CS 490: NATURAL LANGUAGE PROCESSING

Dan Goldwasser, Abulhair Saparov

Spring 2026

WHAT IS NLP?

- Build algorithms that work with natural language.
- How do these algorithms work?
- Why do they sometimes work well?
- Why do they sometimes not work?
- This course is meant to teach about the fundamentals of NLP,
  - And about the tools/methods that practitioners use to solve NLP problems.

NLP TASKS

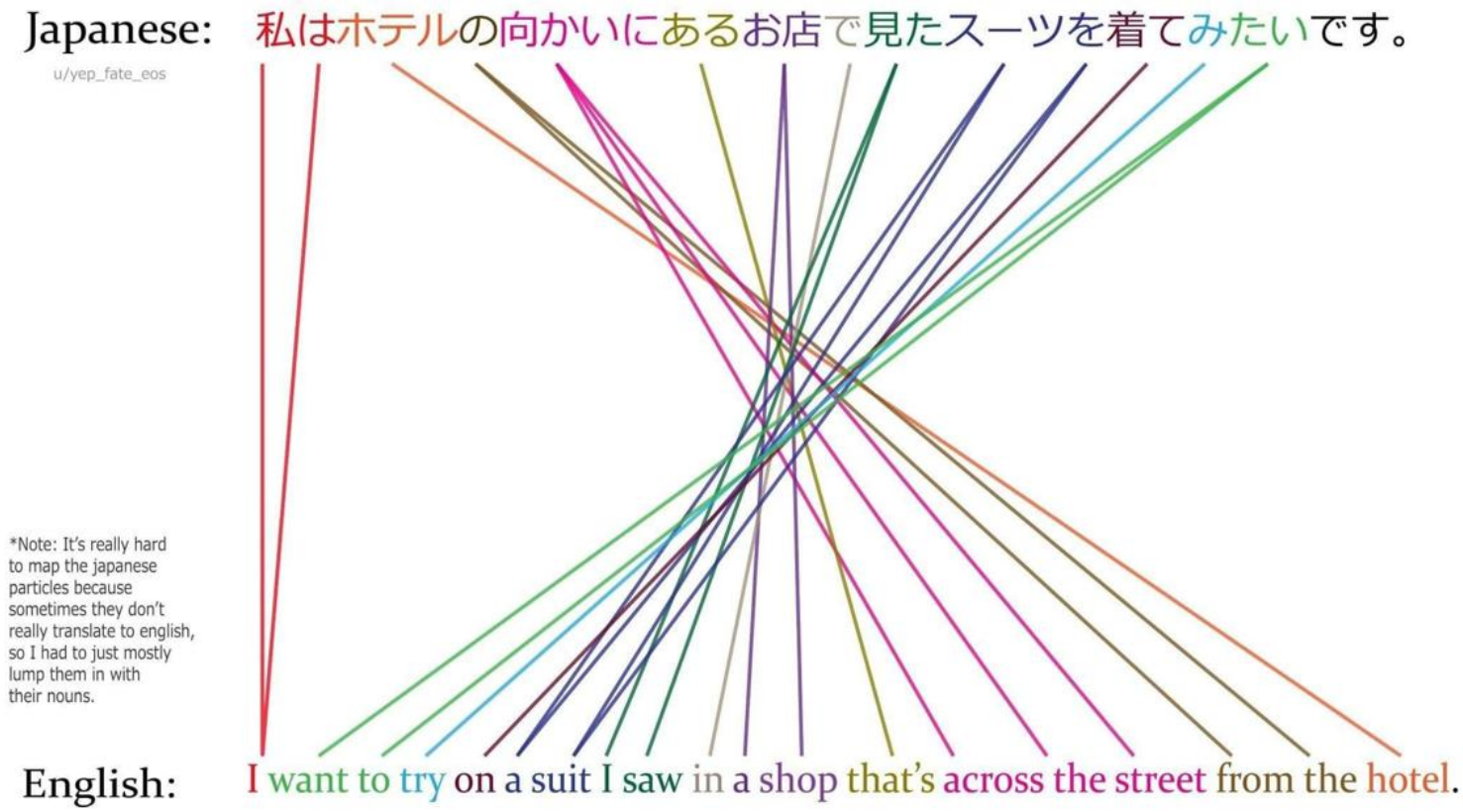
- Machine translation

Input: The quick brown fox jumped over the lazy dog.

