeepMind's AI

Speak-N-Spell / Google WaveNet / The Verge

11

Machine Translation

- Translate text from one languag
- 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

12

Machine Translation

Google Translate 2020

14

Spoken Language Translation

Image: Microsoft Skype via Yejin Choi

16

Summarization

- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative
- Very context-dependent!
- An example of analysis with generation

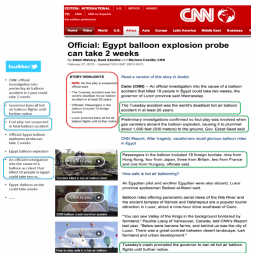


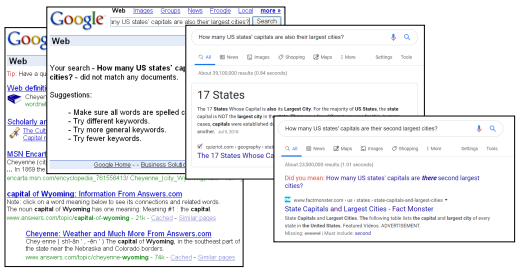
Image: CNN via Wei Gao

17

Understanding Language

18

Search, Questions, and Reasoning



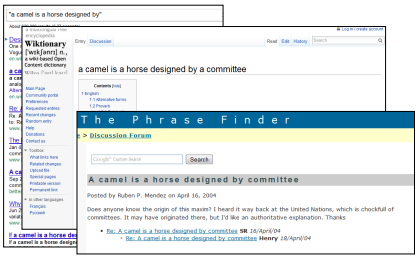
19

Jeopardy!

Images: Jeopardy Productions

20

Question Answering: Watson



21

Question Answering: Watson

Jeopardy!
World Champion

US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.

Slide: Yejin Choi

22

Language Comprehension?

"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 hands always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia 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

25

Interactive Language

26

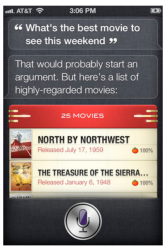
Example: Virtual Assistants

- Virtual assistants do
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech
  - … and back-end actions!

Image: Wikipedia

27

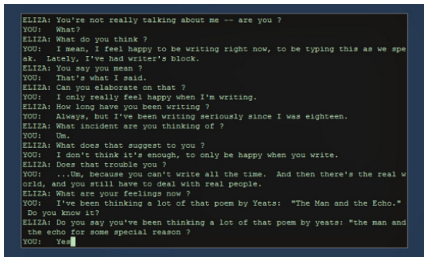
Conversations with Devices?



Slide: Yejin Choi

28

ELIZA



29

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

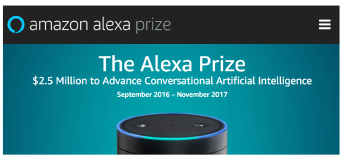
1 [Human] Have you watched Inuyasha? (Note: A Japanese comic)
[Xiaoice] Every time Kagome comes back to the modern world, a math test is waiting for her
2 [Human] Bravo, well, do you know how Naraku died?
[Xiaoice] Naraku is not dead and Shikon Jewel does not disappear.
3 [Human] So amazing!
[Xiaoice]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

Microsoft's Xiaolce

Source: Microsoft

30

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

31

SoundingBoard Example

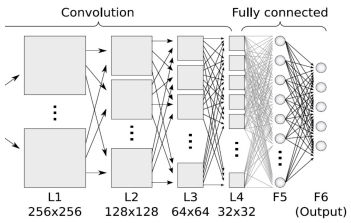
Turn	User	Sounding Board
1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
2	I'm five.	Glad to hear it! We could talk about robots, batman, or superman. Say "next", to chat about other things.
3	Superman.	It looks like you wanna chat about superman. I was high up in the cloud when I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

Negotiation
Thought
Movie
Fact

Source: Mari Ostendorf

32

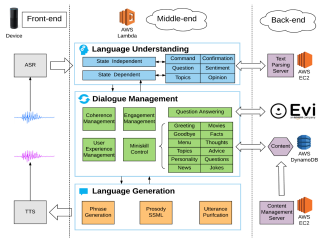
Sounding Board's Architecture



Source: Yejin Choi

33

Sounding Board's Architecture



Source: Yejin Choi

34

Related Areas

35

What is Nearby NLP?

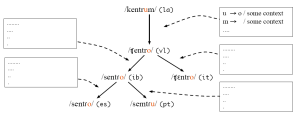
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it

- 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

36

Example: NLP Meets CL



Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Center	centrum	centro	centro	centro

- Example: Language change, reconstructing ancient forms, phylogenies

... just one example of the kinds of linguistic models we can build

37

Why is Language Hard?

