CS11-711 Advanced NLP

Introduction to Natural Language Processing

Daniel Fried and Robert Frederking
with slides from Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Site

https://cmu-anlp.github.io/

What is NLP Anyway?

- Technology to handle human language (usually text) using computers

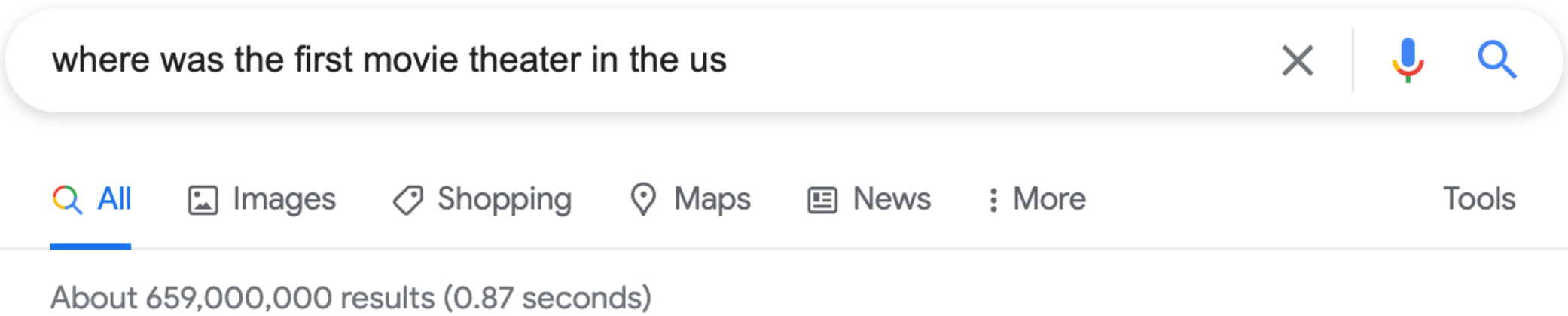
- Aid **human-human communication** (e.g. machine translation)

- Aid **human-machine communication** (e.g. question answering, dialog)

- **Analyze/understand language** (syntactic analysis, text classification, entity/relation recognition/linking)

- We now use NLP several times a day, sometimes without knowing it!

NLP can Answer our Questions



Retrieved Aug. 29, 2021

NLP can Translate Text

Aus den „Nägeln mit Köpfen“, welche die Grünenpolitikerin Katrin Habenschaden noch im Dezember 2022 im Münchner Stadtrat machen wollte, ist nichts geworden. Und das, obwohl die Zweite Bürgermeisterin eigentlich richtig Tempo auf dem Max-Joseph-Platz machen will, ist doch der repräsentative Platz im Herzen der Altstadt in den Fokus ihres Umgestaltungswillens gerückt.

