Course introduction

CS 685, Spring 2024

Advanced Natural Language Processing

http://people.cs.umass.edu/~miyyer/cs685/

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

Course logistics

- Follow along w/ the lectures either in-person or online via YouTube
- There will normally be a short quiz about the week's topics to be submitted on Gradescope (none for the first week!)
- Gradescope for all assignment submissions

who?

TAs:

Chau Pham
Yekyung Kim
Katherine Thai
Saurabh Bajaj

Check out nlp.cs.umass.edu for news/info on NLP research going on at UMass!

email all of us (including me!) at cs685instructors@gmail.com

course website:
https://people.cs.umass.edu/~miyyer/cs685

Office hours (in-person and on zoom)

Monday w/ Katherine: 10-11AM on Zoom
Tuesday w/ Yekyung: 2-3PM in CS207 Cube 3
Wednesday w/ Mohit: 4-5PM in CS232
Thursday w/ Chau: 1-2PM in CS207 Cube 1
Friday w/ Saurabh: 11AM-12PM in CS207
Zoom links on Piazza

If necessary, TA office hours will be extended by one hour during homework / exam weeks

Office hours will begin next Monday 2/12 (none before then)

waitlist override pass/fail etc.

- don’t email us about getting into the class because we can’t help… please contact Jess Kadarisman at jkadarisman@cs.umass.edu with such questions or requests
- Add/drop deadline is **Feb 14** for graduate students and **Feb 7** for undergrads

anonymous questions / comments?

- submit questions/concerns/feedback to https://forms.gle/wtSgjAQ3aa9z29ux5
- we will go over some/all submitted responses at the start of every class
- does this course require prior knowledge of NLP? *No, but basic ML/probability/stats/ programming will help a lot*
- Size of final project groups? *4-5*
- Will we have notes? *Slides will be posted before the lecture, any notes will be posted after*

No official prereqs, but the following will be useful:

- comfort with programming
  - We'll be using Python (and PyTorch) throughout the class
- comfort with probability, linear algebra, and mathematical notation
- Some familiarity with matrix calculus
- Excitement about language!
- Willingness to learn

Please brush up on these things as needed!

Grading breakdown

- 5% weekly quizzes
- 30% problem sets (hw0, hw1, hw2, *hw3*)
  - Written: math & concept understanding
  - Programming: in Python
- 25% exam (~April 10, **in-class exam**)
- 40% final projects (groups of 4-5)
  - Choose any topic you want
  - Project proposal (10%)
  - Final report / presentation (30%)

Extra credit

- There will be many NLP seminar talks this semester
  - Schedule at: https://people.cs.umass.edu/~miyyer/nlpseminar/
- Remotely attend up to **five** of these talks (or watch their recordings) and then complete a writeup about each
- In total, earn up to 3% on top of your final grade

Readings

- No need to buy any textbooks!
- Readings will be provided as PDFs on website
  - Usually NLP research papers / notes

Previous class videos / material

- Fall 2020: https://people.cs.umass.edu/~miyyer/cs685_f20
- Fall 2021: https://people.cs.umass.edu/~miyyer/cs685_f21/
- Fall 2022: https://people.cs.umass.edu/~miyyer/cs685_f22/
- Spring 2023: https://people.cs.umass.edu/~miyyer/cs685_s23/
  - Feel free to use these materials / videos to study!
  - This course will have a lot of overlap with the S23 edition
  - That said, there will be quite a bit of interesting new stuff later in the semester!

natural language processing

natural language processing

languages that evolved naturally through human use

e.g., Spanish, English, Arabic, Hindi, etc.

natural language processing

supervised learning: *map text to* **X**
unsupervised learning: *learn* **X** *from text*
generate *text from* **X**