Output: 素早い茶色のキツネは怠け者の犬を飛び越えました。

NLP TASKS

- Machine translation



[source: Reddit u/yep_fate_eos]

NLP TASKS

- Machine translation
- Named entity recognition

Input: Honda

Output: company

NLP TASKS

- Machine translation
- Named entity recognition

Input: Noam Chomsky

Output: person

NLP TASKS

- Machine translation
- Named entity recognition

Input: Apple

Output: company

NLP TASKS

- Machine translation
- Named entity recognition

Input: apple

Output: not named entity

NLP TASKS

- Machine translation
- Named entity recognition
- Coreference resolution

Input: The souvenir didn't fit into the suitcase because it was too big.

Output: "it" = "souvenir"

NLP TASKS

- Machine translation
- Named entity recognition
- Coreference resolution

Input: The souvenir didn't fit into the suitcase because it was too small.

Output: "it" = "suitcase"

NLP TASKS

- Machine translation
- Named entity recognition
- Coreference resolution
- Question answering

Input: You are in the middle of a circular lake. You can swim at 1 m/s. A dog is at the edge of the lake. The dog can run on land at x m/s, but cannot swim, and you can run faster. What is the highest value for x such that you can still escape?

Output: 4.6033

Reasoning is needed to solve this task.

NLP TASKS

- Machine translation
- Named entity recognition
- Coreference resolution
- Question answering
- Image description

Input:

Output: This image features a framed painting of Purdue University. The artwork is displayed on a white marble wall...

Multi-modal NLP includes the study of tasks involving other modalities, such as vision, sound, speech, motion, etc.

NLP AS MACHINE LEARNING

- It is infeasible to write a function to solve these tasks directly.
- So we rely on machine learning to learn this function from data.
  - We use a dataset containing many input-output examples.
  - We train a machine learning model predict the output from the input.
- The specific choice of model and training regimen is the "method."

LANGUAGE IS AMBIGUOUS

"Teachers strike idle kids."

- Interpretation 1: Teachers physically strike kids who are idle.
- Interpretation 2: The teacher's strike is causing the kids to be idle.

LANGUAGE IS AMBIGUOUS

"Time flies like an arrow."

- Interpretation 1: Time moves forward similar to how an arrow flies.
- Interpretation 2: This is a command, telling you to measure the speed of the flies similar to how you would measure the speed of an arrow.
- Interpretation 3: Also a command, telling you to measure the speed of the flies, but in a manner similar to how arrows would measure the speed of the flies.
- Interpretation 4: There are things called "time flies" and they like an arrow.
- We will cover probabilistic methods that handle ambiguity.

[source: https://en.wikipedia.org/wiki/Time_flies_like_an_arrow;_fruit_flies_like_a_banana]

NLP ≠ LANGUAGE MODELS

- Language modeling is a task in NLP.

Input: The quick brown fox jumped over the lazy

Output: dog

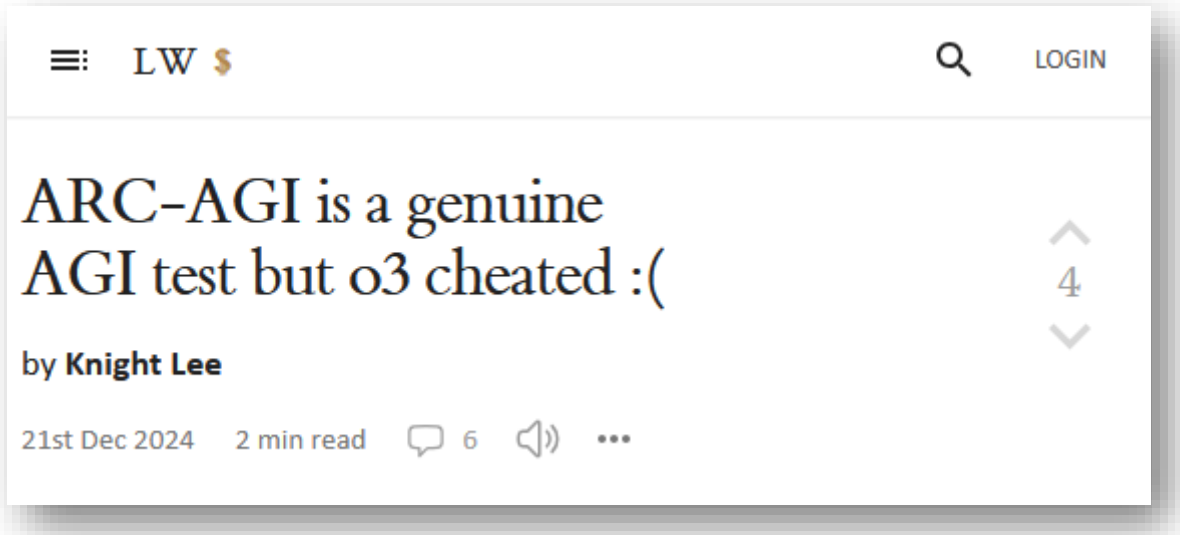
- These days, “language models” almost exclusively refers to large-scale machine learning models that are trained on the language modeling task.

