CMSC 723: NLP

- Gradescope for HW
- course email: cmsc723instructors@gmail.com

- Piazza for QA
- course website:
  - cs.umd.edu/~miyyer/cmsc723

TAs + office hours

- Jenna
- Rishabh
- Mohit

Thurs 4-5 pm } AVW 4424
Mon 12-1pm }

Tues 11am-12pm IRB 4142

Other logistics

- no textbooks, all readings on website
- Prev. class (S2024) on YouTube

Preregs:

- basic ML
- programming (Python)
- math: linear algebra, probability, calculus

Grading:

NO AI:

- 25% exam 1
- 25% exam 2

} in-class, mix of multiple choice and free response

AI allowed:

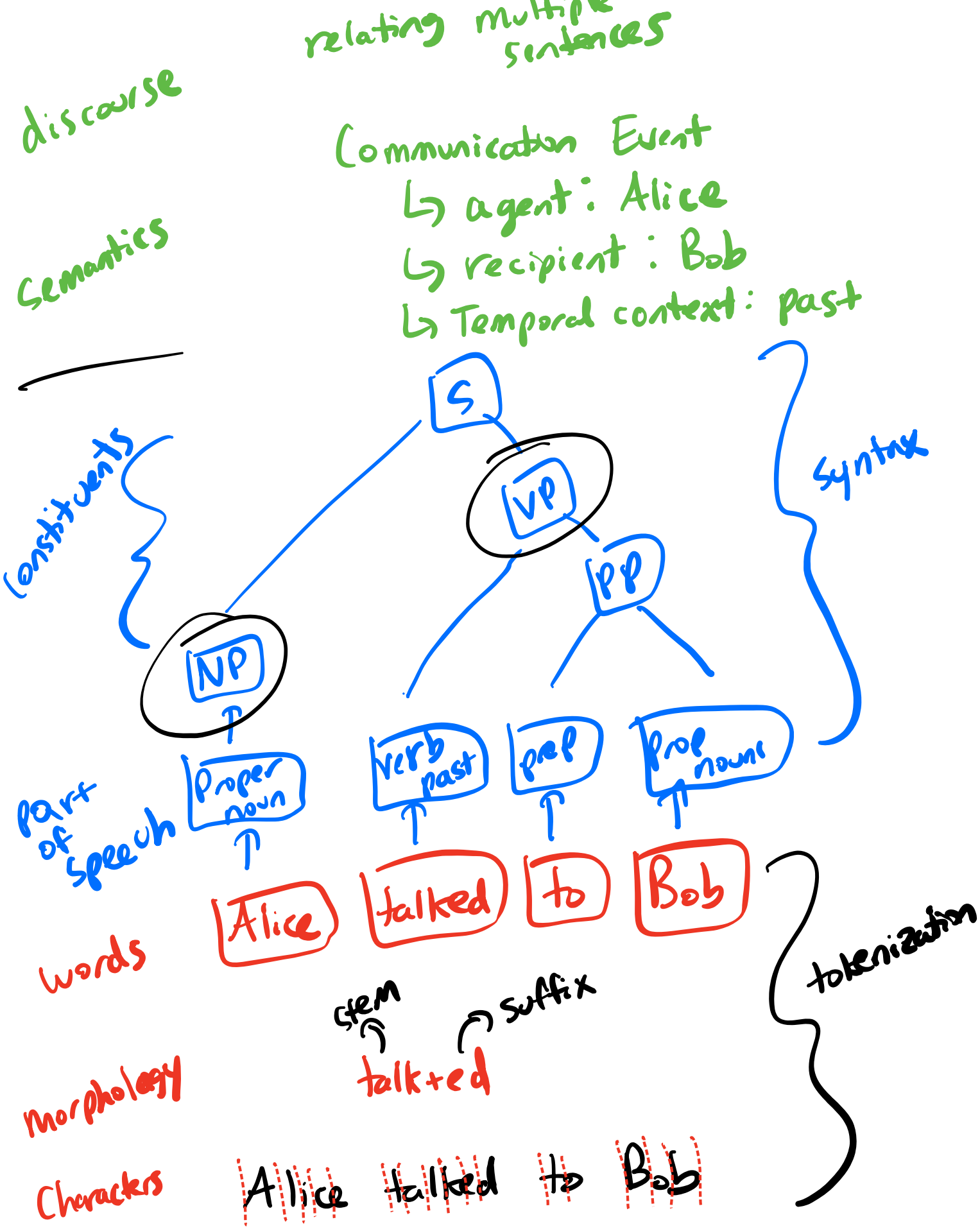
- 20% problem sets
- 30% final project
  - 5% proposal
  - 25% final report
    - groups of 3-4 students
      - choose any NLP-related project

natural language processing

languages that evolved thru human use
→ we will also consider code

↳ supervised learning
→ map text to X
↳ unsup learning
→ learn X from text
↳ generation of text
↳ reinforcement learning

Most of these applications are performed today w/ large language models (LLMs)


relating multiple sentences
discourse
Communication Event
↳ agent: Alice
↳ recipient: Bob
↳ Temporal context: past
Semantics
S
VP
PP
NP
constituents
syntax
Proper noun
Verb past
prep
Pronoun
part of speech
Alice
talked
to
Bob
words
stem
suffix
talk+ed
morphology
tokenization
Characters
Alice talked to Bob

language modeling:

- given the beginning of a sentence/doc (prefix/prompt), predict the next word
- "self-supervised" learning
  - just from raw text (internet/books/repos) we can create trillions of training examples
- common paradigms:
  1. **pretraining**: train from scratch on trillions of words
     - OpenAI, Anth, Google, DeepSeek, ...
  2. **finetuning**: specializing the model on a set of tasks
  3. **prompt** model to solve new tasks