NLP 201: Natural Language Processing 1

Introduction

Jeffrey Flanigan
October 1, 2020

University of California Santa Cruz
jmflanig@ucsc.edu
Plan for today

- Administrative information
- Introductions
- Begin introduction to NLP
Your Instructors

Jeff (instructor):

- UCSC professor since 2019, Ph.D. from CMU in 2018
- Research: machine learning for structured problems in NLP, deep learning, semantics in NLP

TAs:

- Geetanjali Rakshit
- Rongwen Zhao
- Zekun Zhao (shared among the classes)

For each assignment, either Geetanjali and Rongwen will be responsible for questions (as well as Jeff)
Resources

- Piazza - please sign up
- Course website: https://courses.soe.ucsc.edu/courses/nlp201/Fall20/01
- Canvas (for exams, assignment turn-in, and some materials)
- Classes will be recorded
- Assignments will be done either locally or on Google Colab
- We accommodate disabilities. If you require DRC accommodations (https://drc.ucsc.edu/), please let me know ASAP
Outline of NLP 201

1. Introduction to NLP
2. Text classifiers, probabilistic language models
3. Sequence models
4. Syntax and parsing
5. TBD

Don’t expect a detailed calendar since plans may change as we go. This year-long series (NLP 201-3) is new, with no comparable course offering anywhere in the world (as far as I know). We will be developing the course materials as we go along.

If you have taken my undergraduate NLP course CSE 143 Winter 2020, the first quarter may be familiar, but at a slower, more in-depth pace.
Evaluation

- 4 assignments (A1–4), completed individually (60%)
- Midterm exam (20%), towards the middle of the quarter on Canvas
- Final exam (20%), to take place at the end of the quarter on Canvas
• Let’s turn on our cameras
• We’ll go around the room and share one of these
  • What you hope to learn about NLP *OR*
  • What excites you most about NLP *OR*
  • An experience you’ve had with NLP that you enjoyed
(Short 3 min break)
What is Natural Language Processing (NLP)?

- The set of methods for making human language accessible to computers (Eisenstein, 2018).

Why do we want this?
What is Natural Language Processing (NLP)?

• The set of methods for making human language accessible to computers (Eisenstein, 2018).
What is Natural Language Processing (NLP)?

- The set of methods for making human language accessible to computers (Eisenstein, 2018).
- Why do we want this?
Communication with Machines

~ 50s-70s

~ 80s

~ today
NLP Application: Machine translation
NLP Application: Question Answering

- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?
NLP has many end-user tasks (downstream tasks or applications)

- Machine translation
- Summarization
- Question answering
- Conversational agents
- Search (information retrieval)
- Recommender systems
- Document classification
NLP has many end-user tasks *(downstream tasks or applications)*

- Machine translation
- Summarization
- Question answering
- Conversational agents
- Search (information retrieval)
- Recommender systems
- Document classification

These two tasks are **supertasks**.
Downstream tasks sometimes benefit from **intermediate tasks**

- Knowing a word’s sense (i.e. duck – animal vs duck – action) could help translate it. This is **sense disambiguation**.

- Knowing if a word is a verb or noun (its part of speech) could help translate it (duck – noun vs duck – verb). This is **part-of-speech tagging**.

- Splitting text into sentences is often required before processing. This is **sentence segmentation**.

- Deciding what should count as a word ($100 vs $\_100$ or it’s vs it\_‘s) (**tokenization**) usually has a very large effect on performance.
Examples of intermediate tasks

- Tokenization
- Language modeling
- POS tagging
- Synactic parsing
- Entity recognition
- Entity linking
- Relation extraction
- Semantic role labeling
- Semantic parsing
The traditional NLP pipeline

1. Tokenization
2. Morphological analysis
3. Part-of-speech tagging
4. Syntactic Parsing
5. Semantic Parsing
6. Downstream task: QA, summarization, etc

With deep learning, sometimes tasks are done \textit{end-to-end}, without any intermediate steps.
Large growth in NLP in recent years
NLP applications are now commonplace

- Spam email filtering
- Google translate
- Built-in recommender systems (in Amazon, Ebay, Netflix, etc)
- Siri, Amazon Alexa
- Auto-completion suggestions
- Grammar checking
- Automatic essay grading (used by ETS)
- Inappropriate social media post filtering
- Fake news detection
- Lots we probably don’t even realize!
Ethics

- Can run into issues like censorship, bias, security, etc
- Active area of research
• Speech (both recognition and generation) are separate, not an NLP tasks
• Machine learning (computers learn from experience or examples)
• Linguistics (the study of language).
• Computational linguistics (CL)
  • Sometimes synonymous with NLP
  • In practice, CL often has larger emphasis on linguistics and linguistic theories. CL degree programs often have a different curriculum than NLP degree programs