Levels of linguistic structure

Discourse

Semantics

Syntax: Constituents

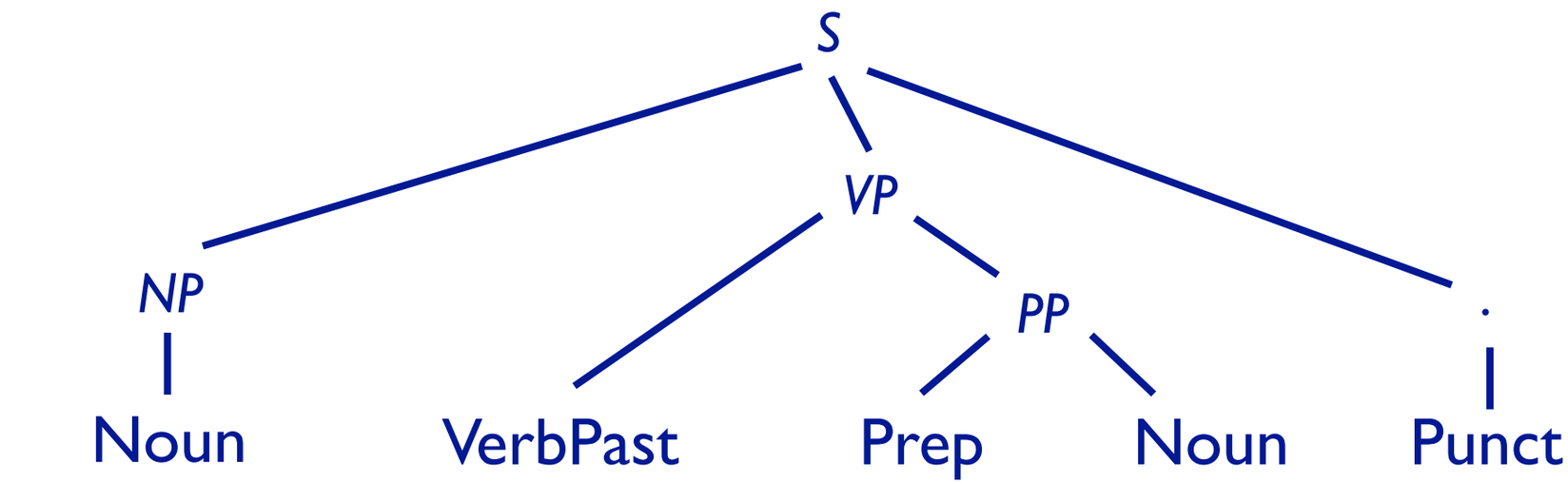
Syntax: Part of Speech

Words

Morphology

Characters

CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)
SpeakerContext(s)
TemporalBefore(e, s)



Alice talked to Bob .

talk -ed [VerbPast]

Alice talked to Bob.

supervised learning: given a collection of labeled examples (where each example is a text *X* paired with a label *Y*), learn a mapping from *X* to *Y*

Example: given a collection of 20K movie reviews, train a model to map review text to review score (*sentiment analysis*)

self-supervised learning: given a collection of *just text,* without extra labels, create labels out of the text and use them for *pretraining* a model that has some general understanding of human language

- **Language modeling**: given the beginning of a sentence or document, predict the next word
- **Masked language modeling**: given an entire document with some words or spans masked out, predict the missing words

How much data can we gather for these tasks?

transfer learning: first *pretrain* a large self-supervised model, and then *fine-tune* it on a small labeled dataset using supervised learning

Example: pretrain a large language model on hundreds of billions of words, and then fine-tune it on 20K reviews to specialize it for sentiment analysis

in-context learning: first *pretrain* a large self-supervised model, and then *prompt* it in natural language to solve a particular task without any further training

Example: pretrain a large language model on hundreds of billions of words, and then feed in “what is the sentiment of this sentence: <insert sentence>”

Language models!

api.together.xyz

What are people using LLMs for?

