Natural Language Processing

Katharina Kann — CSCI/LING5832

Watson Video

https://youtu.be/P18EdAKuC1U

Natural Language Processing

- **Goal**: Build machines that can perform useful and interesting tasks involving human language

- We will study the algorithms used to process language, the formal basis for those algorithms, and the relevant facts about human language that allow those algorithms to work.

Natural Language Processing

- **Why?**
  - Enormous amounts of text is available in machine readable form: newspapers, web pages, medical records, financial filings, product reviews, discussion forums, etc.

Natural Language Processing

- **Why?**
  - Enormous amounts of text is available in machine readable form: newspapers, web pages, medical records, financial filings, product reviews, discussion forums, etc.
  - Advances in speech recognition make conversational agents a key form of human—machine interaction

Natural Language Processing

- **Why?**
  - Enormous amounts of text is available in machine readable form: newspapers, web pages, medical records, financial filings, product reviews, discussion forums, etc.
  - Advances in speech recognition make conversational agents a key form of human—machine interaction
  - Much of human—human interaction is now mediated by computers via social media

Natural Language Processing

- **What do you need?**
  - No formal prerequisites… BUT:
  - Basic programming experience will be helpful.
  - Experience with calculus and probability theory will be helpful.
  - Knowledge of linguistics will be helpful.

Logistical Detail

- The slides are online on Canvas.
- You can follow along and make notes.

Today

- Introductions: Who are we?
- What is NLP, and what is this class about?
- How does this course work?

Introductions

Katharina Kann:

- Since January 2020: Assistant Professor of Computer Science at CU Boulder
- Before that:
  - Postdoc at New York University (ML^2 Group)
  - PhD Student at LMU Munich (CIS)
- Main research areas:
  - Natural language processing with neural networks, transfer learning, low-resource settings, morphology

Introductions

Beilei Xiang:

- Graduate student at CU Boulder
- GSS for this class

Introductions

You:

- 25 students signed up for in-person instruction
- 22 students signed up for the distance section

What is this class about?

Goals of this Course

- You should learn to do creative, thoughtful original engineering research on natural language processing, at a level that is worthy of publication in a top-tier NLP conference.
- This is hard. (But I won’t ask for much else…)
- It’s also different from editions of this class from 2019 and before.

NLP: Two Views

There are two possible motivations for research into AI, including NLP:

- Technological: Ultimate goal is *best performance* on some specific applied task.

- Cognitive: Ultimate goal is a single model with *human-like performance* across the board (in some domain), plus biological/cognitive plausibility.

Paraphrasing James Allen (1987)

NLP: Two Views

There are two possible motivations for research into AI, including NLP:

- Technological: Ultimate goal is *best performance* on some specific applied task. **Natural language processing**

- Cognitive: Ultimate goal is a single model with *human-like performance* across the board (in some domain), plus biological/cognitive plausibility. **Computational linguistics**

Paraphrasing James Allen (1987)

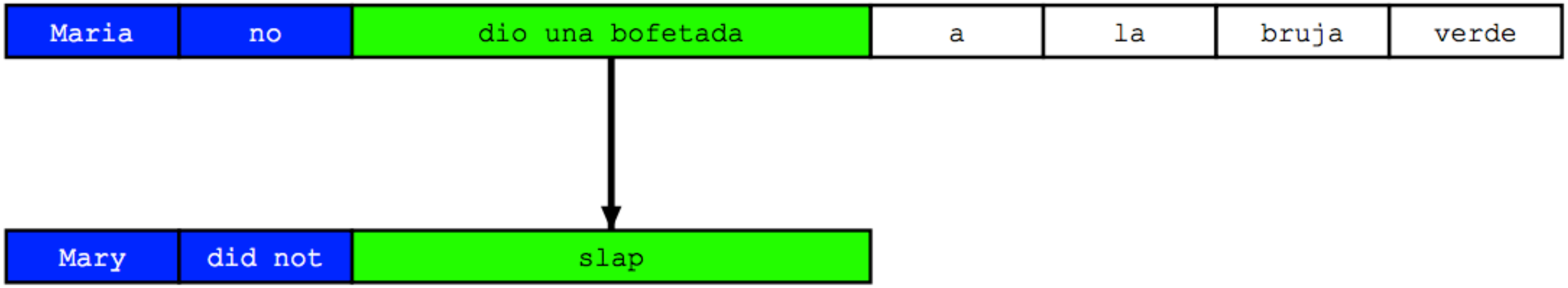
This Class

- The goal of the class is to teach you to do research on NLP.
    - For this, we first learn about the basics.
    - This is not primarily a linguistics class.
- Many of the most important methods in NLU draw on ideas from linguistics, especially computational semantics.
    - We will cover these ideas.

