NLP 201: Introduction

Jeffrey Flanigan
University of California Santa Cruz
jmflanig@ucsc.edu

September 22, 2022

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: core NLP tasks, deep learning, semantics in NLP

TA:

- Zekun Zhao, PhD student

Administrative

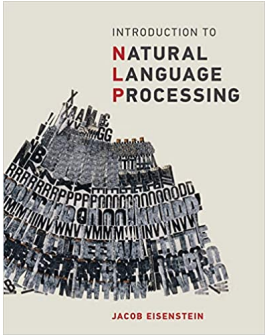
- Classes will be simultaneously on Zoom and recorded
- Assignments will be done either locally or on Google Colab
- We accommodate disabilities. If you require DRC accommodations (`https://drc.ucsc.edu/`), please let me know ASAP

Resources

- Course website: `https://courses.soe.ucsc.edu/courses/nlp201/Fall21`
- Canvas (for videos, exams, assignment turn-in, and some materials)
- NLP wiki: `https://jlab.soe.ucsc.edu/nlp-wiki` Please do not share widely
- No official textbook. Readings posted on Canvas or the website

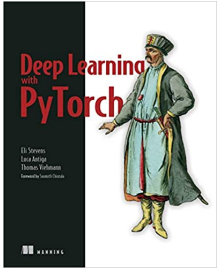
Textbooks and Readings

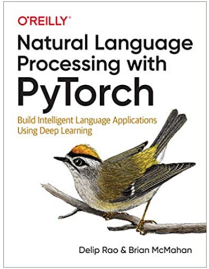
- No official textbook. Readings posted on Canvas or the website
- Highly recommend this book



PyTorch Books

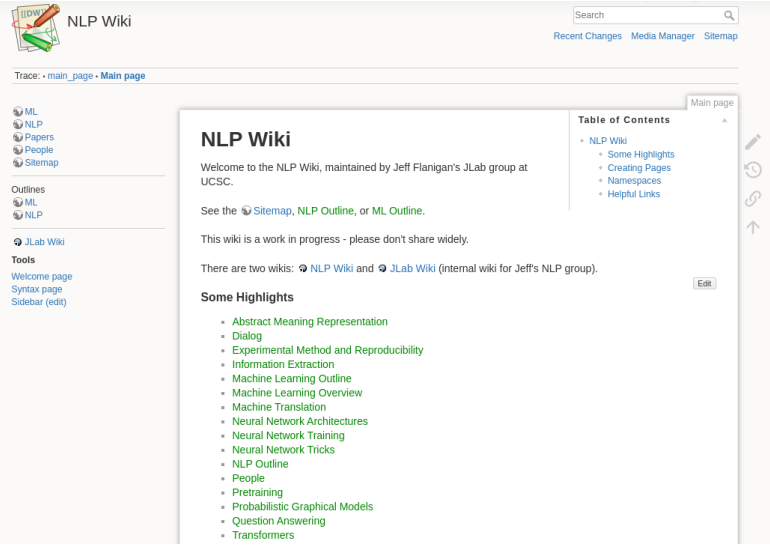
Highly recommend these two books





We won't be using deep learning in this course, but you will in NLP 243 and NLP 202

NLP Wiki



NLP Wiki

Search

Recent Changes Media Manager Sitemap

Trace: • main_page • **Main page**

Main page

- ML
- NLP
- Papers
- People
- Sitemap

Outlines
- ML
- NLP

JLab Wiki

Tools

Welcome page
Syntax page
Sidebar (edit)

NLP Wiki

Welcome to the NLP Wiki, maintained by Jeff Flanigan's JLab group at UCSC.

See the Sitemap, NLP Outline, or ML Outline.

This wiki is a work in progress - please don't share widely.

There are two wikis: NLP Wiki and JLab Wiki (internal wiki for Jeff's NLP group).

Edit

Some Highlights

- Abstract Meaning Representation
- Dialog
- Experimental Method and Reproducibility
- Information Extraction
- Machine Learning Outline
- Machine Learning Overview
- Machine Translation
- Neural Network Architectures
- Neural Network Training
- Neural Network Tricks
- NLP Outline
- People
- Pretraining
- Probabilistic Graphical Models
- Question Answering
- Transformers

Table of Contents

- NLP Wiki
  - Some Highlights
  - Creating Pages
  - Namespaces
  - Helpful Links

Active Learning

My lectures will sometimes incorporate "active learning" (the education term, not the machine learning term).
For example:

- You each have a whiteboard, and I may ask you to write things from time to time, and discuss with a partner or small group
- Studies show: generally students don't like active learning, but it greatly improves learning
- If you are on Zoom or watching the video later, you can post in the chat or email me your responses

Evaluation

- 4 assignments (A1–4), completed individually (45%)
  You get a 24 hour grace period to turn it in late
- Quizes (10%) weekly on Canvas
- Midterm exam (15%), towards the middle of the quarter on Canvas
- Final exam (25%), to take place at the end of the quarter
- Attendance / Participation (5%) For attending each class and participating

Academic Integrity

- Assignments and tests are to be completed individually
- Do not look up or copy either code or solutions from others or the internet
- My assignments are designed to aid learning
- They are not designed to prevent copying
- I expect you to take responsibility for your learning

You are here to learn: plagiarism will only hurt **you**

Assignment "redos"

- Each student gives feedback on two assignments
- Three days for feedback
- One week for changes for final submission which TA grades
- Writeup explanation of what you changed, and where you learned it (This is important. You get marked down without it)
- We check if you copy

For the final deadline, assignments may be turned in up to 24 hours after the deadline for a 10% grade penalty. After 24 hours, assignments will receive zero credit.

