CS378: Natural Language Processing
Lecture 1: Introduction

Greg Durrett

Administivia

- Lecture: Tuesdays and Thursdays 9:30am - 10:45am
- Piazza: link on the course website
- My office hours: Tuesday 1pm-2pm (starting next week), Wednesday 11am-noon (starting tomorrow), GDC 3.420
- TA: Jiacheng Xu; Office hours: Monday + Wednesday, 1pm-2pm GDC 1.302
- TA: Shivangi Mahto; Office hours: Thursday, 2pm-3pm GDC 1.302

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

- I want everyone to be able to take this class!
- 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 make it up
  - 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

- Siri, what’s the most valuable American company?
- Apple

Who is its CEO?
- Tim Cook

Machine Translation

Translate

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

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

NLP Analysis Pipeline

- NLP is about building these pieces! (largely using statistical approaches)
How do we represent language?

**Text**

- Labels
  - the movie was good
  - Beyoncé had one of the best videos of all time
  - subjective
- Sequences/tags
  - Tom Cruise stars in the new Mission Impossible film
- Trees
  - \( \lambda x. flight(x) \land dest(x) = \text{Miami} \)

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

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

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

What do we need to understand language?

‣ Lots of data!

<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] [solutions] [transitoire] [qui] [permettrait] [l] [conduire] à [terme] à [une] charte à [valeur] contraignante [']</td>
</tr>
<tr>
<td>10x DATA</td>
<td>[this] [would] [be] [a] [transitional] [solution] [which] [would] [lead] to [a] [binding] [charter] [legally] [binding] [']</td>
</tr>
<tr>
<td>100x DATA</td>
<td>[this] [would] be [a] [transitional] [solution] [which] [would] [lead] to [a] [binding] [charter] [']</td>
</tr>
<tr>
<td>1000x DATA</td>
<td>[this] [would] be [a] [transitional] [solution] [which] [would] [eventually] [lead] to [a] [binding] [charter] [']</td>
</tr>
</tbody>
</table>
What do we need to understand language?

- World knowledge: have access to information beyond the training data
  - What is a green light? How do we understand what "green lighting" does?

- Grounding: learn what fundamental concepts actually mean in a data-driven way
  - Question: What object is right of [O2]?

What techniques do we use?
(to combine data, knowledge, linguistics, etc.)

- Linguistic 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
  - John has been having a lot of trouble arranging his vacation.
  - He cannot find anyone to take over his responsibilities. (he = John)
  - He called up Mike yesterday to work out a plan. (he = John)
  - Mike has annoyed him a lot recently.
  - He called John at 5 AM on Friday last week. (he = Mike)
  - Centering Theory
  - Grosz et al. (1995)
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.

Computational tools for other purposes: literary theory, political science...

Outline of the Course

- Classification: conventional and neural, word representations (3 weeks)
- Text analysis: tagging, parsing, information extraction (3.5 weeks)
- Generation, applications: language modeling, machine translation, dialogue (4 weeks)
- Other applications: question answering, TBD (3 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 2018?
Coursework

- Five assignments, worth 45% of grade
- Mix of writing and implementation;
- Assignment 0 is out NOW, due Friday
- ~2 weeks per assignment after Assignment 0

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!

Coursework

- Midterm (25% of grade), in class
- Similar to written homework problems
- Final project (30% of grade)
  - Groups of 2 preferred, 1 is possible
  - (Brief!) proposal to be approved by course staff
  - Open-ended *or* there will be a few more structured options (around translation and dialogue)

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

Conduct

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
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. Which of the following have you used?
   1. Python
   2. numpy/scipy/scikit-learn
   3. Tensorflow/PyTorch
5. One interesting fact about yourself, or what you like to do in your spare time