Some major NLP tasks: Applied

Machine Translation

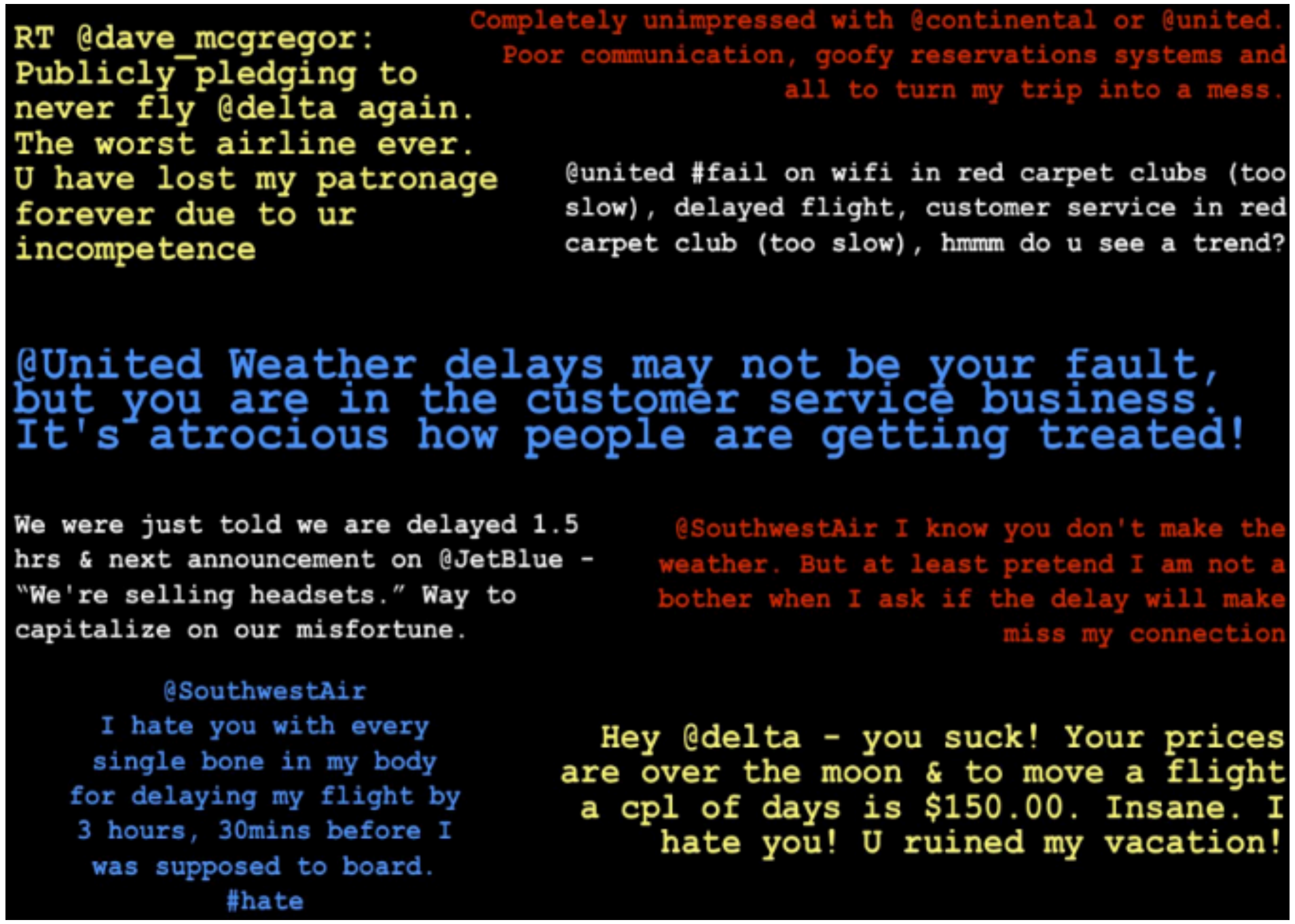
- **Input**: Sentence in a source language
- **Output**: Sentence in a target language



(Stephen Clark and Phillip Koehn)

Sentiment Analysis

- **Input:** Sentence or short document
- **Output**: Score (e.g., 1-10)



RT @dave_mcgregor: Publicly pledging to never fly @delta again. The worst airline ever. U have lost my patronage forever due to ur incompetence

Completely unimpressed with @continental or @united. Poor communication, goofy reservations systems and all to turn my trip into a mess.

