COS 484: Natural Language Processing

L1: Introduction to NLP

Spring 2023
Logistics and course structure
Course staff

Instructor

Danqi Chen

Graduate TAs

Alexander Wettig

Austin Wang

Howard Chen

Samyak Gupta
Logistics

Course webpage: https://nlp.cs.princeton.edu/cos484/

- Contains all the detailed information about the course (slides, reading lists, assignments, policy, office hours, etc).
- Canvas will be only used for announcements - make sure you have notifications turn on!

Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topics</th>
<th>Readings</th>
<th>Assignments</th>
</tr>
</thead>
</table>
| 1    | Mon (1/30) | Introduction to NLP       | 1. Advances in natural language processing  
2. Human Language Understanding & Reasoning | A0 out      |
|      | Wed (2/1) | n-gram language models | J & M 3.1-3.4                                                             |             |
| 2    | Mon (2/7) | Text classification     | J & M 4.1-4.8                                                             | A0 due, A1 out |
|      | Mon (2/8) | Word embeddings 1        | J & M 6.2-6.4, 6.6  
Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors |             |

Slides will be available before the class.

All readings are optional but can help strengthen your understanding.

All assignments are due on Monday at 9:30 before the lecture.
Logistics

- **Lectures**: Monday / Wednesday 9:30-10:50 at Bowen Hall 222 (no Zoom support)

- **Precepts**: 1-hour precept every week taught by TAs (optional)
  - Refresher for basic maths, review course materials, Colab/Pytorch tutorial etc
  - Likely Friday - we will send out a poll later today!
Logistics

• Sign up for **Ed** and **Gradescope** today
  - We will use **iClicker** for polls in the class… Will make an announcement before Wednesday!

• **Ed** will be the main forum for all class-related questions and discussion

• We also provide a mailing list **cos484-2023@googlegroups.com** for emergencies, or personal matters that you don’t wish to put in a private Ed post.
  - The mailing list only has instructor + grad TAs. Please don’t write to the instructor directly unless you HAVE TO.

• We would like to help in our office hours!
  - All the information will be posted this week on the website
  - We will also have a number of undergraduate TAs’ office hours
Course structure

Assignments (40%): 4 + 1 = 5 in total

- **A0 (4%)** is a warm-up assignment and will be released TODAY and due in one week
- **A1, A2, A3, A4**: 9% each - each assignment has 2 or 3 weeks
- Every assignment has a written component and a programming component based on Colab (A4 requires access to GPUs).
- You will need to know how to program in PyTorch/Numpy.

You have **96 free late hours** for all assignments; After that, 10% penalty for each late day (up to a maximum of 3 days beyond which submissions will not be accepted)
Course structure

Midterm (25%)
• A 3-hour timed exam on Gradescope in a 27-hour window (March 8-9)
• All the topics up to March 1 will be covered
• No final exam
Course structure

Final project (35%)

• Complete in a team of 3
• Two options: (a) reproducing a state-of-the-art NLP paper (ACL/EMNLP/NAACL 2020-2022) and coming up with your analysis, ablations or innovations (strongly encouraged) (b) competing a research project
• Proposal (0%) due on March 20
• Poster presentation tentatively scheduled on May 3 (during reading period)
• Final report due on May 9

Extra bonus (5%) - participation in class and Ed discussions

No pre-determined cut-offs for final grades, will be decided at the end taking into account the performance of the entire class and will be fairly assigned to measure your level of understanding of the subject.
Textbooks
(NLP is a rapid-moving field…)

https://web.stanford.edu/~jurafsky/slp3/

Speech and Language Processing (3rd ed. draft)
Dan Jurafsky and James H. Martin

Here’s our Jan 7, 2023 draft! This draft is mostly a bug-fixing and restructuring release, there are no new applications section earlier, reflecting how we and others tend to teach NLP, and combines the linguisti

A good way to learn about state-of-the-art NLP concepts is through research papers and blog posts.
Course goals

- Gain an understanding of the **fundamentals** of different sub-fields within NLP
- Understand **theoretical** concepts and algorithms
- **Hands on experience** building statistical models for language processing
- Carry out an **independent research project** at the end
This is an advanced class
This is a very advanced class
Prerequisites

- **Required:** **COS324**, knowledge of probability, linear algebra, calculus (A0 will give you a sense)
- Be ready to pick up new ML concepts
- Proficiency in Python: programming assignments and projects will require use of Python, Numpy and PyTorch.

