CS388: Natural Language Processing
Lecture 1: Introduction

Greg Durrett

Administrivia

- Lecture: Tuesdays and Thursdays 12:30pm - 1:45pm
- Piazza: link on the course website
- My office hours: Office hours: Wednesday 4pm, Thursday 2pm
- TA: Uday Kusupati. Office hours: Monday 12pm-1pm, Tuesday 11am-12pm, GDC 1.302

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

Enrollment

- We’ll get as many people in as we can
- Mini1 is out now (due September 10), 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)
- Other NLP offerings:
  - CS378 (ugrad course, taught by me in the spring)
  - LIN 393 (taught by Jessy Li): NLP with minimal supervision
What’s the goal of NLP?

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

**Example:**
- Siri, what’s the most valuable American company?
- Apple
- Who is its CEO?
- Tim Cook

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.

But not long after one of New America’s scholars posted a statement on the think tank’s website praising the European Union’s penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group’s president, Anne-Marie Slaughter, according to the scholar.

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 100 full-time employees and unpaid fellows — would be exited from New America.

The writer and his team were dismissed.

Machine Translation

People’s Daily, August 30, 2017

NLP Analysis Pipeline

- NLP is about building these pieces!
- All of these components are modeled with statistical approaches trained with machine learning
How do we represent language?

- **Text**
  - the movie was good
  - Beyoncé had one of the best videos of all time
- **Labels**
  - subjective
- **Sequences/tags**
  - PERSON: Tom Cruise
  - WORK_OF_ART: Mission Impossible
- **Trees**
  - l eat cake with icing
  - flights to Miami

How do we use these representations?

- **Text Analysis**
  - Labels
  - Sequences
  - Trees
- **Applications**
  - Extract syntactic features
  - Tree-structured neural networks
  - Tree transducers (for machine translation)

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?

Language is Ambiguous!

- 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.
- 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!
- World knowledge: have access to information beyond the training data.
What do we need to understand language?

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

McMahan and Stone (2015)
Golland et al. (2010)

...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.
   \( C_b = \text{John}; C_I = \{\text{John}\} \)
b. He cannot find anyone to take over his responsibilities. (he = John)
   \( C_b = \text{John}; C_I = \{\text{John}\} \)
c. He called up Mike yesterday to work out a plan. (he = John)
   \( C_b = \text{John}; C_I = \{\text{John, Mike}\} \) (CONTINUE)
d. Mike has annoyed him a lot recently.
   \( C_b = \text{John}; C_I = \{\text{Mike, John}\} \) (RETAIN)
e. He called John at 5 AM on Friday last week. (he = Mike)
   \( C_b = \text{Mike}; C_I = \{\text{Mike, John}\} \) (SHIFT)

Centering Theory
Grosz et al. (1995)

A brief history of (modern) NLP

- Largely rule-based, expert systems
- Penn treebank
- Collins vs. Charniak parsers
- Unsup: topic models, grammar induction
- Pretraining

earliest stat MT work at IBM Ratnaparkhi tagger Sup: SVMs, CRFs, NER, Sentiment Neural
earliest stat MT work at IBM Ratnaparkhi tagger Semi-sup, structured prediction

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

What do we need to understand language?
Supervised vs. Unsupervised

- Supervised techniques work well on very little data (even neural networks)
  - Supervised learning
  - annotation (two hours!)
  - 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)

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

Less Manual Structure

The yield on the benchmark issue rose to 10% from 5%

- LSTM

( S ( NP ( NP ( DT The ) ) ) NN yield ...)

- No grammars at all!

Sutskever et al. (2015), Bahdanau et al. (2014)

Interpretability

- Translate

  English | French | Spanish | Chinese - detected

  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 \text{ Hawai'i}
  
  0.005 \text{ LA}
  
  0.0001 \text{ class}
\]

- Use this model for other purposes

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

- Model understands some sentiment?

  \[
\]

- 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 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...

Hamilton et al. (2016)


**Outline**

**ML and structured prediction for NLP**

- Aug 29: Introduction (4pp)
- Sept 3: Binary classification
- Sept 5: Multiclass classification
- Sept 10: Sequence Models 1: HMMs (Guest Lecture: Ray Mooney)
- Sept 12: Sequence Models 2: CRFs
- Sept 17: NN1: Feedforward
- Sept 19: NN2: Word embeddings
- Sept 24: NN3: RNNs
- Sept 26: NN4: Language Modeling and Preserving
- Oct 1: NIS: Interpretability/CNNs/Neural CRFs/etc.

**Neural nets** (this part is still in flux)

- Oct 3: Trees 1: Constituency, PCFGs
- Oct 8: Trees 2: Constituency, Dependency
- Oct 10: Trees 3: Dependency
- Oct 15: Semantics 1
- Oct 17: Semantics 2 / Seq2seq 1
- Oct 22: Seq2seq 2: Attention and Pointers

---

**Outline: Syntax + Semantics**

- Oct 3: Trees 1: Constituency, PCFGs
- Oct 8: Trees 2: Constituency, Dependency
- Oct 10: Trees 3: Dependency
- Oct 15: Semantics 1
- Oct 17: Semantics 2 / Seq2seq 1
- Oct 22: Seq2seq 2: Attention and Pointers

---

**Outline: Applications**

- Oct 24: Machine Translation 1
- Oct 29: Machine Translation 2 / Transformers
- Oct 31: Pretrained Transformers / BERT
- Nov 5: Information Extraction / SRL
- Nov 7: Question Answering 1
- Nov 12: Question Answering 2
- Nov 14: Dialogue
- Nov 19: Summarization
- Nov 21: Multilinguality and morphology
- Nov 26: Wrapup + Ethics

---

**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 2019?
- 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, 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 (October 15)
- Written in the style and tone of an ACL paper

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

Survey (Optional)

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