→

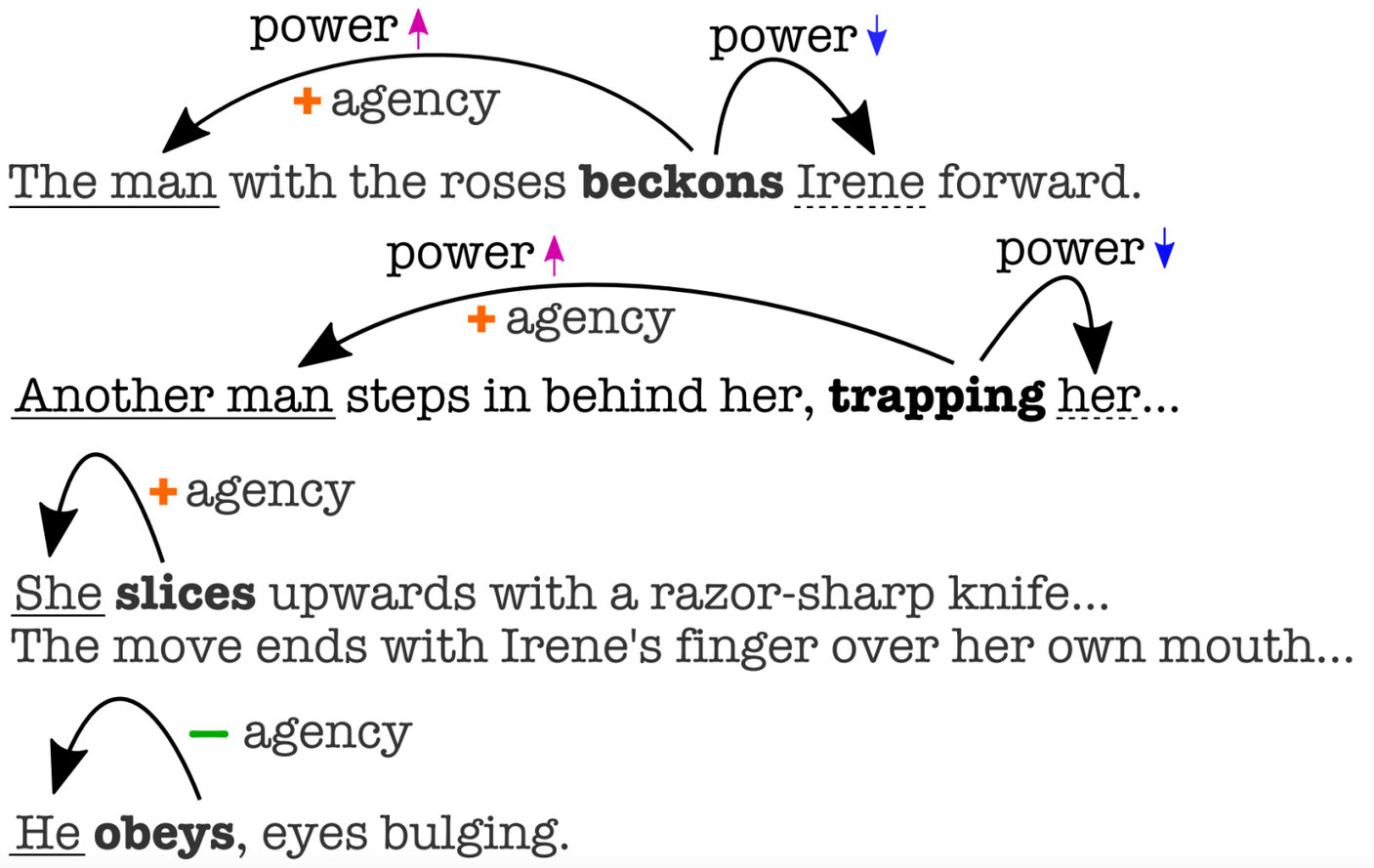
Nothing has come of the “nails with heads” that the Green Party politician Katrin Habenschaden wanted to do in December 2022 in the Munich City Council. And that, although the deputy mayoress really wants to speed things up on Max-Joseph-Platz, the representative square in the heart of the old town has become the focus of her will to redesign.

Front page news from Frankfurter Allgemeine, translated by Google Aug 20, 2023

NLP can Aid Scientific Inquiry

- e.g. *computational social science*, answering questions about society given observational data

- example: "do movie scripts portray female or male characters with more power or agency?" [Sap+ 2017]



Frame	β	gender
$agency(\text{AG})=+$	-0.951	**M****
$power(\text{AG}>\text{TH})$	-0.468	**M****
$agency(\text{AG})=-$	0.277	**F****
$power(\text{AG}<\text{TH})$	*not sig.*	

Sap et al. "Connotation Frames of Power and Agency in Modern Films" EMNLP 2017.

NLP can Answer our Questions

1 / 2

Who won the 2021 Pittsburgh mayoral democratic primary?

The 2021 Pittsburgh mayoral Democratic primary was won by Ed Gainey. He defeated incumbent mayor Bill Peduto in the primary election. Gainey went on to win the general election and became the mayor of Pittsburgh. Please note that my knowledge is accurate up until September 2021, so I recommend verifying this information with more recent sources to ensure its accuracy.

Checked Aug. 20, 2023

NLP can (sometimes) Answer our Questions

‹ 2 / 2 ›

Who won the 2019 Pittsburgh mayoral democratic primary?

The 2019 Pittsburgh mayoral Democratic primary was won by Bill Peduto. He secured the Democratic nomination for the mayoral race in Pittsburgh, Pennsylvania. Bill Peduto was the incumbent mayor at the time and went on to win the general election in November 2019 to serve a second term as mayor.

