course introduction

CS 685, Fall 2021

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:
Tu Vu
Katherine Thai

The TAs are my own PhD students and have lots of NLP research experience!

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)

Tuesday w/ Katherine: 11am-12pm in CS207 Cube 1
Thursday w/ Mohit: 3-4PM in CS258
Friday w/ Tu: 2-3pm in CS207 Cube 4
Zoom links on Piazza

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

Office hours will begin next Thursday 9/9 (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 Sep 15 for grad students, Sep 8 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
  - From this week: 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*
  - 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

- 10% weekly quizzes
- 30% problem sets (hw0, hw1, hw2*)
  - Written: math & concept understanding
  - Programming: in Python
  - All HWs will be on Google Colab
- 25% exam (late Oct or early Nov, open book/ internet, 24 hours to complete)
- 35% final projects (groups of 4)
  - Choose any topic you want
  - Project proposal (10%)
  - Final report / presentation (25%)

Readings

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

F2020 class videos / material

- The Fall 2020 version of 685 was completely remote. All of its lecture videos / materials are at https://people.cs.umass.edu/~miyyer/cs685_f20
- Feel free to use these materials / videos to study!
- This course will obviously have a lot of overlap with the prior iteration
- That said, there will be some interesting new stuff not covered last year!

natural language processing

natural language processing

languages that evolved naturally through human use
e.g., Spanish, English, Arabic, Hindi, etc.

NOT: controlled languages (e.g., Klingon)
NOT: programming languages

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

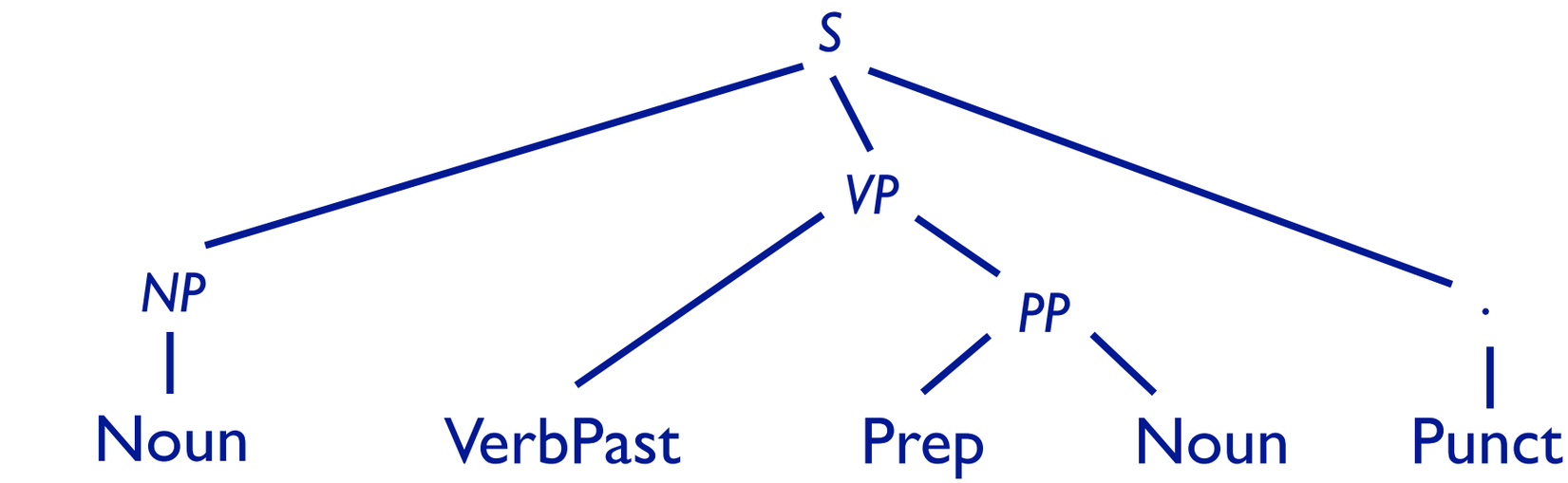
Syntax: Part of Speech

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 (each example is a document X paired with a label Y), learn a mapping from X to Y

Tasks commonly tackled in a supervised setting:

- **Sentiment analysis**: map a product review to a sentiment label (positive or negative)
- **Question answering**: given a question about a document, provide the location of the answer within the document
- **Textual entailment**: given two sentences, identify whether the first sentence entails or contradicts the second one
- **Machine translation**: given a sentence in a source language, produce a translation of that sentence in a target language

self-supervised learning: given a collection of just text (no extra labels), create labels out of the text and use them for *representation learning*

- **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?

representation learning: given some text, create a representation of that text (e.g., real-valued, low-dimensional vectors) that capture its linguistic properties (syntax, semantics)

word	**dim0**	**dim1**	**dim2**	**dim3**
today	0.35	-1.3	2.2	0.003
cat	-3.1	-1.7	1.1	-0.56
sleep	0.55	3.0	2.4	-1.2
watch	-0.09	0.8	-1.8	2.9

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

- Transfer learning has recently (last ~2 years) become the method of choice for most downstream NLP tasks.