@united #fail on wifi in red carpet clubs (too slow), delayed flight, customer service in red carpet club (too slow), hmmm do u see a trend?

@United Weather delays may not be your fault, but you are in the customer service business. It's atrocious how people are getting treated!

We were just told we are delayed 1.5 hrs & next announcement on @JetBlue - "We're selling headsets." Way to capitalize on our misfortune.

@SouthwestAir I know you don't make the weather. But at least pretend I am not a bother when I ask if the delay will make miss my connection

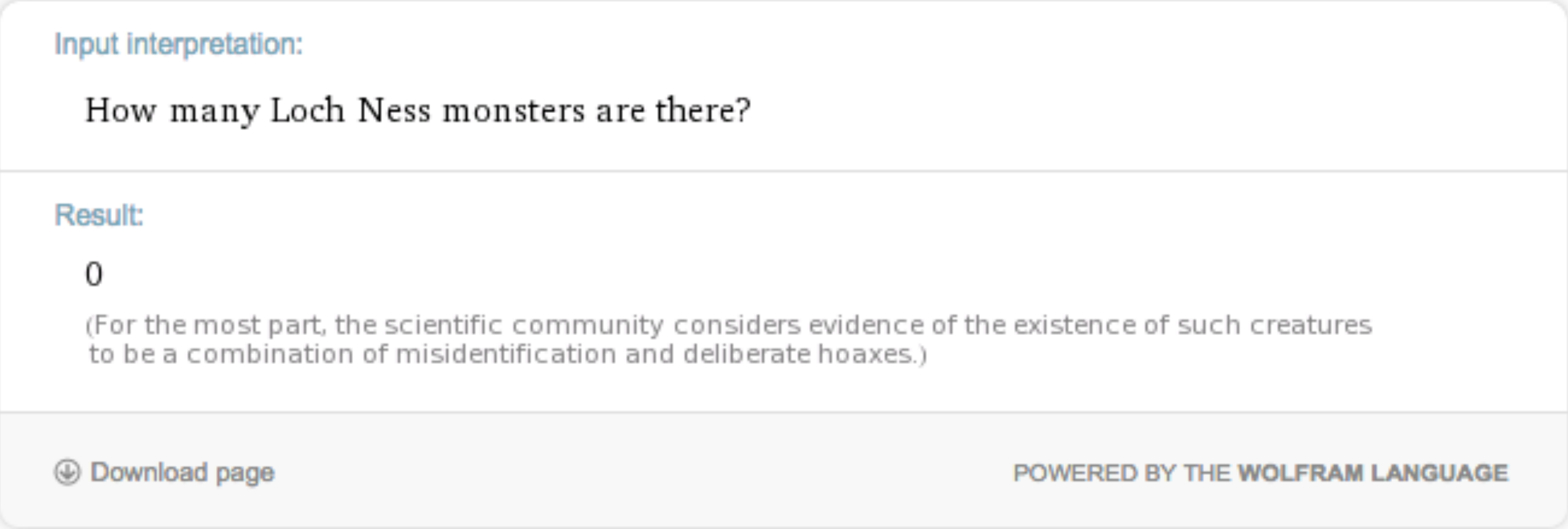
@SouthwestAir I hate you with every single bone in my body for delaying my flight by 3 hours, 30mins before I was supposed to board. #hate

Hey @delta - you suck! Your prices are over the moon & to move a flight a cpl of days is $150.00. Insane. I hate you! U ruined my vacation!

(from Stanford CS 224U)

Question Answering

- **Input:** Question
- **Output:** Short Answer



Input interpretation:

How many Loch Ness monsters are there?

Result:

0

(For the most part, the scientific community considers evidence of the existence of such creatures to be a combination of misidentification and deliberate hoaxes.)

