

Natural Language Processing



Dan Klein, GSI: Nick Tomlin
UC Berkeley

Logistics



Logistics

Enrollment

- Class is “full” but we’re scaling up
- We will process waitlist after P1
- No materials require enrollment

Course expectations

- Readings, lectures, ~8 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

PL: Ready to work in PyTorch (on colab)

NL: Care a lot about natural language

COVID Policies

- Remote lecture / chat format for now
- Expanded late day policy (14 day)



Resources and Readings

Resources

- Webpage (syllabus, readings, slides, links)
- Piazza (course communication)
- Gradescope (submission and grades)
- Compute via Colab notebooks
- Berkeley-internal webcasts / recordings

Readings (see webpage)

- Individual papers will be linked
- Optional text: Jurafsky & Martin, 3rd (more NL)
- Optional text: Eisenstein (more ML)

CS 288	Lectures
Home	Jan 18: Introduction
Course Info	Jan 20: Language Modeling
Instructors	Jan 25: Neural Language Modeling

A screenshot of the Piazza platform interface. At the top, there's a navigation bar with 'New Post', 'Search or add a post...', and other account-related options. Below the navigation, a sidebar on the left lists course announcements and posts. The main area shows a feed of posts from students, with one prominent post by 'Dan Klein' titled 'Welcome to Piazza!' containing introductory text and a link to the syllabus. The interface is clean with a white background and blue/purple accents for buttons and links.



Projects and Infrastructure

- Projects

- P1: Language Models
- P2: Question Answering
- P3: Machine Translation
- P4: Speech
- P5: Syntax and Parsing
- P6: Semantics
- P7: Grounding
- P8: Historical Linguistics / TBD

```

File Edit View Insert Runtime Tools Help
+ Code + Text
[7]: def evaluate(network, data, eval_data):
    predictions = get_predictions(network, data)
    return f1score(predictions, eval_data)

Now we are ready to run everything and get our outputs. A correct implementation should get a validation score around 70%.
❶ trained_network = train()
test_predictions = get_predictions(trained_network, test_inputs)
save_predictions(test_predictions, predicted_test_outputs.txt")
...
83.300390096595
86.0000000000000
87.16791279949875
88.559812889769

```

- Infrastructure

- Python / PyTorch
- Compute via Colab notebooks
- Grading via Gradescope

What is NLP?



Natural Language Processing



Goal: Deep Understanding

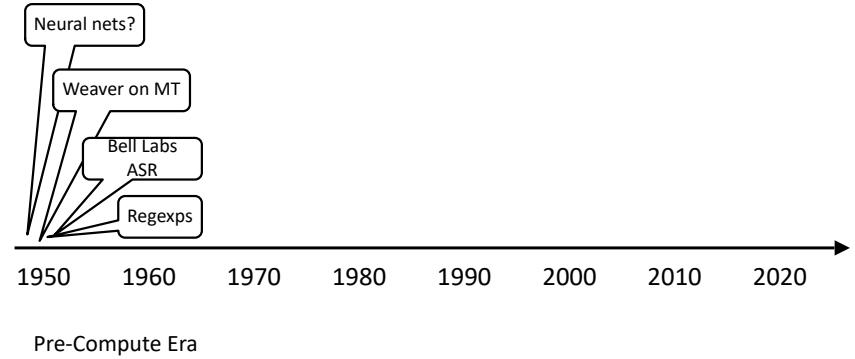
- Requires knowledge, context, and grounding
- Just starting to see successes

