CS388: Natural Language Processing
Lecture 1: Introduction

Greg Durrett
Lecture: Tuesdays and Thursdays 9:30am - 10:45am

Course website:

Gradescope: you should’ve gotten an email

Piazza: link on the course website

My office hours: Office hours: Tuesday 1pm-2pm, Wednesday 3:30pm-4:30pm

Note: my OHs today are 12:30pm-1:30pm

TA: Xi Ye. See course website for OHs
Course Requirements

- 391L Machine Learning (or equivalent)
- 311 or 311H Discrete Math for Computer Science (or equivalent)
- Python experience
- Additional prior exposure to probability, linear algebra, optimization, linguistics, and NLP useful but not required
- Mini1 is out now (due January 28), please look at it soon
  - If this seems like it’ll be challenging for you, come and talk to me (this is smaller-scale than the projects, which are smaller-scale than the final project)
What’s the goal of NLP?

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

Siri, what’s your favorite kind of movie?

I like superhero movies!

What’s come out recently?

The Avengers
Question Answering

When was Abraham Lincoln born?

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lincoln, Abraham</td>
<td>2/12/1809</td>
</tr>
<tr>
<td>Washington, George</td>
<td>2/22/1732</td>
</tr>
<tr>
<td>Adams, John</td>
<td>10/30/1735</td>
</tr>
</tbody>
</table>

February 12, 1809

How many visitor centers are there in Rocky Mountain National Park?

The park has a total of five visitor centers

five
中共中央政治局于7月30日召开会议，会议分析研究当前经济形势，部署下半年经济工作。

The Political Bureau of the CPC Central Committee held a meeting on July 30 to analyze and study the current economic situation and plan economic work in the second half of the year.

People’s Daily, August 10, 2020
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.

What does this mean? What is the context? How can we summarize this?
NLP Analysis Pipeline

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

- NLP is about building these pieces!
- All of these components are modeled with statistical approaches trained with machine learning

Applications
- Summarize
- Extract information
- Answer questions
- Identify sentiment
- Translate
How do we represent language?

Text:
- the movie was good
- Beyoncé had one of the best videos of all time
- Tom Cruise stars in the new Mission Impossible film
- I eat cake with icing
- flights to Miami

Labels:
- the movie was good
- Beyoncé had one of the best videos of all time
- Tom Cruise stars in the new Mission Impossible film
- I eat cake with icing
- flights to Miami

Sequences/tags:
- PERSON: Tom Cruise
- WORK_OF_ART: Mission Impossible

Trees:
- S
  - VP
    - VBZ: eat
    - NP: cake
    - PP: with icing
- lambda: x. flight(x) \land dest(x)=Miami
Main question: 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

  example credit: Dan Klein
Language is **Really** Ambiguous!

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

Department of Justice

**DOJ** greenlights Disney-Fox merger

metaphor; “approves”

- 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 $O_2$?

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)
      \( C_b = \text{John}; C_f = \{\text{John, Mike}\} \) (CONTINUE)

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

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

   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
  - S
  - NP
  - VP
- Collins vs. Charniak parsers
- Ratnaparkhi's tagger
  - NNP
  - VBZ
- Unsup: topic models, grammar induction
- Sup: SVMs, CRFs, NER, Sentiment
- Semi-sup, structured prediction
- Pretraining

- 1980: Earliest statistical MT work at IBM
- 1990: Ratnaparkhi's tagger
- 2000: Neural
- 2010: Semi-sup, structured prediction
- 2020: Pretraining
Supervised vs. Unsupervised

- Supervised techniques work well on very little data (even neural networks)
  - Annotation (two hours!)
  - Unsupervised learning
  - Better system!

- Fully unsupervised techniques have fallen out of favor

“Learning a Part-of-Speech Tagger from Two Hours of Annotation”
Garrette and Baldridge (2013)
Pretraining

- Language modeling: predict the next word in a text \( P(w_i | w_1, \ldots, w_{i-1}) \)

\[
P(w | \text{I want to go to}) = 0.01 \text{ Hawai'i}
\]
\[
0.005 \text{ LA}
\]
\[
0.0001 \text{ class}
\]

- Model understands some sentiment?

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

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

Peters et al. (2018), Devlin et al. (2019)
Interpretability

- When we have complex models, how do we understand their decisions?

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

  The movie is mediocre, maybe even bad.  

Wallace, Gardner, Singh  
Interpretability Tutorial at EMNLP 2020
Interpretability

- When we have complex models, how do we understand their decisions?

- “Attribution”: understand what parts of the input contribute to a prediction

  - Why was it class A instead of class B?

  - What is the “counterfactual” scenario we are considering (the foil)?
    
    - I drank tea because I don’t like coffee
    - I drank tea because I was thirsty (Jacovi and Goldberg, 2020)

- Dataset biases: does our data have flaws that prevent the model from doing the right thing?