Summarization

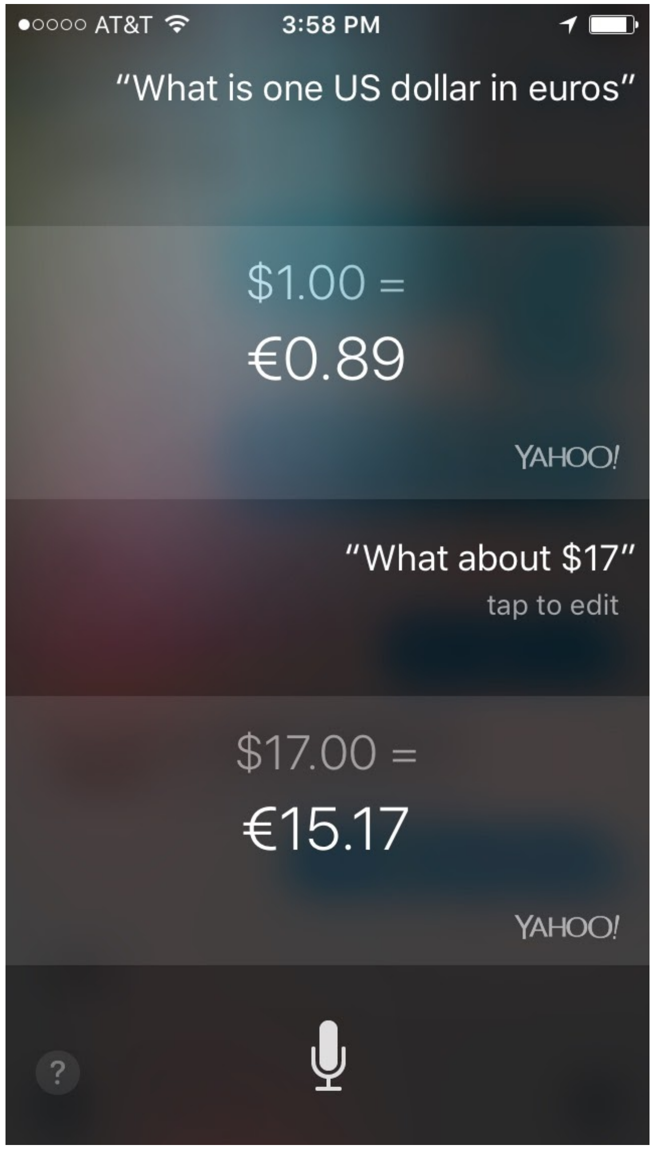
- **Input**: Long text
- **Output**: Short text

Many of us live by our to-do lists. That makes these scribbled or typed lists incredibly important — they dictate what gets done! So what's the most effective way to organize your priorities? I find that thinking in terms of a week — 168 hours — is better than day by day!

=> Planning by the week, rather than by the day, keeps you focused on important things.

Dialogue Systems

- **Input**: User utterance
- **Output**: System response



“What is one US dollar in euros”

$1.00 =
€0.89

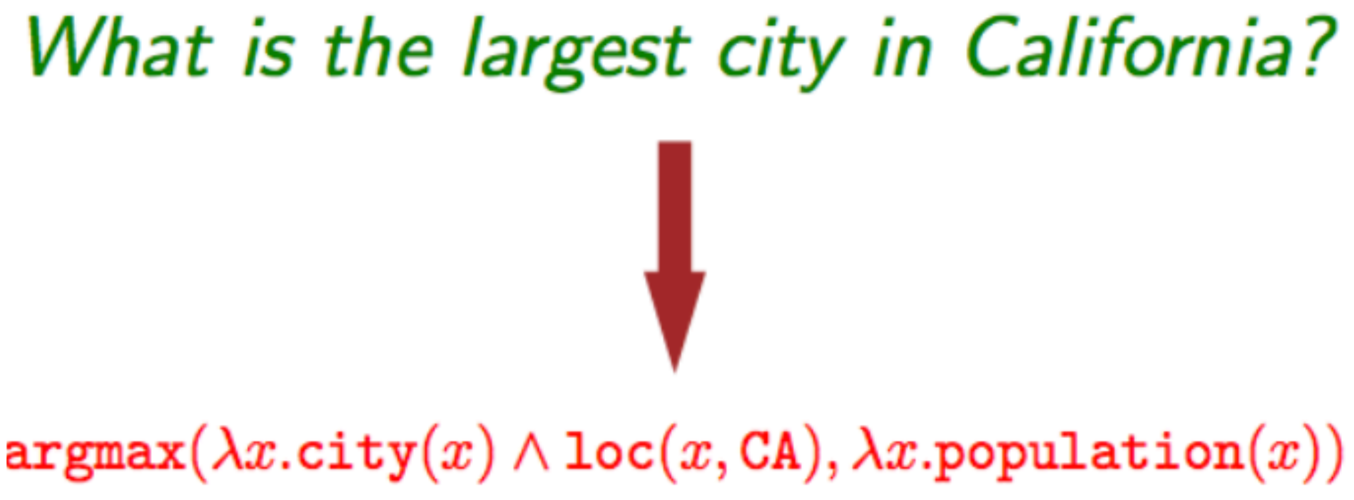
“What about $17”