NLP ≠ LANGUAGE MODELS

- It seems as though large language models (LLMs) have “taken over” NLP.
  - And we will discuss how they work.
- Language modeling task has a nice property:
  - Many (all?) other NLP tasks can be written as a language modeling task.
  - So if you train a good language model, you train it to perform many NLP tasks simultaneously.
  - Valid question: Can LLMs “solve” all NLP tasks?
- But this property is not unique to language modeling.
  - E.g., many NLP tasks can also be phrased as question-answering.

NLP ≠ LANGUAGE MODELS

NLP ≠ LANGUAGE MODELS



- LLMs are making evaluation very difficult.
- We will discuss how NLP models are evaluated.
  - And how their evaluation differs from before LLMs.

NLP ≠ DEEP LEARNING

- Modern approaches in NLP rely heavily on deep learning methods.
- But NLP is “method-agnostic”:
  - Many different kinds of methods can be used to solve various problems in NLP.
- NLP is not only the study of the methods for solving natural language tasks.
- NLP includes the modeling of natural language itself:
  - What is language?
  - How can we describe it?
- Studying the nature of language itself will help us build better models and implement better methods to solve NLP tasks.

LANGUAGE HAS STRUCTURE

- Language is more than just a sequence of words.
- There is recursive structure:
  - “Fae sees Alex.”
  - “Fae sees the person sitting under the tree.”
  - “Fae sees the person sitting under the tree that had been planted 10 years ago.”
  - etc...
- There is structure *within* words too:
  - “Recalculating” -> “re”- “calculate”- “ing”
  - “Sleeplessness” -> “sleep”- “less”- “ness”
  - Other languages have much more complex morphologies.

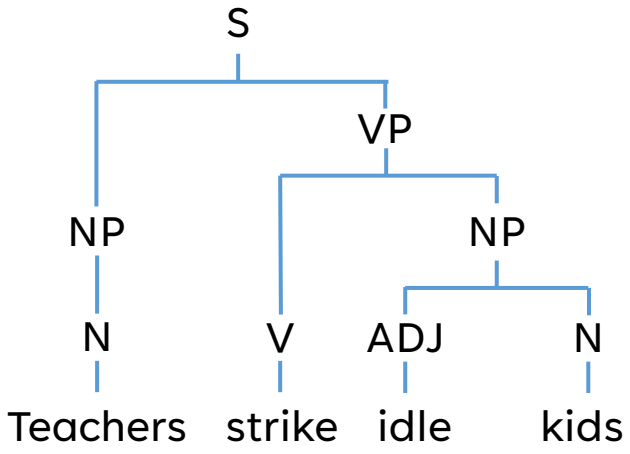
LANGUAGE HAS STRUCTURE

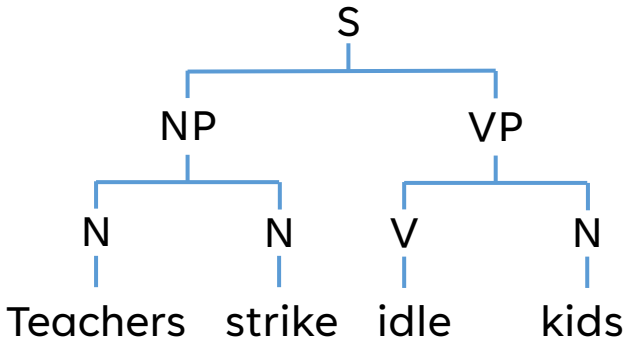
- Most NLP tasks require understanding the structure of language.
- How can we evaluate how well an NLP method performs on a task (or build better methods) if we don't understand the structure underlying the task?
  - E.g., I can train a model to land a rocket on the moon by having it try many attempts.
  - The model will try different rocket shapes, fuels, maneuvers, etc.
  - But if I understand Newtonian gravity, I can build a better model/rocket.
- Understanding linguistic theory is similarly important.
- We will cover foundational concepts of linguistic theory, such as morphology, syntax, semantics, pragmatics, etc.

LANGUAGE HAS STRUCTURE

"Teachers strike idle kids."

- Interpretation 1: Teachers physically strike kids who are idle.
- Interpretation 2: The teacher's strike is causing the kids to be idle.





COURSE OUTLINE

The course is divided into three modules.