Reality: Shallow Matching

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

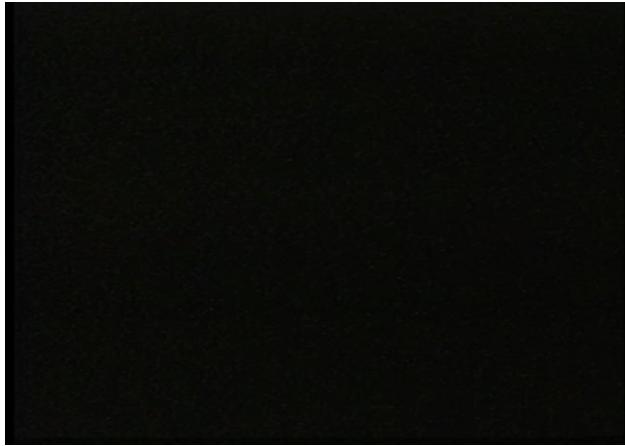


NLP History

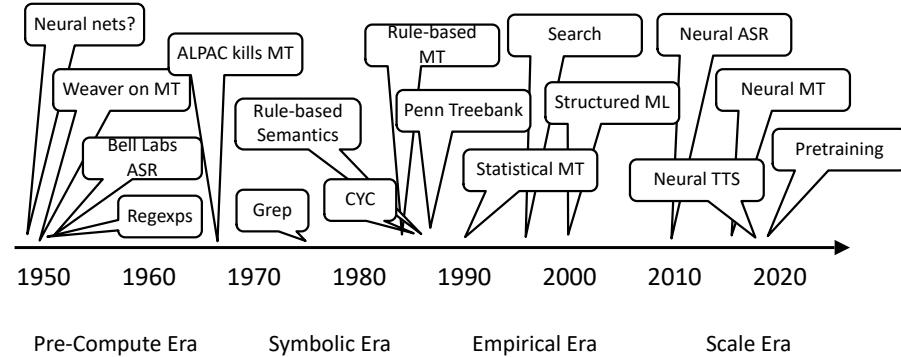




NLP History



NLP History



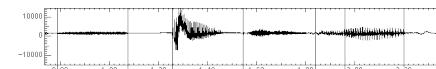
Transforming Language



Speech Systems

Automatic Speech Recognition (ASR)

- Audio in, text out
- SOTA: <<1% error for digit strings, 5% conversational speech, still >>20% hard acoustics



Microsoft reaches 'human parity' with new speech recognition system

“Speech Lab”

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

Machine Translation

EN CE MOMENT Impôts Kenya Syrie Pakistan Emploi Scandale Prism

Impôt sur le revenu : combien paierez-vous en 2014 ?

IMPOTS

Saisissez votre revenu et votre situation familiale pour découvrir si vous bénéficierez de la prime fiscale.

Comment le budget 2014 est réparti ? à visuel interactif

Un budget 2014 soumis aux critiques

Budget : ces dépenses ou aucun gouvernement ne peut diminuer

Budget 2014 : le détail des économies à visuel interactif

AT THIS MOMENT Taxes Kenya Syria Pakistan Use Prism scandal

Income tax: how much do you pay in 2014?

IMPOTS

Member(s) of Europe Ecology-Greens, do you share the finding of severe Christmas Mamère EELV?

Share your experience

Continuous

Géolocalisation Paris Actual Values' he arrested digital game of the NBA in Brussels i

Select your income
How is the budget allocated?
Le budget 2014 : le détail des économies à visuel interactif

- Translate text from one language
- 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

Machine Translation

CINÉMA - MÉDIAS

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.

Le Monde avec AFP - Publié le 18 janvier 2020 à 06h49

Lecture 1 min.

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.

Google Translate 2020

Partage

CINEMA - MEDIA

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.

Le Monde avec AFP - Posted on January 18, 2020 at 06:49

Reading time 1 min.

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.

Sharing



Spoken Language Translation

This call may be recorded for quality or training purposes.

skype

Hi grandma. Happy birthday. Are you having a great day?

Hola abuela. Feliz cumpleaños. ¿Tienes un buen día?

Hi Dylan. Yes, a great day thanks. And the flowers that you have sent are precious.

Hola Dylan. Sí, un día estupendo gracias. Y las flores que has enviado son preciosas.

I'm glad they arrived.

Me alegra que llegaron.

They arrived first thing in the morning.

Llegaron a primera hora de la mañana.

Type a message here

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

Official: Egypt balloon explosion probe can take 2 weeks

By Adam Mosley, Sean Aronoff and Marlene Cuthbert, CNN

Read a version of this story in Arabic.

Egypt (CNN) -- An official investigation into the cause of a hot air balloon explosion that killed 19 people in Egypt could take two weeks, the governor of Luxor province said Wednesday.