$17.00 =
€15.17

Some major NLP tasks: Others

Semantic Parsing

- **Input**: Sentence in natural language
- **Output**: Logical expression



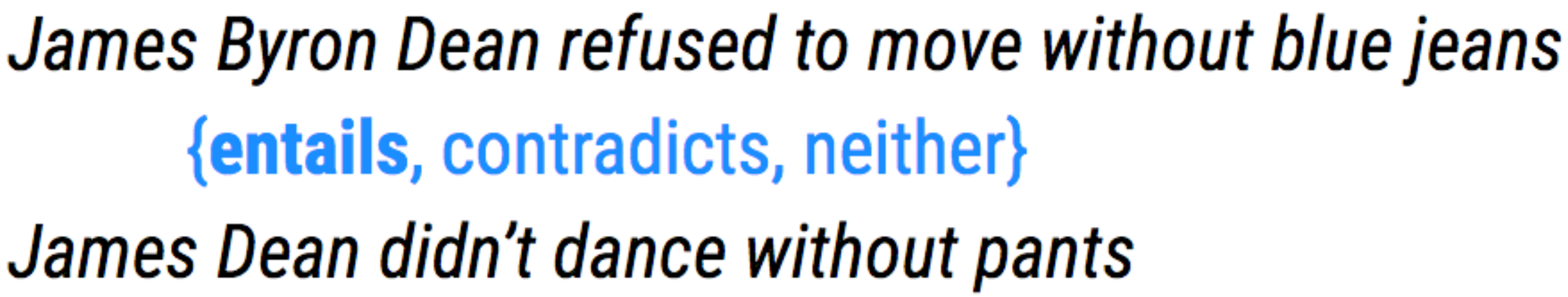
What is the largest city in California?

$\mathtt{argmax}(\lambda x.\mathtt{city}(x) \wedge \mathtt{loc}(x, \mathtt{CA}), \lambda x.\mathtt{population}(x))$

(Percy Liang)

Natural Language Inference

- **Input**: Two sentences
- **Output**: The relationship between their meanings (if any)



James Byron Dean refused to move without blue jeans

{**entails**, contradicts, neither}

James Dean didn't dance without pants

Paraphrase Identification

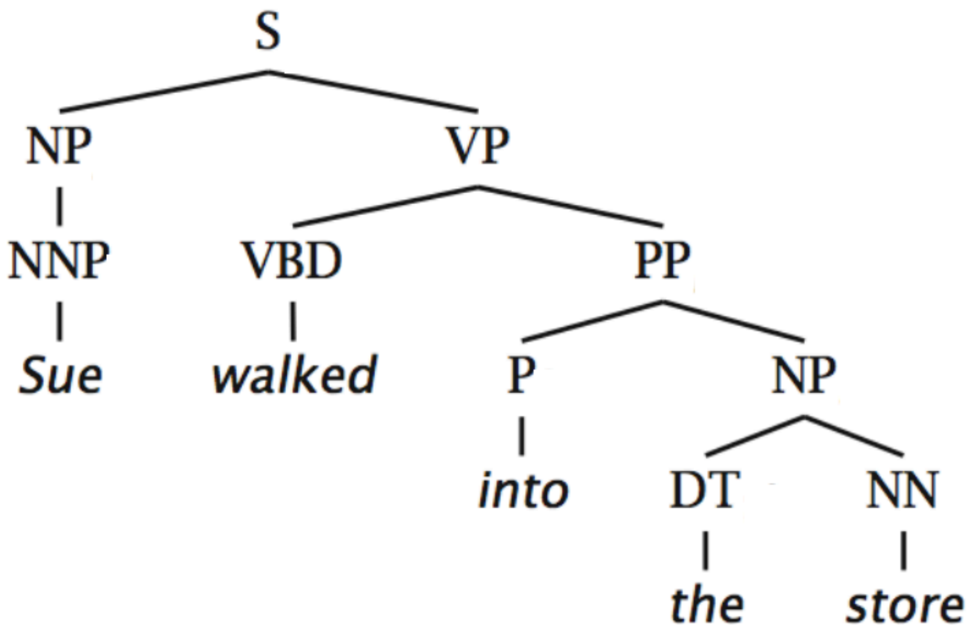
- **Input**: Two sentences
- **Output**: Binary decision (are these sentences paraphrases?)