- Probing: what representations get learned in deep models?
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)
NLP vs. Computational Linguistics

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

Bamman, O’Connor, Smith (2013)
### Outline

**ML and structured prediction for NLP**

<table>
<thead>
<tr>
<th>Date</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 19</td>
<td>Introduction</td>
</tr>
<tr>
<td>Jan 21</td>
<td>Binary Classification</td>
</tr>
<tr>
<td>Jan 26</td>
<td>Multiclass Classification</td>
</tr>
<tr>
<td>Jan 28</td>
<td>Sequence Models 1: HMMs</td>
</tr>
<tr>
<td>Feb 2</td>
<td>Sequence Models 2: CRFs</td>
</tr>
<tr>
<td>Feb 4</td>
<td>Neural 1: Feedforward</td>
</tr>
<tr>
<td>Feb 9</td>
<td>Neural 2: Word Embeddings; Bias</td>
</tr>
<tr>
<td>Feb 11</td>
<td>Neural 3: RNNs</td>
</tr>
<tr>
<td>Feb 16</td>
<td>Neural 4: Language Modeling, ELMo</td>
</tr>
<tr>
<td>Feb 18</td>
<td>Neural 5: Interpreting NNs</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
</tr>
<tr>
<td>----------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>Feb 23</td>
<td>Trees 1: Constituency, PCFGs</td>
</tr>
<tr>
<td>Feb 25</td>
<td>Trees 2: Better grammars, Dependency</td>
</tr>
<tr>
<td>Mar 2</td>
<td>Trees 3: Shift-reduce, State-of-the-art parsers</td>
</tr>
<tr>
<td>Mar 4</td>
<td>Semantics 1</td>
</tr>
<tr>
<td>Mar 9</td>
<td>Semantics 2 / Seq2seq 1</td>
</tr>
<tr>
<td>Mar 11</td>
<td>Seq2seq 2: Attention</td>
</tr>
<tr>
<td>Mar 16</td>
<td>NO CLASS</td>
</tr>
<tr>
<td>Mar 18</td>
<td>NO CLASS</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
</tr>
<tr>
<td>----------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Mar 23</td>
<td>Seq2seq 3: Degeneration / Annotation, Dataset Bias</td>
</tr>
<tr>
<td>Mar 25</td>
<td>MT 1: Phrase-based</td>
</tr>
<tr>
<td>Mar 30</td>
<td>MT 2: Neural, Transformers</td>
</tr>
<tr>
<td>April 1</td>
<td>Pre-training 1: BERT, GPT</td>
</tr>
<tr>
<td>April 6</td>
<td>Pre-training 2: BART/T5 and beyond</td>
</tr>
<tr>
<td>April 8</td>
<td>Generation 1: Dialogue, Ethics</td>
</tr>
<tr>
<td>April 13</td>
<td>Generation 2: Summarization</td>
</tr>
<tr>
<td>April 15</td>
<td>QA 1: Reading comprehension</td>
</tr>
<tr>
<td>April 20</td>
<td>QA 2: Multi-hop, etc.</td>
</tr>
<tr>
<td>April 22</td>
<td>Guest Lecture: Jason Baldridge (Google)</td>
</tr>
<tr>
<td>April 27</td>
<td>Multilingual / Cross-lingual models</td>
</tr>
<tr>
<td>April 29</td>
<td>Wrapup + Ethics</td>
</tr>
<tr>
<td>May 4</td>
<td>FP presentations 1</td>
</tr>
<tr>
<td>May 6</td>
<td>FP presentations 2</td>
</tr>
</tbody>
</table>
E.g., “toxic degeneration”: systems can generate {racist, sexist, …} content

We will touch on ethical issues throughout the course
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 2021?
- Make you a “producer” rather than a “consumer” of NLP tools
  - The four assignments should teach you what you need to know to understand nearly any system in the literature (e.g.: state-of-the-art NER system = project 1 + mini 2 + BERT, basic MT system = project 2)
Assignments

- Two minis (10% each), two projects (20% each)
  - Implementation-oriented, with an open-ended component to each
  - Mini 1 (classification) is out NOW
  - 1 week for minis, ~2 weeks per project, 5 “slip days” for automatic extensions

- Grading:
  - Minis: largely graded based on code performance
  - Projects: graded on a mix of code performance, writeup, extension

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 (40%)
  - Groups of 2 preferred, 1 is possible
  - (Brief!) proposal to be approved by me by the midpoint of the semester
  - Written in the style and tone of an ACL paper
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.
Survey (on Instapoll)

1. Name
2. Fill in: I am a [CS / ____] [PhD / masters / undergrad] in year [1 2 3 4 5+]
3. Write one reason you want to take this class or one thing you want to get out of it
4. One interesting fact about yourself, or what you like to do in your spare time