Assignment zero

- Assignment zero out today or tomorrow
- Won't count towards your grade
- TA will go over it in section

Outline of NLP 201

Basics

1. **Introduction to NLP**
2. **Finite-state methods**
3. **Language models**
4. **Sequence models**
5. **Graphical models**
6. **Conditional random fields (CRFs)**
7. **Overview of NLP tasks**

Outline of NLP 202

Syntax and Advanced topics

1. **Deep learning on GPUs**
2. **Syntax and parsing**
3. **Dependency Parsing**
4. **Structured prediction and loss functions**
5. **Optimization for deep learning**
6. **Neural network tricks**

Outline of NLP 203 (taught by another instructor)

Applications

1. **Evaluation and statistical significance**
2. **Ethics**
3. **Summarization**
4. **Machine translation**
5. **Question answering**
6. **Semantic parsing**
7. **Information extraction**

Questions?

Introductions

- If you're on Zoom, try to turn on your cameras
- Lets 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

What is Natural Language Processing (NLP)?

- (3 Minutes) On your whiteboard or in the chat, write
  - 2-3 things (tasks, areas of study, etc) you think that are NLP
  - 2-3 things you think are **not** NLP (but potentially confused with NLP)
- (3 minutes) Discuss with the person next to you or with the others in Zoom
- Hold up your boards and we'll discuss

What is Natural Language Processing (NLP)?

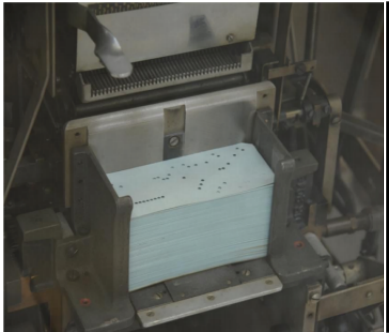
What is Natural Language Processing (NLP)?

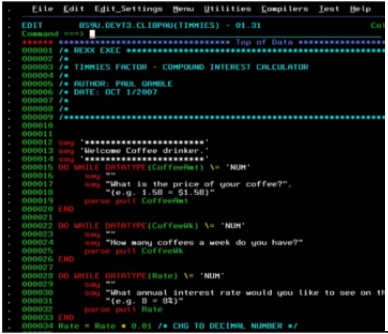
- The set of methods for making human language accessible to computers (Eisenstein, 2018).

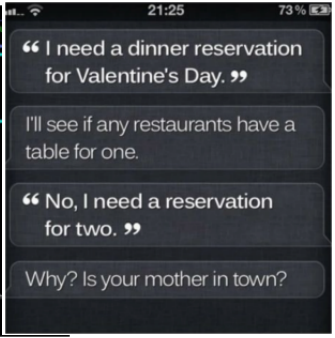
What is Natural Language Processing (NLP)?

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

Communication with Machines





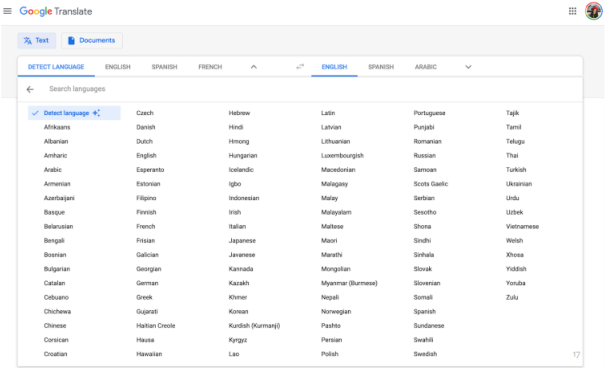


~ 50s-70s

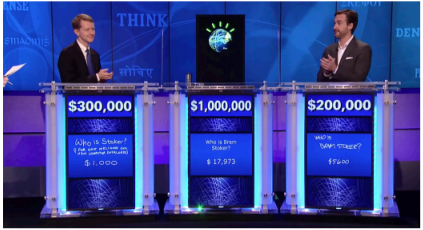
~ 80s

~ 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

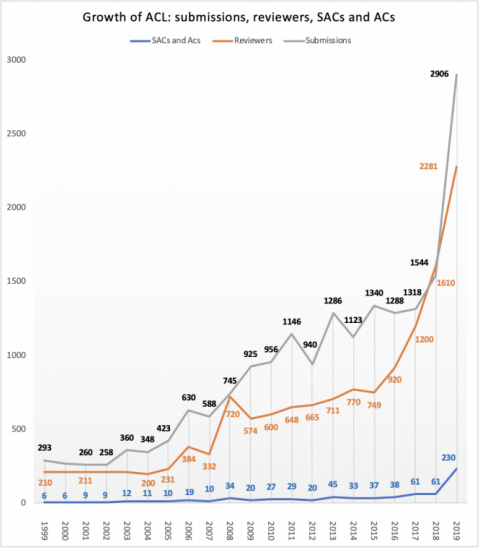
- Tokenization
- Language modeling
- POS tagging
- Synactic parsing
- Entity recognition
- Entity linking
- Relation extraction
- Semantic role labeling
- Semantic parsing

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 **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

Relation of NLP to other fields

- 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

References I

Jacob Eisenstein. *Natural Language Processing*. 2018. URL https://github.com/jacobeisenstein/gt-nlp-class/raw/master/notes/eisenstein-nlp-notes.pdf.