The **2017 Pittsburgh mayoral election** took place on November 7, 2017. The primary election was held on May 16, 2017. Incumbent Democratic Mayor Bill Peduto successfully ran for re-election to a second term.[1][2]

Incumbent Ed Gainey since January 3, 2022	
Style	"The Honorable"
Term length	4 years

60	Bill Peduto ♥	2014–2022	Democrat	Former city councilman	Joshua Wander Unopposed

Checked Aug. 20, 2023

NLP cannot Translate Text

بەڵام 3 توێژەر بە سەرۆکایەتی زانای سەربەخۆی بەبەردبوو گریگۆری پاولّ لە شاری بالتیمۆر لە ویلایەتی میریلاند لە مانگی 3ی ٢٠٢٢دا ئاماژەیان بەوە کرد کە پێویستە وەک سێ T. rex سالّی جۆر بناسرێت.

However, three researchers, led by independent fossil scientist Gregory Paul of Baltimore, Maryland, argued in March 2022 that T. rex should be recognized as three species.

جگە لە جۆری T. rex، کە بە واتای "پاشای مارمێلکەی درِندە" دێت، سەرباری ئەوە 2 جۆری تریان پێشنیار کرد.

In addition to the T. rex species, which means "king of ferocious lizards", they also proposed two other species.

T. imperator بە واتای "ئیمپراتۆری مارمێلکەی درِندە دێت

T. imperator means "emperor of the savage lizard

T. regina بە واتای "شاژنی مارمێلکەی درِندە.

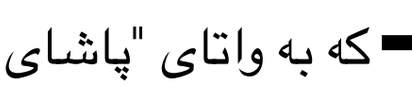
T. regina means "Queen of the ferocious snail.

Front page news from Voice of America Kurdish, translated by Google Aug 20., 2023

NLP cannot Translate Text

بەڵام 3 توێژەر بە سەرۆکایەتی زانای سەربەخۆی بەبەردبوو گریگۆری پاولْ لە شاری بالتیمۆر لە ویلایەتی میریلاند لە مانگی 3ی وەک سێ T. rex ساڵی ٢٠٢٢دا ئاماژەیان بەوە کرد کە پێویستە جۆر بناسرێت.

However, on three occasions, in the spring of 2022, the leadership of paleontological knowledge was shaken in the city of Baltimore, Maryland, when Gregory Paul presented their homage, stating that T. rex should be classified into three species.



کە بە واتای "پاشای مارمێلکەی درِندە" دێت، T. rex جگە لە جۆری سەرباری ئەوە 2 جۆری تریان پێشنیار کرد.

According to the classification of T. rex based on "chewing marbles," their proposal suggests two additional species.

بە واتای "ئیمپراتۆری مارمێلکەی درِندە دێت T. imperator

T. imperator, meaning "emperor of chewing marbles."

بە واتای "شاژنی مارمێلکەی درِندە T. regina.

T. regina, meaning "queen of chewing marbles."

Front page news from Voice of America Kurdish, translated by ChatGPT Aug 20., 2023

NLP Fails at Even Basic Tasks

First sentence of first article in NY Times Aug 29., 2021, recognized by Stanford CoreNLP

Hurricane (CAUSE_OF_DEATH) Ida battered Louisiana (STATE_OR_PROVINCE) on Sunday (DATE 2021-08-29) making landfall as a Category 4 (NUMBER 4.0) storm (CAUSE_OF_DEATH), delivering an onslaught of harsh winds , floodwaters and power outages and threatening to assail Baton Rouge (ORGANIZATION) and New Orleans (CITY) as one (NUMBER 1.0) of the most devastating storms (CAUSE_OF_DEATH) to strike the region since Hurricane (CAUSE_OF_DEATH) Katrina .

recognized by spaCy

Hurricane Ida ORG battered Louisiana GPE on Sunday DATE making landfall as a Category 4 storm, delivering an onslaught of harsh winds, floodwaters and power outages and threatening to assail Baton Rouge GPE and New Orleans GPE as one of the most devastating storms to strike the region since Hurricane Katrina.

In this Class, we Ask:

- Why do current state-of-the-art NLP systems **work uncannily well** sometimes?
- Why do current state-of-the-art NLP systems still **fail**?
- How can we
  - **create systems for various tasks**,
  - **identify their strengths and weaknesses**,
  - **make appropriate improvements**,
  - and **achieve whatever we want to do with NLP**?

