CSE 5525: Foundations of Speech and Language Processing

Lecture 1: Introduction

Huan Sun (CSE@OSU)

Many thanks to Prof. Greg Durrett @ UT Austin for sharing his slides.
Logistics

- Lectures (online): Wednesdays and Fridays 12:45pm - 2:05pm

- Course website:  
  http://web.cse.ohio-state.edu/~sun.397/courses/au2020/cse5525.html

- Piazza: Signup link is on Carmen

- My office hours: Wednesdays 4:30 - 5:30PM & additional hours by appointment

- TA: Dingkang Wang; Office hours: Fridays 4:30 - 5:30PM

All Zoom links for lectures and office hours can be found on Carmen.
Course Requirements

- Prereq: (CSE 3521 or CSE 5521) and (CSE 5522 or Stat 3460 or Stat 3470). Not open to students with credit for CSE 733 (Prior course number).

- Python experience

- Additional prior exposure to probability, linear algebra, optimization, linguistics, and NLP useful but not required

- Textbooks:
  - Speech and Language Processing (3rd Edition), Dan Jurafsky and James H. Martin
  - Natural Language Processing, Jacob Eisenstein
  - Both are available online and linked on our website too!
Grading Plan

- Participation (10%)

- Homework (50%)
  
  HW #1 is out! If you find it challenging, please email or talk to me (this is smaller-scale and relatively easier, compared with other HWs/Project or midterm exam)

- Midterm Exam (20%): On OCT 14th, 2020

- Final Project (20%)
  
  Form teams (2-3 people with Diverse* Background) from now!

  *: e.g., A group of mixed undergrads and grads (e.g., 1 grad + 1 undergrads, or, 2 grads + 1 undergrad) are expected! Exceptions can be made, i.e., 2 grads with distinct research backgrounds are also OK (but talk to the instructor first); 2 grads with both AI/NLP majors are not OK.
Assignments

- 3 Assignments
  - Implementation-oriented, with one or more open-ended components to each
  - HW #1 (classification) is out NOW
  - No late assignments

These projects 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!
Assignments

- Final project (20%)
  - Form a group of 2-3 people with **diverse** background
    - e.g., 1 grad + 1-2 undergrads (start forming teams from now)
  - (Brief! 1-page) proposal to be approved by the instructor by the midpoint of the semester (10/21/2020)
  - Start thinking about topics to work on from now
- Due at the end of the semester:
  - A roughly 4-page report (**Written in the style and tone of an ACL paper; more detailed requirements to come**)
  - Final project presentation (~ 5 min for each group depending on how many groups)
What’s the goal of NLP?

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

Siri, what’s the most valuable American company?

- Apple
- Who is its CEO?
- Tim Cook

Recognize `marketCap` is the target value

Do computation

Recognize predicate

Resolve references
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.

Ms. Slaughter told Mr. Lynn that “the time has come for Open Markets and New America to part ways,” according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be exiled from New America.
Trump Pope family watch a hundred years a year in the White House balcony
NLP Analysis Pipeline

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

- Annotations

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

- NLP is about building these pieces!

- All of these components can be modeled with statistical approaches trained using machine learning
How do we represent language?

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

**Sequences/tags**
- PERSON
  - Tom Cruise
- WORK_OF_ART
  - Mission Impossible _film_

**Trees**
- \(\lambda x. \text{flight}(x) \land \text{dest}(x)=\text{Miami}\)
- flights to Miami
- \(I\) eat cake with icing
Main questions: What representations do we need for language? What do we want to know about it?

Boils down to: what ambiguities do we need to resolve?
Why is language hard?
(and how can we handle that?)
Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

The city council refused the demonstrators a permit because they advocated violence

The city council refused the demonstrators a permit because they feared violence

The city council refused the demonstrators a permit because they ______ violence

>5 datasets in the last two years examining this problem and commonsense reasoning

Referential ambiguity
Language is Ambiguous!

Syntactic and semantic ambiguities: parsing needed to resolve these, but need context to figure out which parse is correct.
Language is **Really** Ambiguous!