Q. Why is COS324 a prerequisite?

We assume you have learned the following concepts already:
- Language models
- Logistic regression w/ regularization
- Unsupervised vs supervised learning
- Feedforward neural networks, convolutional neural networks
- PyTorch programming
- (A little bit of reinforcement learning)
What is NLP and what is the course about
Natural Language processing

• NLP = building computer programs to analyze, understand and generate human language - either spoken or written (informal)

• NLP is an interdisciplinary field
Natural Language processing

- NLP = building **computer programs** to analyze, understand and generate **human language - either spoken or written** (informal)

Communication with humans (ex. personal assistants, customer service)

Access the wealth of information about the world — crucial for AI systems

- Banking assistant
  - **ONLINE**

- Offline
  - **OFFLINE**
Computer learns to play Civilization by reading the instruction manual

By Matthew Rogers on July 14, 2011 at 5:03 pm 16 Comments

MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

The MIT Computer Science and Artificial Intelligence lab has a computer that now plays Civilization
Turing Test

Ability to understand and generate language ~ intelligence


COMPUTING MACHINERY AND INTELLIGENCE
By A. M. Turing

1. The Imitation Game
A brief history of NLP

1949
Weaver’s memorandum

1960s
Grammar Theories

1970s
Conceptual Ontologies

1980s
Symbolic Models

1990s
Statistical Models

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbc937ebce
A brief history of NLP

2003: Neural Language Models
2008: Multi-task Learning
2013: Word Embeddings
2013: NLP Neural Nets
2014: Seq-to-seq Learning
2015: Attention
2018: Pretrained Models
2018: BERT
2019: T5, RoBERTa
2020: GPT-3
2022: ChatGPT

How “neural” is this course?

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37
How it started

“Within three or five years, machine translation will be a solved problem”

Georgetown experiment
1954

How it’s going
<table>
<thead>
<tr>
<th>English</th>
<th>Hindi</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man shot the elephant while wearing his pajamas</td>
<td>आदमी ने अपना पजामा पहने हुए हाथी को गोली मार दी</td>
<td>Man shoots elephant wearing his pajamas</td>
</tr>
</tbody>
</table>

(Based on Google Translate result in 2023-01)
> How many red blocks are there?  
- THREE OF THEM

> Pick up the red block on top of a green one  
OK.

- Rule-based, requiring extensive programming
- Limited domain
Statistical learning

- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

Unsupervised vs. supervised?
Statistical learning

IBM translation models

Speech recognition

Anytime a linguist leaves the group the (speech) recognition rate goes up

- Fred Jelinek 1998
The era of deep learning

- Significant advances in core NLP technologies
- **Essential ingredient**: large-scale supervision, lots of compute
- Reduced manual effort - less/zero **feature engineering**

36M sentence pairs

**Russian**: Машинный перевод - это круто!

**English**: Machine translation is cool!
The era of pre-training / LLMs

- BERT, ELMo, ERNIE...
  - Leverages a lot of unlabeled text
  - Model size increased by $10^3 - 10^5$ x in parameters
Here are some NLP topics that are popular and relevant in 2023:

1. Transformers (BERT, GPT, etc.)
2. Transfer Learning in NLP
3. Multi-lingual NLP
4. Attention Mechanisms
5. Adversarial Training in NLP
6. NLP for Question Answering
7. Sentiment Analysis
8. Text Generation (summarization, translation, etc.)
9. Named Entity Recognition
10. Dialogue Systems

Note: The field of NLP is constantly evolving, so it's important to stay up-to-date with recent advances and adjust your teaching content accordingly.
ChatGPT examples

what do you know about the Princeton NLP group?