NLP System Building

Overview

A General Framework for NLP Systems

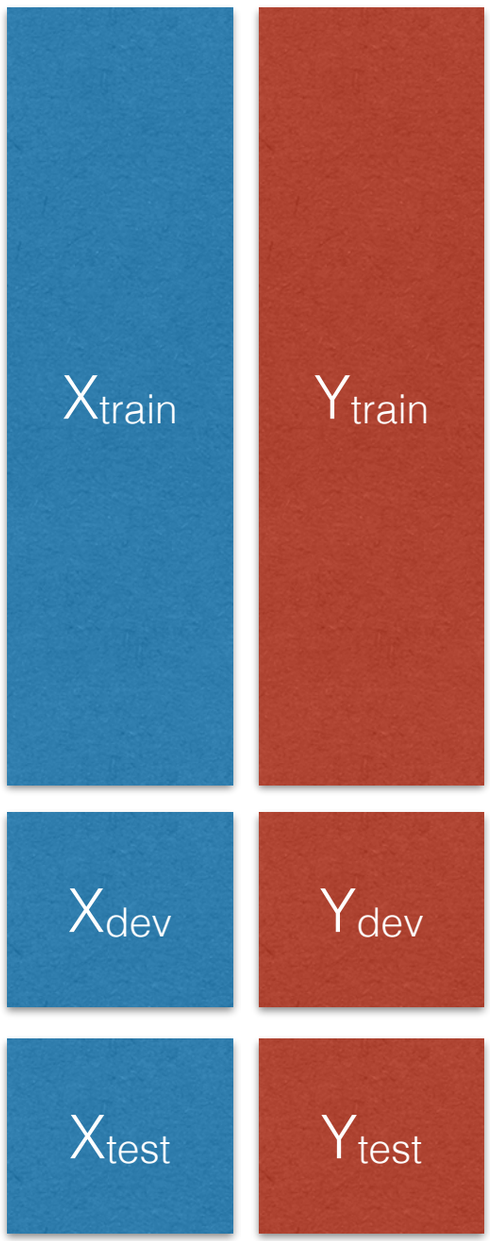
- Formally, create a function to map an input *X (usually language)* into an output *Y*. Examples:

Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Text	Label	Text Classification
Text	Linguistic Structure	Language Analysis

- To create such a system, we can use
  - Manual creation of rules
  - Machine learning from paired data <*X*, *Y*>

Train, Development, Test

- When creating a system, use three sets of data



Training Set: Generally larger dataset, used during system design, creation, and learning of parameters.

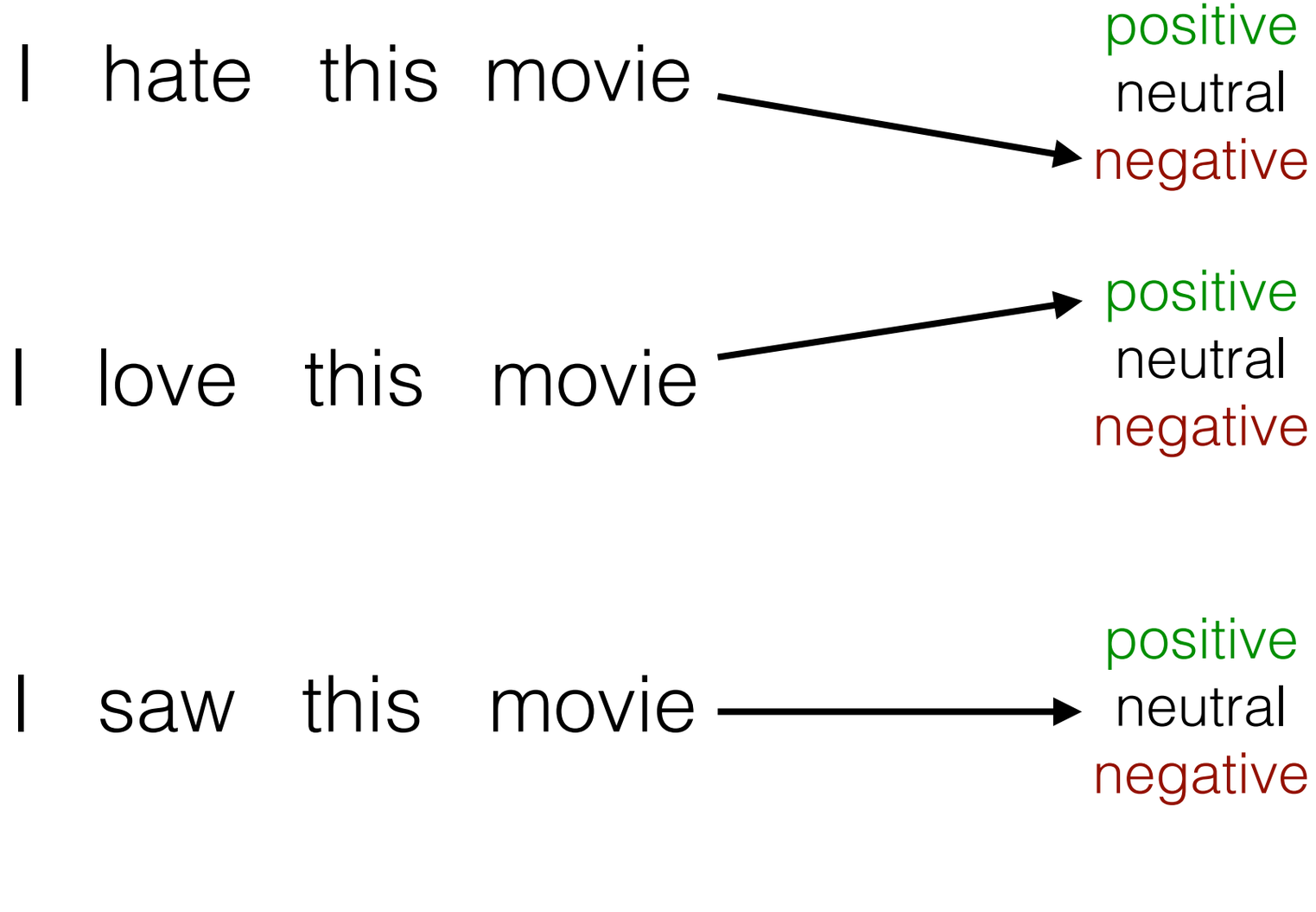
Development ("dev", "validation") Set: Smaller dataset for testing different design decisions ("hyper-parameters").