1. The judge also refused to postpone the trial date of Sept. 29.

2. Obus also denied a defense motion to postpone the September trial date.

Syntactic Parsing and Part-of-Speech Tagging

- **Input**: Sentence
- **Output**: Tree structure, with optional annotations



(Dan Jurafsky and Chris Manning)

Tokenization

- **Input**: String
- **Output**: String with marked word boundaries

Computational linguistics

Semantics

Computational semantics applies methods (and theoretical frameworks) from computer science to answer scientific questions about natural language meaning:

- How do we represent the meanings of words and sentences?
- How do we derive the meaning of a sentence from the meanings of its words?
- How do we evaluate whether a sentence is true in a situation?
- How do we evaluate whether a sentence follows from another sentence?

Syntax

Computational syntax is concerned with questions about the structures of natural language sentences (i.e., grammar):

- What kinds of representations allow you to best describe what sentences are grammatical/possible in a language?
- What kinds of sentence structure are and are not possible in any given human languages?
- To what extent are these grammatical rules language-specific and learned, and to what extent do they represent something universal about human cognition?

Machine learning and NLP: Background and context

History of NLP in One Slide

- **1960s**: Pattern-matching with small rule-sets; huge (unrealistic) ambitions.
- **1970-80s**: Linguistically rich, logic-driven systems; labor-intensive successes on a few narrow tasks.
- **1990-2000s**: Statistical modeling revolution, machine learning becomes a central part of NLP, systems start to be deployed for practical tasks.
- **2010s**: Deep learning (artificial neural networks) takes off, accelerates progress on most tasks.
- **2018**: New surge of progress in semi-supervised learning with unlabeled data sparks progress on less data-rich tasks.

Machine Learning in this Class

- Machine learning (especially deep learning/artificial neural networks) represent a huge part of modern NLP research.

- This class does not cover any area of machine learning in depth.

- This class will briefly cover key points, and will refer to ideas in machine learning where appropriate.

Machine Learning in this Class

For the final project:

- If you have machine learning experience: Use it!
- If you want to learn by yourself: Do it!
- If you don’t know machine learning, you can still do an excellent project that doesn’t require the creation of new models.

Some issues to keep in mind

What Does Success Look Like?

Most NLU research takes a *behavioral* approach. A model is assumed to learn/understand language in some setting iff:

- It can perform the task that it was built to perform
- ...on data that wasn’t used to build or train it.
- (optional) ...and it makes errors and uses resources in a similar way to humans.
- Note: Ignoring internal representations here.

What Does Success Look Like?

The question of whether a computer is playing chess, or doing long division, or translating Chinese, is like the question of whether robots can murder or airplanes can fly — or people; after all, the "flight" of the Olympic long jump champion is only an order of magnitude short of that of the chicken champion (so I'm told). These are questions of decision, not fact; decision as to whether to adopt a certain metaphoric extension of common usage.

Chomsky (1996)

How Much Linguistic Knowledge Do NLP Systems Need?

Two questions:

- How much knowledge does a system need?
- If a system can learn from data, how much built-in knowledge does it need?

Sometimes it's more efficient to have your model learn language from data than trying to build in explicit knowledge.

Every time I fire a linguist, the performance of the speech recognizer goes up.