The Princeton NLP group is a research group at Princeton University that focuses on the study of natural language processing (NLP) and its applications. The group is led by Professor Christopher Manning, who is a leading expert in the field of NLP. The group has published numerous research papers and has developed several widely-used NLP tools, including the Stanford Parser and Stanford Named Entity Recognizer. The group also collaborates with other research groups and organizations, such as the Allen Institute for Artificial Intelligence and Google, on NLP research projects.

Why do you hallucinate so much?

As an AI language model, I don't have the capacity to hallucinate as I don't possess emotions, sensations, or consciousness. I can only generate responses based on patterns in the text I was trained on.
NLP subfields

- Machine translation
- Question answering
- Information extraction
- Text summarization
- Dialogue
- Semantic parsing
- Syntactic parsing
- Language grounding
- …. 

The Massachusetts Institute of Technology (MIT) is a private research university in Cambridge, Massachusetts, often cited as one of the world's most prestigious universities. Founded in 1861 in response to the increasing industrialization of the United States, …

City: Cambridge, MA
Founded: 1861
Mascot: Tim the Beaver

Article

Database

Tell my wife I love her
Action: Text

Arg1: Sarah Fox
Arg2: I love you

Question: What is the capital of Germany?
Context: The capital of Germany is the city state of Berlin. It is the seat of the President of Germany, …

“Berlin”
Topics we will cover in this class

- We will focus on fundamentals, important concepts rather than concrete applications
- The class will be a mix of statistical and neural NLP approaches (>70% neural)

Before midterm
- N-gram language models
- Text classification
- Word embeddings $\times 2$
- Sequence models $\times 2$
- Parsing $\times 2$
- Neural networks for NLP

After midterm
- Recurrent neural networks $\times 2$
- (neural) machine translation $\times 2$
- **Transformers** $\times 2$
- BERT, ELMo, …
- **Large language models** + more guest lectures!
- Question answering
- Language grounding
Why is language difficult to understand?
Why is language difficult to understand?

- Ambiguous
- Dialects
- Accents
- listener has to infer - pragmatics
- humor, sarcasm, irony
- context, dependencies
Lexical ambiguity

The fisherman went to the *bank*.

**bank**¹
/baNGk/  
*noun*  
plural noun: banks  
1. the land alongside or sloping down to a river or lake.  
   "willows lined the bank"  
   *synonyms:* edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe.  
   More  
1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.  
   "I paid the money straight into my bank"  
   *synonyms:* financial institution, merchant bank, savings bank, finance company, trust company.
Lexical ambiguity

The fisherman went to the *bank*. He deposited some money.

*bank*¹

/ˈbæŋk/ (ˈbæŋk/)
noun
plural noun: banks

1. the land alongside or sloping down to a river or lake.
   "willows lined the bank"
   synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe. More

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.
   "I paid the money straight into my bank"
   synonyms: financial institution, merchant bank, savings bank, finance company, trust company,

Word sense disambiguation
Lexical variations

According to the thesaurus, “They’re humid, prepossessing homosapiens with full sized aortic pumps” means “They’re warm, nice people with big hearts.”

Several words can mean the same thing!
Distributed representations

Project words onto a continuous vector space

Similar words closer to each other

\[ v(\text{king}) - v(\text{man}) + v(\text{woman}) = v(\text{queen}) \]
Comprehending word sequences

- My brother went to the park near my sister’s house

- Park my went house near to sister’s my brother the

- “My brother went park near sister’s house”?  

- The old man the boat

- Fat people eat accumulates

Garden Path sentence

- Implicit structure in all languages
- Coarse-to-fine levels (recursive)
- What are some good data structures to represent this?
Syntactic ambiguity

I shot an elephant in my pajamas

Human language is full of such examples!
Syntactic parsing

Parse tree

Sentence

His duties will be assumed by John Smith who has been elected deputy chairman

Online tools: http://nlp.stanford.edu:8080/corenlp/
Discourse ambiguity

- The man couldn’t lift his son because he was so heavy.
- The man couldn’t lift his son because he was so weak.

  What does “he” refer to?

- The city councilmen refused the demonstrators a permit because they feared violence.
- The city councilmen refused the demonstrators a permit because they advocated violence.

  What does “they” refer to?

Anaphora resolution
Next lecture: n-gram language models