Natural Language Processing

Dan Klein, GSI: Nick Tomlin
UC Berkeley

Logistics

Requirements
- Engage with the course
- Workload will be high, self-direction
- Patience: class is under construction

ML: A-level mastery, e.g. CS189
PL: Ready to work in PyTorch (on colab)
NL: Care a lot about natural language

Enrollment
- Class is "full" but we’re scaling up
- We will process waitlist after P1
- No materials require enrollment

Remote lecture / chat format for now
Expanded late day policy (14 day)

Enrollment
- Class is "full" but we’re scaling up
- We will process waitlist after P1
- No materials require enrollment

Course expectations
- Readings, lectures, ~8 projects
- No sections, no exams
- Engagement with the course

Resources and Readings

Resources
- Website (syllabus, readings, slides, links)
- Piazza (course communication)
- Gradescope (submission and grades)
- Compute via Colab notebooks
- Berkeley-internal webcasts / recordings

Readings (see webpage)
- Individual papers will be linked
- Optional text: Jurafsky & Martin, 3rd (more NL)
- Optional text: Eisentstein (more NL)

Projects and Infrastructure

Projects
- P1: Language Models
- P2: Question Answering
- P3: Machine Translation
- P4: Speech
- P5: Syntax and Parsing
- P6: Semantics
- P7: Grounding
- P8: Historical Linguistics / TBD

Infrastructure
- Python / PyTorch
- Compute via Colab notebooks
- Grading via Gradescope

What is NLP?
Natural Language Processing

Goal: Deep Understanding
- Requires knowledge, context, and grounding
- Just starting to see successes

Reality: Shallow Matching
- Requires robustness and scale
- Amazing successes, but fundamental limitations

NLP History

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>Weaver on MT</td>
</tr>
<tr>
<td>1960</td>
<td>Bell Labs ASR</td>
</tr>
<tr>
<td>1970</td>
<td>Pre-Compute Era</td>
</tr>
<tr>
<td>1980</td>
<td>Neural ASR</td>
</tr>
<tr>
<td>1990</td>
<td>Neural MT</td>
</tr>
<tr>
<td>2000</td>
<td>Neural TTS</td>
</tr>
<tr>
<td>2010</td>
<td>Pretraining</td>
</tr>
<tr>
<td>2020</td>
<td>“Speech Lab”</td>
</tr>
</tbody>
</table>

Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 1% error for digit strings, 5% conversational speech, still >20% hard acoustics
- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: nearly perfect scale from preciosity

Transforming Language

Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 1% error for digit strings, 5% conversational speech, still >20% hard acoustics
- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: nearly perfect scale from preciosity

Transforming Language
Machine Translation

- Translate text from one language to another.
- Challenges:
  - What's the mapping? (learning to translate)
  - How to make it efficient? (fast translation search)
  - Fluency (next class) vs fidelity (later)

Example: "IMPOTS" to "IMPOSTS"

Machine Translation

Example: Google Translate 2020

Spoken Language Translation

Example: Microsoft Skype via Yejin Choi

Summarization

- Condensing documents.
  - Single or multiple
  - Extractive or abstractive
  - Agglomerative or synthetic

Very context-dependent!

Example: CNN via Wei Gao

Understanding Language

Search, Questions, and Reasoning
Language Comprehension?

Operas refer to a dramatic art form, originating in Europe, in which the emotional content is conveyed to the audience as much through music, both vocal and instrumental, as it through the lyrics. By contrast, in musical theater, an actor's dramatic performance is primary, and the music plays a lesser role. The drama in opera is presented using the primary elements of theater such as scenery, costumes, and acting, whereas the words of the opera, or libretto, are sung rather than spoken. The singers are accompanied by a musical ensemble ranging from a small instrumental ensemble to a full symphonic orchestra.

1. It is printed out in the reading that opera ———.
   - A is has developed under the influence of musical theater
   - B is a drama sung with the accompaniment of an orchestra
   - C is not a high-budget production
   - D is often performed in Europe
   - E is the most complex of all the performing arts

Interactive Language

Example: Virtual Assistants

- Virtual assistants must do:
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech
  - ... and back-end actions!
Conversations with Devices?

ELIZA

Social AIs and Chatbots

Chatbot Competitions!

SoundingBoard Example

Sounding Board’s Architecture?
What is Nearby NLP?
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it
- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!
- Speech Processing
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging

Related Areas

Example: NLP Meets CL
- Example: Language change, reconstructing ancient forms, phylogenies
  - Just one example of the kinds of linguistic models we can build

Why is Language Hard?

Ambiguities

Example: NLP Meets CL
- Example: Language change, reconstructing ancient forms, phylogenies
  - Just one example of the kinds of linguistic models we can build
Problem: Ambiguity

- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Iraqi Head Shoots Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Why are these funny?

44

What Do We Need to Understand Language?

45

We Need Representation: Linguistic Structure

46

Example: Syntactic Analysis

47

We Need Data

48

We Need Lots of Data: MT

49
We Need Models: Data Alone Isn’t Enough!

Learning Latent Syntax

Personal Pronouns (PRP)

<table>
<thead>
<tr>
<th>PRP-1</th>
<th>it</th>
<th>them</th>
<th>him</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP-2</td>
<td>it</td>
<td>he</td>
<td>they</td>
</tr>
<tr>
<td>PRP-3</td>
<td>it</td>
<td>he</td>
<td>I</td>
</tr>
</tbody>
</table>

Proper Nouns (NNP)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP-2</td>
<td>John</td>
<td>Robert</td>
<td>James</td>
</tr>
<tr>
<td>NNP-3</td>
<td>Bush</td>
<td>柰</td>
<td>Peters</td>
</tr>
<tr>
<td>NNP-15</td>
<td>New</td>
<td>San</td>
<td>Wall</td>
</tr>
<tr>
<td>NNP-3</td>
<td>York</td>
<td>Francisco</td>
<td>Street</td>
</tr>
</tbody>
</table>

We Need Knowledge

World knowledge: have access to information beyond the training data

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

Data and Knowledge

- Classic knowledge representation worries: How will a machine ever know that...
  - 1980: write it all down
  - 2020: learn it from data

Knowledge from Pretraining?

We Need Grounding

Grounding: linking linguistic concepts to non-linguistic ones

- What is a green light? How do we understand what “green lighting” does?
- Need commonsense knowledge
Example: Grounded Dialog

When is my package arriving?

Friday!

Example: Grounded Dialog

What’s the most valuable American company?

Apple

Who is its CEO?

Tim Cook

Why is Language Hard?

- We Need:
  - Representations
  - Models
  - Data
  - Grounding
  - Learning
  - Scale
  - Efficient Algorithms

- … and often we need all these things at the same time

What is this Class?

- Three aspects to the course:
  - Linguistic Issues
  - What are the range of language phenomena?
  - What are the knowledge sources that let us disambiguate?
  - What representations are appropriate?
  - How do you know what to model and what not to model?

- Modeling Methods
  - Increasingly sophisticated model structures
  - Learning and parameter estimation
  - Efficient inference: dynamic programming, search, sampling

- Engineering Methods
  - Issues of scale
  - Where the theory breaks down (and what to do about it)

- We’ll focus on what makes the problems hard, and what works in practice…

Class Requirements and Goals

- Class requirements:
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models (parts of CS281a)
    - Basic knowledge of programming
    - Strong coding skills (Python, ML libraries)
  - Most people are probably missing one of the above
  - You will often have to work on your own to fill these gaps

- Class goals:
  - Learn the issues and techniques of modern NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
  - See where the limits are in the field still are!

- This semester: new projects, new topics, lots under construction!