The Tuesday accident was the world's deadliest hot air balloon crash in at least 20 years.

Egypt (CNN) -- An official investigation into the cause of a hot air balloon explosion that killed 19 people in Egypt could take two weeks, the governor of Luxor province said Wednesday.

The Tuesday accident was the world's deadliest hot air balloon crash in at least 20 years.

Passengers on the balloon included 10 foreign tourists: nine from Hong Kong, four from Japan, three from Britain, two from France and one from Hungary, officials said.

How safe is hot air ballooning?

An Egyptian company offering hot air balloon flights over Luxor, which is known for its ancient pyramids and temples, confirmed no hot air was involved when the balloons above the balloon exploded, causing it to plummet about 1,000 feet (300 meters) to the ground, Gov. Ezzat Saad said.

CNN Report: After tragedy, vacationers recall giddy balloon rides in Egypt

Tuesday's crash prompted the governor to ban all hot air balloon flights until further notice.

Image: CNN via Wei Gao

Understanding Language



Search, Questions, and Reasoning

The screenshot shows a Google search results page. The query is "How many US states' capitals are also their largest cities?". The search bar has a note: "Your search - How many US states' capitals are also their largest cities? - did not match any documents." Below the search bar are suggestions and a link to "The 17 States". The main search results include links to "capital of Wyoming: Information From Answers.com", "Cheyenne: Weather and Much More From Answers.com", and "www.factmonster.com/us/states/state-capitals-and-largest-cities". A sidebar on the right shows a related search for "How many US states' capitals are their second largest cities?" with results from "quizlet.com" and "factmonster.com".

Jeopardy!



Images: Jeopardy Productions



Question Answering: Watson

The screenshot shows a Wiktionary entry for the phrase "a camel is a horse designed by committee". The entry is defined as a wikt:Open quote. It includes a sidebar with recent changes, donations, and a discussion forum. Below the Wiktionary entry is a screenshot of "The Phrase Finder" forum thread. The thread asks if anyone knows the origin of the maxim and receives a response from Ruben P. Mendez on April 16, 2004.



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



Language Comprehension?

Opera refers to a dramatic art form, originating in Europe, in which the emotional content is conveyed to the audience as much through music, both vocal and instrumental, as it is through the lyrics. By contrast, in musical theater an actor's dramatic performance is primary, and the music plays a lesser role. The drama in opera is presented using the primary elements of theater such as scenery, costumes, and acting. However, the words of the opera, or libretto, are sung rather than spoken. The singers are accompanied by a musical ensemble ranging from a small instrumental ensemble to a full symphonic orchestra.

1. It is pointed out in the reading that opera ----.

- A) has developed under the influence of musical theater
- B) is a drama sung with the accompaniment of an orchestra
- C) is not a high-budget production
- D) is often performed in Europe
- E) is the most complex of all the performing arts

[From GrammarBank]

Interactive Language



Example: Virtual Assistants



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

Image: Wikipedia

Conversations with Devices?



Slide: Yejin Choi

ELIZA

```
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.
```

[Weizenbaum 1968]. Image: <https://thoughtcatalog.com/oliver-miller/2012/08/a-conversation-with-eliza/>

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

....



Xiaoice, developed by Microsoft

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

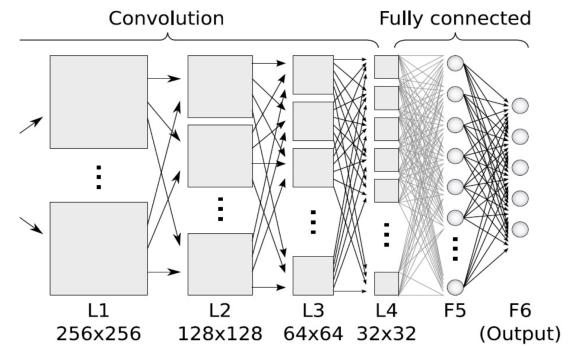


SoundingBoard Example

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.	Negotiation
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?	Thought
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.	Moving
6	Yes, it was hilarious.	I'm glad you feel this is hilarious. Which part of 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?	Fact
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.	