- There aren’t just one or two possibilities which are resolved pragmatically

  - *il fait vraiment beau*

    - It is really nice out
    - It’s really nice
    - The weather is beautiful
    - It is really beautiful outside
    - He makes truly beautiful
    - It fact actually handsome

- Combinatorially many possibilities, many you won’t even register as ambiguities, but systems still have to resolve them
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] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] .</td>
</tr>
<tr>
<td>10x DATA</td>
<td>[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] .</td>
</tr>
<tr>
<td>100x DATA</td>
<td>[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] .</td>
</tr>
<tr>
<td>1000x DATA</td>
<td>[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] .</td>
</tr>
</tbody>
</table>

slide credit: Dan Klein
What do we need to understand language?

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

- DOJ greenlights Disney - Fox merger

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

- Need commonsense knowledge
What do we need to understand language?

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

**Question:** What object is right of 02?

- Golland et al. (2010)

- McMahan and Stone (2015)
What do we need to understand language?

- 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

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

b. He cannot find anyone to take over his responsibilities. (he = John)
   \[ C_b = \text{John}; C_f = \{\text{John}\} \]

c. He called up Mike yesterday to work out a plan. (he = John)
   \[ C_b = \text{John}; C_f = \{\text{John, Mike}\} \text{ (CONTINUE)} \]

d. Mike has annoyed him a lot recently.
   \[ C_b = \text{John}; C_f = \{\text{Mike, John}\} \text{ (RETAIN)} \]

e. He called John at 5 AM on Friday last week. (he = Mike)
   \[ C_b = \text{Mike}; C_f = \{\text{Mike, John}\} \text{ (SHIFT)} \]

Centring Theory
Grosz et al. (1995)
What techniques do we use?
(to combine data, knowledge, linguistics, etc.)
A brief history of (modern) NLP

Largely rule-based, expert systems

Penn treebank

Collins vs. Charniak parsers

Unsup: topic models, grammar induction

earliest stat MT work at IBM

Ratnaparkhi tagger

Sup: SVMs, CRFs, NER, Sentiment

Semi-sup, structured prediction

Pretraining


Largely rule-based, expert systems

Penn treebank

Collins vs. Charniak parsers

Unsup: topic models, grammar induction

earliest stat MT work at IBM

Ratnaparkhi tagger

Sup: SVMs, CRFs, NER, Sentiment

Semi-sup, structured prediction

Pretraining
Less Manual Structure

- Training is supervised but models still rely less on manual structure

Klein and Manning (2003)
Manually-constructed grammars

Petrov et al. (2006)
Induced grammars

Hall, Durrett, Klein (2014)
Basic grammar + features
“The yield on the benchmark issue rose to 10% from 5%”

- No grammars at all!
Interpretability

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

- Hard to analyze why these errors happen in neural models (but people are trying)
- Models with more manual structure might be more interpretable
Pretraining

- Language modeling: predict the next word in a text $P(w_i | w_1, \ldots, w_{i-1})$
  
  $P(w | \text{I want to go to}) = 0.01$ Hawai’i
  
  $0.005$ LA
  
  $0.0001$ class

- Model understands some sentiment?

  $P(w | \text{the acting was horrible, I think the movie was}) = 0.1$ bad
  
  $0.001$ good

- Train a neural network to do language modeling on massive unlabeled text, fine-tune it to do {tagging, sentiment, question answering, ...}

  Peters et al. (2018), Devlin et al. (2019)
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 to 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
NLP vs. Computational Linguistics

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

Bamman, O'Connor, Smith (2013)
Course Goals

- Cover fundamental machine learning 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 the last 3 years?
- Make you a “producer” rather than a “consumer” of NLP tools
  - The 3 assignments should teach you what you need to know to understand nearly any system in the literature
Preview what to come

- 3 Assignments
  - HW #1 (classification) is out NOW and due on 09/09/2020! Start early!
  - HW #2 and #3 would be like implementing Conditional Random Fields and Neural Networks

Similar courses and HWs/Final Project Requirements can be found here: https://www.cs.utexas.edu/~gdurrett/courses/fa2019/cs388.shtml

These projects 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!