- **Module 1**: NLP fundamentals
- Tasks: Text classification
  - E.g., spam detection, author identification, document classification
- Methods: Perceptron, logistic regression, neural networks
  - How to train NLP models
  - Optimization, gradient descent, regularization
- Representation learning, word embeddings
- Attention and transformers

COURSE OUTLINE

- **Module 2**: NLP foundations
- Morphology and lexical semantics
  - Lexical relations, tokenization, byte-pair encoding
- Syntax
  - Context-free grammars (CFGs), dependency grammars, parsing
- Semantics
  - Compositional semantics, lexical semantics
  - Reasoning, natural language understanding, machine comprehension
- Discourse and pragmatics
  - Conversational NLP

MORPHOLOGY

- Morphology is the study of how words are constructed from smaller components.
  - E.g., verb conjugation: “I walk,” “she walks,” “We walked yesterday,” …
  - “I sit,” “she sits,” “We sat,” …
  - “I am,” “she is,” “We were,” …
- Simply adding/deleting endings is not sufficient:
  - “gorge” vs “gorgeous”
  - “good” vs “goods”
  - “arm” vs “army”

SYNTAX

- Syntax describes the structural relationship between words in a sentence.

"Sally caught the butterfly with a net."

vs

"Sally caught the butterfly with a spot."

- However, syntax is not enough to capture the meaning of the sentence.

SEMANTICS

- Semantics describes the meaning of words, phrases, and sentences.
- Compositional semantics describes how the meaning of smaller phrases combines to form the meaning of larger phrases:
  - “Sally caught the butterfly” + “with a net”
- Meaning can be represented in a formal language, such as logic, programming languages, or math.
  - “Mary gave 10 apples to Bob.”

    can be semantically-parsed into:

    ```
    bob['apple'] += 10
    mary['apple'] -= 10
    ```

COURSE OUTLINE

- **Module 3**: Large language models
- Classical language models, transformer language models
- Scaling laws
- Prompting
  - Few-shot prompting, in-context learning, chain-of-thought prompting
- Retrieval-augmented generation, LLM agents
- Fine-tuning, model compression
- Reinforcement learning
  - RLHF, RLVR
- Multi-modal NLP

COURSE LOGISTICS

COURSE EXPECTATIONS

- Evaluation:
  - Four assignments (30%)
    - First assignment will check whether you have sufficient technical background
  - Final project (30%)
  - Final exam (40%)

GUIDELINES

- Working in groups
  - We encourage you to work in groups on the assignments and final project.
  - Groups should have 2-4 people.
  - You are free to collaborate.
  - But to state the obvious: No cheating or plagiarism
  - You can discuss homeworks with others but must write up your own solution.

- Late policy: 5 late days total
  - We strongly recommend you start assignments early.

- Attend office hours to seek guidance, and to discuss homeworks and projects.

USE OF GENERATIVE AI

- If you find generative AI useful (e.g., ChatGPT), you are permitted to use it.
- However, do not simply copy the output of AI into your assignments.
- You should write your own solutions.
- When coding, you may use AI to generate snippets of code (e.g., boilerplate).
  - But be wary of over-relying on/putting too much trust in the AI.
  - We will design assignments that are not as easily solved by current AI models.
- AI will not help you on the final exam.
  - Relying too much on AI for assignments will hinder your preparation for the exam.

ONLINE DISCUSSION

- We will be using Ed Discussion as the online platform for discussion.
- Join the discussion forum using the following link:

https://edstem.org/us/join/SnZKcE

- If you have any questions, please make a post there!
- I will announce this link on Brightspace.
- If you are not registered on Brightspace, send me an email and I will add you.

FINAL PROJECT

- Find a topic you care about!
  - Something you always wanted to build.
  - E.g., applications of NLP to other domains.
- Key points:
  - Identify a language-related problem and define it precisely.
  - Interesting approach in tackling the problem
    - We will cover several different kinds of methods
    - You will have to choose the methods and justify your choice
  - What *not* to do: avoid generic problems and generic solutions
    - Apply LLM with chain-of-thought is not novel or interesting

FINAL PROJECT

- Proposal:
  - Define the problem you aim to solve/answer
  - Basic intuitions and proposed method
  - Describe datasets (if applicable)
  - No more than 5 pages!
- Final report: due at end of class
  - Short report describing your findings
  - Presentations? (depending on class size and time availability)

FINAL PROJECT IDEAS

- Question-answering
  - Maybe in a new domain, where domain-specific information can be exploited.
- Multi-modal NLP
  - Combine NLP and vision or robotics
- Analysis or rigorous evaluation of NLP models
  - Including LLMs
- Conversational agent to solve a problem in a specific domain (e.g., medicine, law, finance)
- Metaphor, poetry, humor, non-literal language
- Your own application

QUESTIONS?