Rough list of topics

- **Background**: language models and neural networks
- **Models**: RNNs > Transformers, ELMo > BERT > GPT3, also many others
- **Tasks**: text generation (e.g., translation, summarization), classification, sequence labeling, retrieval, etc.
- **Data**: annotation, evaluation, artifacts
- **Ethics**: bias amplification, privacy issues
- **Methods**: transfer learning, few-shot learning, prompt-based learning

New topics for Fall 2021

- Prompt-based learning
- Efficient Transformer variants
- Large-scale multilingual pretrained models
- Tokenization-free approaches to NLP
- New language+vision approaches (e.g., CLIP, DALL-E)
- *… potentially others! feel free to suggest things too*

Final projects

Timeline

- All groups should be formed by Sep 15
  - Groups of 4, either form them yourselves and tell us, or we will randomly assign you on 9/15
- Only two deliverables:
  - project proposal: 3+ pages, due 9/24
  - final report: 12+ pages, due last day of classes
- Almost completely open-ended!
  - All projects must involve natural language data
  - All projects should include at least some amount of model implementation

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.

The Heilmeier Catechism

- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What is new in your approach and why do you think it will be successful?
- Who cares? If you are successful, what difference will it make?
- What are the risks?
- How much will it cost?
- How long will it take?
- What are the mid-term and final “exams” to check for success?

https://en.wikipedia.org/wiki/George_H._Heilmeier#Heilmeier.27s_Catechism

NLP Research

- All the best publications in NLP are open access!
  - Conference proceedings: ACL, EMNLP, NAACL (EACL, LREC...)
  - Journals: TACL, CL
  - "aclweb": ACL Anthology-hosted papers http://aclweb.org/anthology/
  - NLP-related work appears in other journals/conferences too: data mining (KDD), machine learning (ICML, NIPS), AI (AAAI), information retrieval (SIGIR, CIKM), social sciences (Text as Data), etc.
- Reading tips
  - Google Scholar
    - Find papers
    - See paper's number of citations (imperfect but useful correlate of paper quality) and what later papers cite it
    - [... or SemanticScholar...]
  - For topic X: search e.g. [[nlp X]], [[aclweb X]], [[acl X]], [[X research]]...
  - Authors' webpages
    find researchers who are good at writing and whose work you like
  - Misc. NLP research reading tips:
    http://idibon.com/top-nlp-conferences-journals/

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

A few examples

We will post some sample project reports from previous semesters after getting student permission

- Detection tasks
  - Sentiment detection
  - Sarcasm and humor detection
  - Emoticon detection / learning
- Structured linguistic prediction
  - Targeted sentiment analysis (i liked __ but hated __)
  - Relation, event extraction (who did what to whom)
  - Narrative chain extraction
  - Parsing (syntax, semantics, discourse...)
- Text generation tasks
  - Machine translation
  - Document summarization
  - Story generation
  - Text normalization / "style transfer" (e.g. translate online/Twitter text to standardized English)
- End to end systems
  - Question answering
  - Conversational dialogue systems (hard to eval?)
- Predict external things from text
  - Movie revenues based on movie reviews ... or online buzz? http://www.cs.cmu.edu/~ark/movie$-data/
- Visualization and exploration (harder to evaluate)
  - Temporal analysis of events, show on timeline
  - Topic models: cluster and explore documents
- Figure out a task with a cool dataset
  - e.g. Urban Dictionary

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

Tools

- Tagging, parsing, NER, coref, ...
  - Stanford CoreNLP http://nlp.stanford.edu/software/corenlp.shtml
  - spaCy (English-only, no coref) http://spacy.io/
  - Twitter-specific tools (ARK, GATE)
- Many other tools and resources

  tools ... word segmentation ... morph analyzers ...

  resources ... pronunciation dictionaries ... wordnet, word embeddings, word clusters ...
- Long list of NLP resources

  https://medium.com/@joshdotai/a-curated-list-of-speech-and-natural-language-processing-resources-4d89f94c032a
- Deep learning? Try out AllenNLP, PyTorch, Tensorflow (https://allennlp.org, https://pytorch.org/, https://www.tensorflow.org/)

Be on the lookout for

- **HW0:** released today, due Sep 13 (11:59pm) on Gradescope
- Readings on language models for next week
- **Final project:** Organize into groups of 4 by 9/15
- **Final project:** project proposal due 9/24

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

demos!

(allennlp.org)

demos!

(https://beta.openai.com/playground)