Source: Mari Ostendorf

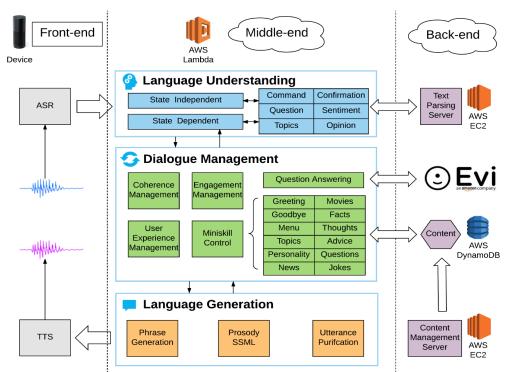
Sounding Board's Architecture?



Source: Yejin Choi



Sounding Board's Architecture



Source: Yejin Choi

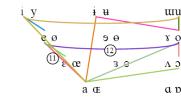
Related Areas



What is Nearby NLP?

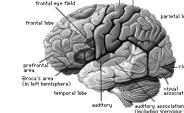
▪ Computational Linguistics

- 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!

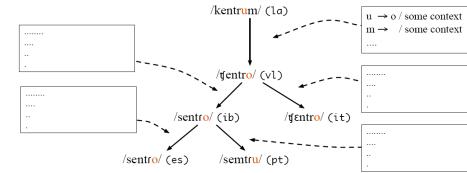


▪ Speech Processing

- Mapping audio signals to text
- Traditionally separate from NLP, converging



Example: 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

Why is Language Hard?



Ambiguities



Live TV

Stevie Wonder announces he'll be having kidney surgery during London concert

By Amit Verma, CNN

Updated 11:16 PM EDT, Sat July 06, 2013



(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?

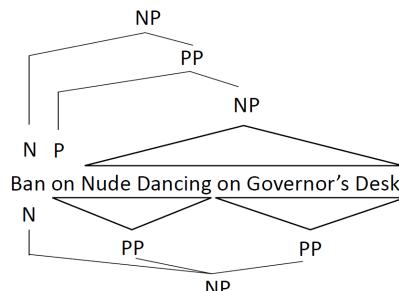
What Do We Need to Understand Language?



We Need Representation: Linguistic Structure

N N V N
N V ADJ N
Teacher Strikes Idle Kids

body/ body/
position position
Iraqi Head Seeks Arms

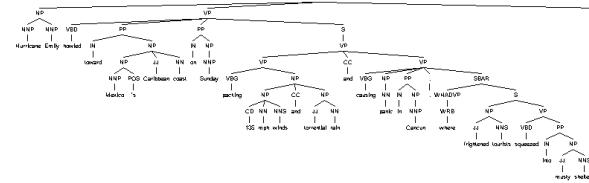


- Syntactic and semantic ambiguities: parsing needed to resolve these, but need context to figure out which parse is correct

Slide: Greg Durrett



Example: Syntactic Analysis

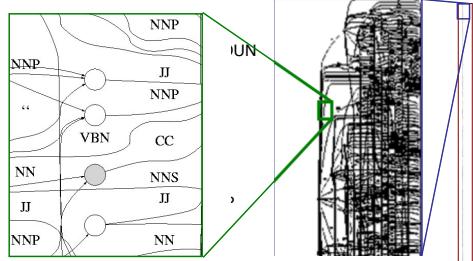


Hurricane Emily howled toward Mexico's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun, where frightened tourists squeezed into musty shelters .

Accuracy: 95+



We Need Data



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] [.]



We Need Models: Data Alone Isn't Enough!