Figure 3: Topic distribution of 100K sampled conversations. Manual inspection of cluster centroids

 https://arxiv.org/pdf/2309.11998.pdf

Rough list of topics

- **Background**: language models and neural networks
- **Models**: RNNs > Transformers
  - ELMo > BERT > GPT3 > ChatGPT > today's LLMs
- **Tasks**: text generation (e.g., translation, summarization), classification, retrieval, etc.
- **Data**: annotation, evaluation, artifacts
- **Methods**: pretraining, finetuning, preference tuning, prompting

Final projects

Timeline

- All groups should be formed by **2/16**
  - Groups of 4, either form them yourselves and tell us, or we will randomly assign you on 2/17
- Only two deliverables:
  - project proposal: 3+ pages, due **3/8**
  - final report/code: 8+ pages, due last day of classes
- Almost completely open-ended!
  - All projects must involve natural language data
  - There should be a significant coding component of every project

Project

- Either *build* natural language processing systems, or *apply* them for some task.
- Use or develop a dataset. Report empirical results or analyses with it.
- Different possible areas of focus
  - Implementation & development of algorithms
  - Defining a new task or applying a linguistic formalism
  - Exploring a dataset or task

Formulating a proposal

- What is the ***research question***?
- What's been done before?
- What experiments will you do?
- How will you know whether it worked?
  - If data: held-out accuracy
  - If no data: manual evaluation of system output. Or, annotate new data

Feel free to be ambitious (in fact, we explicitly encourage creative ideas)! Your project doesn't necessarily have to "work" to get a good grade.

NLP Research

- All of the best NLP publications are open access!
  - The ACL Anthology (https://aclanthology.org/) contains papers from all of the top NLP conferences (e.g., ACL, EMNLP, NAACL) spanning many decades
  - Machine learning conferences (ICLR, NeurIPS, ICML)
  - Check out arXiv CS-CL (https://arxiv.org/list/cs.CL/recent) for the most recent papers!
  - This is a fast-moving field, so follow NLP researchers on Twitter for discussion on the latest advances
- Use Google Scholar and Semantic Scholar to search for relevant papers

An example proposal

- Introduction / problem statement
- Motivation (why should we care? why is this problem interesting?)
- Literature review (what has prev. been done?)
- Possible datasets
- Evaluation
- Tools and resources
- Project milestones / tentative schedule

Sample projects from last year

Taller, Stronger, Sharper:
Probing Comparative Reasoning Abilities of Vision-Language Models

Examining Medical Narratives of Eating Disorder Recovery on Reddit

Replication of TagRec, a Hierarchical Taxonomy Tagging Model

Learning Schematic and Contextual Representations for Text-to-SQL Parsing

Syllamo: Generating Keyword Mnemonics for Vocabulary Acquisition

Broader ideas

https://2024.aclweb.org/calls/main_conference_papers/#call-for-main-conference-papers

https://colmweb.org/cfp.html

Sources of data

- All projects must use (or make, and use) a textual dataset. Many possibilities.
  - For some projects, creating the dataset may be a large portion of the work; for others, just download and more work on the system/modeling side

- SemEval and CoNLL Shared Tasks:
  dozens of datasets/tasks with labeled NLP annotations
  - Sentiment, NER, Coreference, Textual Similarity, Syntactic Parsing, Discourse Parsing, and many other things...
  - e.g. SemEval 2015 ... CoNLL Shared Task 2015 ...
  - https://en.wikipedia.org/wiki/SemEval (many per year)
  - http://ifarm.nl/signll/conll/ (one per year)

- General text data (not necessarily task specific)
  - Books (e.g. Project Gutenberg)
  - Reviews (e.g. Yelp Academic Dataset https://www.yelp.com/academic_dataset)
  - Web
  - Tweets

Be on the lookout for

- **HW0:** released today, due 2/16 (11:59pm) on Gradescope
- Readings on language models for Wednesday
- **Final project:** Organize into groups of 4 or 5 by 2/16
- **Final project:** project proposal due 3/8

Having issues accessing Piazza/Gradescope/videos? Email the instructors account!