38

Ambiguities

39

Problem: Ambiguity

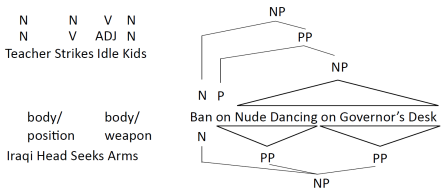
- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor's Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Why are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

40

What Do We Need to Understand Language?

41

We Need Representation: Linguistic Structure



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

Slide: Greg Durrett

42

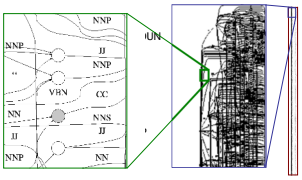
Example: Syntactic Analysis

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 .

Accuracy: 95+

43

We Need Data



44

We Need Lots of Data: MT

SOURCE	Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.
HUMAN	That would be an interim solution which would make it possible to work towards a binding charter in the long term .
1x DATA	[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]
10x DATA	[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]
100x DATA	[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]
1000x DATA	[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]

45

We Need Models: Data Alone Isn't Enough!

CLASSIC SOUPS

			Sm.	Lg.
57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)		1.50	2.75
58.	Chicken Rice Soup		1.85	3.25
59.	Chicken Noodle Soup		1.85	3.25
60.	Cantonese Wonton Soup		1.50	2.75
61.	Tomato Clear Egg Drop Soup		1.65	2.95
62.	Regular Wonton Soup		1.10	2.10
63.	Hot & Sour Soup		1.10	2.10
64.	Egg Drop Soup		1.10	2.10
65.	Egg Drop Wonton Mix		1.10	2.10
66.	Tofu Vegetable Soup		NA	3.50
67.	Chicken Corn Cream Soup		NA	3.50
68.	Crab Meat Corn Cream Soup		NA	3.50
69.	Seafood Soup		NA	3.50

Example from Adam Lopez

46

Learning Latent Syntax

Personal Pronouns (PRP)

PRP-1	it	them	him
PRP-2	it	he	they
PRP-3	It	He	I

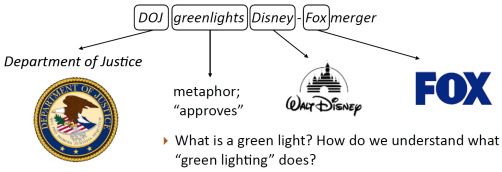
Proper Nouns (NNP)

NNP-14	Oct.	Nov.	Sept.
NNP-12	John	Robert	James
NNP-2	J.	E.	L.
NNP-1	Bush	Noriega	Peters
NNP-15	New	San	Wall
NNP-3	York	Francisco	Street

47

We Need World Knowledge

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



- What is a green light? How do we understand what "green lighting" does?
- Need commonsense knowledge

Slide: Greg Durrett

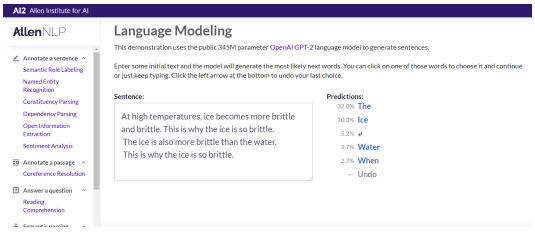
48

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

49

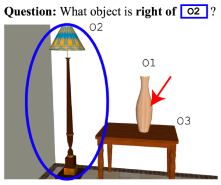
Knowledge from Pretraining?



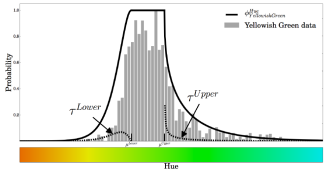
50

We Need Grounding

Grounding: linking linguistic concepts to non-linguistic ones



Golland et al. (2010)

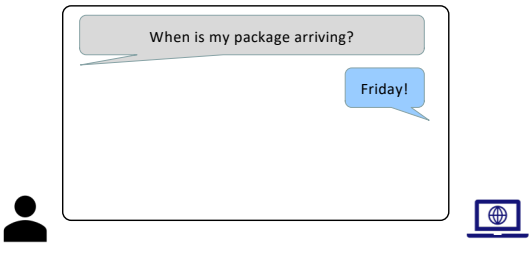


McMahan and Stone (2015)

Slide: Greg Durrett

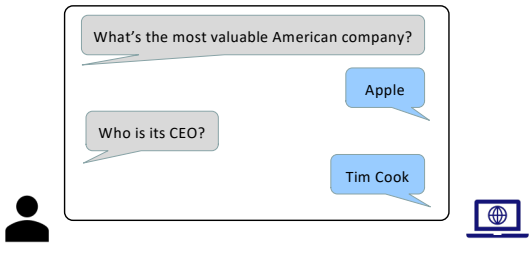
51

Example: Grounded Dialog



52

Example: Grounded Dialog



53

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

54

What is this Class?

55

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…

56

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!

57