# CS378: Natural Language Processing

## Lecture 1: Introduction

**Greg Durrett**

## Administrivia

- **Piazza:** link on the course website
- **My office hours:** Tuesday 1pm-2pm, Wednesday 10am-11am, GDC 3.812
- **TA:** Yasumasa Onoe; **Proctor:** Shrey Desai. See website for OHs

## 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)

## Enrollment

- We’ve already expanded the cap by 25 students by moving to this room, more is not possible
- **Assignment 0** is out now (due Friday):
  - Please look at the assignment well before then
  - 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)
  - If you get in and didn’t do the assignment because you weren’t registered, you will be able to submit it late
- If you are past 20 on the waitlist, you have a low chance of getting into the class, but we have to see how it progresses
What’s the goal of NLP?

- Be able to solve problems that require deep understanding of text
- Example: dialogue systems

Automatic Summarization

Google Critic Ousted From Think Tank Funded by the Tech Giant
WASHINGTON — In the hours after European antitrust regulators levied a record $2.7 billion fine against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

One of New America’s writers posted a statement critical of Google. Eric Schmidt, Google’s CEO, was displeased.

The writer and his team were dismissed.

Machine Translation

Translate

Trump Pope family watch a hundred years a year in the White House balcony

She emphasized that it is necessary to thoroughly implement the spirit of General Secretary Xi Jinping’s important instructions, implement the instructions of Premier Li Keqiang and the implementation of the executive meeting of the State Council, consolidate territorial control responsibilities, strengthen the implementation of prevention and control measures, and effectively protect the people’s health and maintain normal production and living order.
NLP Analysis Pipeline

- Text Analysis
  - Syntactic parses
  - Coreference resolution
  - Entity disambiguation
  - Discourse analysis

- Annotations
  - Summarize
  - Extract information
  - Answer questions
  - Identify sentiment
  - Translate

- Applications
  - Summarize
  - Extract information
  - Answer questions
  - Identify sentiment
  - Translate

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

How do we represent language?

- Labels
  - the movie was good
  - Beyoncé had one of the best videos of all time

- Sequences/tags
  - Tom Cruise stars in the new Mission Impossible film

- Trees
  - I eat cake with icing
  - λx. flight(x) ∧ dest(x)=Miami

How do we use these representations?

- Text Analysis
  - Labels
  - Sequences
  - Trees

- Applications
  - Learn tree-to-tree machine translation models
  - Build tree-structured neural networks over text

Why is language hard?
(And how can we handle that?)

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

- Lots of data!

| 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] [permisstrait] [licences] [to] [term] [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] [ ] |

What do we need to understand language?

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

Department of Justice

metaphor; “approves”

- What is a green light? How do we understand what “green lighting” does?

What do we need to understand language?

- Grounding: learn what fundamental concepts actually mean in a data-driven way

Grounding: What object is right of O2?

- Linguisic structure

...but computers probably won't understand language the same way humans do

However, linguistics tells us what phenomena we need to be able to deal with and gives us hints about how language works:

a. John has been having a lot of trouble arranging his vacation.

b. He cannot find anyone to take over his responsibilities. (he = John)

C_p = John; C_I = (John)

c. He called up Mike yesterday to work out a plan. (he = John)

C_p = John; C_I = (John, Mike) (CONTINUE)

d. Mike has annoyed him a lot recently.

C_p = John; C_I = (Mike, John) (RETAI)

e. He called John at 5 AM on Friday last week. (he = Mike)

C_p = Mike; C_I = (Mike, John) (SHIFT)
What techniques do we use? 
(to combine data, knowledge, linguistics, etc.)

Where are we?

- NLP consists of: analyzing and building representations for text, solving problems involving text
- These problems are hard because language is ambiguous, requires drawing on data, knowledge, and linguistics to solve
- Knowing which techniques use requires understanding dataset size, problem complexity, and a lot of tricks!
- NLP encompasses all of these things

---

NLP vs. Computational Linguistics

- NLP: build systems that deal with language data
- CL: use computational tools to study language

---

Hamilton et al. (2016)
Outline of the Course

- Classification: linear and neural, word representations (3.5 weeks)
- Text analysis: tagging and parsing (3 weeks)
- Generation, applications: language modeling, machine translation (3 weeks)
- Question answering, pre-training (2 weeks)
- Applications and miscellaneous (3.5 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 50% of grade (A0, A5: 5%; all others 10%)
- Mix of writing and implementation;
- Assignment 0 is out NOW, due Friday (extensions granted if you get in the class late)
- ~2 weeks per assignment after Assignment 0
- 2 “slip days” throughout the semester to turn in assignments 24 hours late. Otherwise, you lose 15% credit per day the assignment is late

These assignments require understanding of the concepts, ability to write performant code, and ability to think about how to debug complex systems. They are challenging, so start early! The course staff are not here to debug your code! We will help you understand the concepts from lecture and come up with debugging strategies, but we won’t read through your code to spot your bug.

Coursework

- Midterm (25% of grade), in class
- Similar to written homework problems
- Final project (25% of grade)
  - Groups of 2 preferred, 1 is possible
  - Standard project: neural network models for question answering
  - Independent projects are possible: these must be proposed by March 24 (to get you thinking early) and will be held to a high standard!

Academic Honesty

- Assignments and exams are to be completed independently (except for the group final project)
- Don’t share code with others — we will be running Moss
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. If you notice an incident that causes concern, please contact the Campus Climate Response Team: diversity.utexas.edu/ccrt

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

### Conduct

YOU BELONG HERE

### Survey

1. Your name
2. Fill in: I am a [CS / ____] undergrad in year [1 2 3 4 5+]
3. Which of the following have you taken?
   1. CS 342/343/363
   2. Another class which taught classification
   3. A class which taught SVD
4. One reason you want to take this class or one thing you want to get out of it
5. One interesting fact about yourself, or what you like to do in your spare time