(probably apocryphal, attributed to Fred Jelinek, 1988)

Things to Know About NLP

- Natural language is:
  - Highly ambiguous at all levels
  - Complex and subtle in its use of context to convey meaning
  - Fuzzy, probabilistic
  - Involves reasoning about the world
  - A key part of people interacting with other people (deeply tied to social interaction)

Things to Know About NLP

- Natural language is:
  - Highly ambiguous at all levels
  - Complex and subtle in its use of context to convey meaning
  - Fuzzy, probabilistic
  - Involves reasoning about the world
  - A key part of people interacting with other people (deeply tied to social interaction)
- But NLP can also be surprisingly easy
  - Sometimes simple text features can do more than half the job

How does this class work?

Grading

- Homework: **30%**
- Final project: **70%**
- Piazza: **+3%**
- Participation will *not* be graded
- There will be *no* written exam

Grading

- Homework: **30%**
  - Homework assignments will make sure that everyone is following the lectures.
  - These assignments are *short* exercises: We expect these to take less than two hours each.
  - Deadlines: 10 AM on days marked in the syllabus

Grading

- Final project: **70%**
  - Requirement: Write an original research paper about a topic of your choice within NLP.
  - To earn an A, your idea, your execution, and your writing must be up to the same standard you'll see in the published papers you read.
  - Many projects yield negative results or unclear conclusions. This might make your work harder to publish, but it won't impact your grade.
  - It'll be short: 4 pages, in (dense) NLP conference paper format.

Grading

- Final project: 70%
  - **Proposal (10%)**: September 28
    - One page: What’s your idea, and how do you plan to pursue it?
  - **Proposal presentation (5%)**: October 07
    - Presentation and discussion of you idea with the class
  - **Partial draft (20%)**: November 23
    - 3 pages: Introduction, background, baseline results, and a literature review.
    - You must have made substantial progress by this point.
  - **In-class presentation (5%)**: December 02 and 07
  - **Final paper (30%)**: December 07
    - 4 pages

Reading

- **Assigned readings**:
  - There will be assigned readings (papers or book chapters).
  - You won’t be tested on these, but it will be hard to keep in with the class if you don’t read them.
  - Useful practice for the *main* readings for the course:

Reading

- **Assigned readings**:
  - There will be assigned readings (papers or book chapters).
  - You won’t be tested on these, but it will be hard to keep in with the class if you don’t read them.
  - Useful practice for the *main* readings for the course:
- **Project background readings**:
  - Your should build on recent, important prior work in NLP, and it should be novel (i.e., no one has done it before).
  - To do this, you’ll need to do a lot of background reading!
  - Start as soon as you have formed groups and decided on a research topic.
  - More guidance in a few weeks.

Collaboration and Teams

- Projects must be done in teams of 4 (or as close to that number as possible).

- Every team should ideally have at least one linguist and one computer scientist.

- You should discuss the homeworks with your teammates (or other classmates), but you must write up and submit your work on your own.

In-Class Exercises

- Some lectures will contain small exercises to deepen understanding.
- Bring your laptops and a pen and paper to class!

Course Sites

- **Piazza:**
    - https://piazza.com/colorado/fall2020/csciling5832
    - Questions and discussions
    - Team matching

Course Sites

- **Piazza:**
    - https://piazza.com/colorado/fall2020/csciling5832
    - Questions and discussions
    - Team matching
- **Canvas:** canvas.colorado.edu
    - Assignment submission
    - Gradebook
    - Slides
    - Recordings of lectures

Additional Policies

- Turn things in on time! Anything turned in late will not receive points.
- Combining projects across classes is *great*!
    - ...but you have to contact me (and your other instructor) by the draft deadline to confirm that the project is big enough for both classes.
- Plagiarism will not be tolerated.
- When in doubt, read the syllabus or ask (at office hours or on Piazza).

What’s next?

What’s Next?

- Wednesday: Lexical semantics (first real class!)
- Sign up on Piazza
- Find yourself a group
- Decide on a research project

Project Tips

- Think about why you enrolled in this class, and what kinds of skills you want to develop.
    - Do you want to technological research or cognitive science?
    - Are there specific methods you want to learn?
- Skim through the textbooks and class topics to see what we will cover.
- Once you have some rough ideas:
    - Search for recent work on Google Scholar to see what has been done.
    - Even if you don't understand the papers you find, look for examples of what computational models can and can’t do.

Questions?