CLASSIC SOUPS		Sm.	Lg.
清 嫩 雞 湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50 2.75
雞 飯 湯	58.	Chicken Rice Soup	1.85 3.25
雞 麵 湯	59.	Chicken Noodle Soup	1.85 3.25
廣 東 雪 菜 湯	60.	Cantonese Wonton Soup	1.50 2.75
蕃 茄 湯	61.	Tomato Clear Egg Drop Soup	1.65 2.95
雪 菜 湯	62.	Regular Wonton Soup	1.10 2.10
酸 辣 湯	63.	Hot & Sour Soup	1.10 2.10
蕃 芥 湯	64.	Egg Drop Soup	1.10 2.10
雪 菜 湯	65.	Egg Drop Wonton Mix	1.10 2.10
豆 腐 素 湯	66.	Tofu Vegetable Soup	NA 3.50
雞 玉 米 湯	67.	Chicken Corn Cream Soup	NA 3.50
蟹 肉 玉 米 湯	68.	Crab Meat Corn Cream Soup	NA 3.50
海 鮮 湯	69.	Seafood Soup	NA 3.50

Example from Adam Lopez



Learning Latent Syntax

Personal Pronouns (PRP)

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

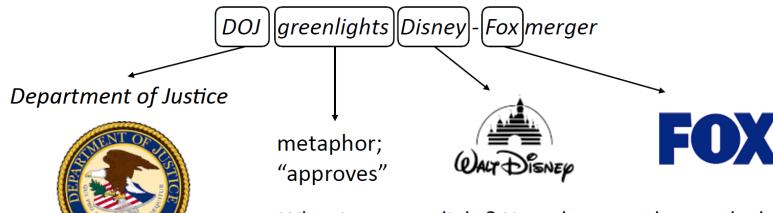
Proper Nouns (NNP)

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



We Need 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



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?

AI2 Allen Institute for AI

AllenNLP

- Annotate a sentence
 - Semantic Role Labeling
 - Named Entity Recognition
 - Constituency Parsing
 - Dependency Parsing
 - Open Information Extraction
 - Sentiment Analysis
- Annotate a passage
 - Coreference Resolution
- Answer a question
 - Reading Comprehension
- Commonsense

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:

Predictions:

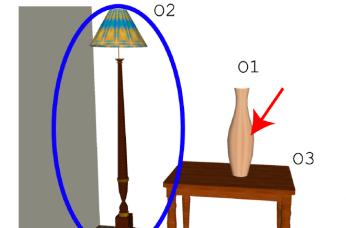
- 32.0% The
- 10.3% Ice
- 9.2% a
- 3.7% Water
- 2.7% When
- Undo



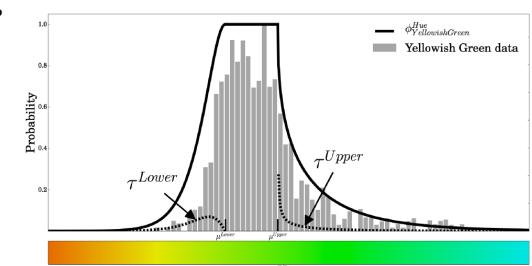
We Need Grounding

Grounding: linking linguistic concepts to non-linguistic ones

Question: What object is right of O2 ?



Golland et al. (2010)



McMahan and Stone (2015)

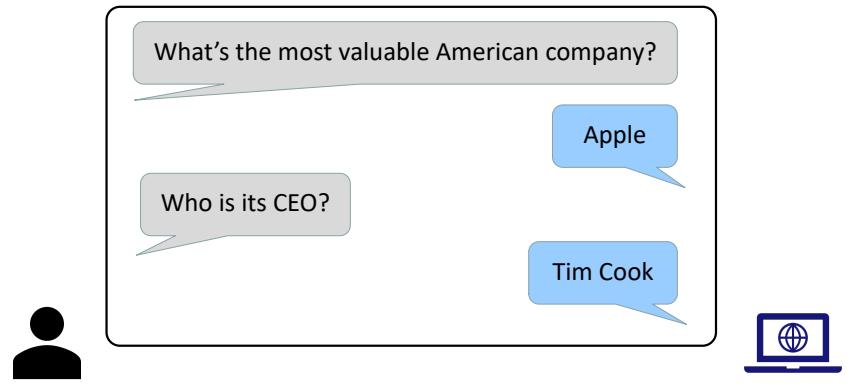
Slide: Greg Durrett



Example: Grounded Dialog



Example: Grounded Dialog



Why is Language Hard?

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

- ... and often we need all these things at the same time

What is this Class?



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!