Test Set: Dataset reflecting the final test scenario, do not use for making design decisions.

Let's Make a Rule-based NLP System!

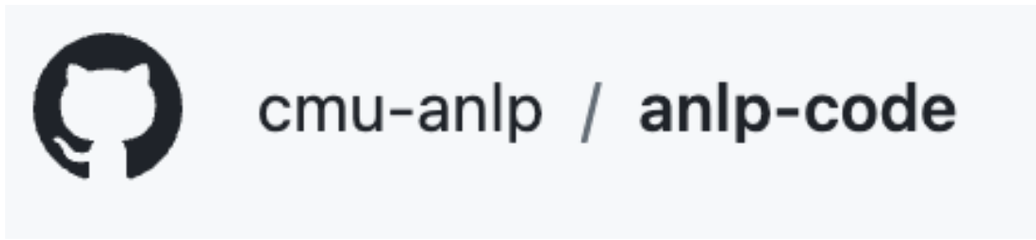
Example Task: Review Sentiment Analysis

- Given a review on a reviewing web site (*X*), decide whether its label (*Y*) is positive (1), negative (-1) or neutral (0)



Let's Look at Data

https://github.com/cmu-anlp/anlp-code



data/sst-sentiment-text-threeclass

- Remember: look at "train", not "dev" or "test"

A Three-step Process for Making Predictions

- **Feature extraction:** Extract the salient features for making the decision from text

- **Score calculation:** Calculate a score for one or more possibilities

- **Decision function:** Choose one of the several possibilities

Formally

- **Feature Extraction:** $\mathbf{h} = f(\mathbf{x})$
- **Score Calculation:** binary, multi-class

$$s = \mathbf{w} \cdot \mathbf{h} \qquad \mathbf{s} = W\mathbf{h}$$

- **Decision:** $\hat{y} = \text{decide}(\mathbf{s})$

Sentiment Classification Code Walk!

https://github.com/cmu-anlp/anlp-code/tree/main/01-rulebasedclassifier

- See code for all major steps:
  1. Featurization
  2. Scoring
  3. Decision rule
  4. Accuracy calculation
  5. Error analysis

Now Let's Improve!

1. What's going wrong with my system?
   → Look at error analysis

2. Modify the model (featurization or scoring function)

3. Measure accuracy improvements, accept/reject change

4. Repeat from 1

5. Finally, when satisfied with train/dev accuracy, evaluate on test!

Some Difficult Cases

Low-frequency Words

The action switches between past and present , but the material link is too **tenuous** to anchor the emotional connections that **purport** to span a 125-year divide .

negative

Here 's yet another studio horror franchise **mucking** up its storyline with **glitches** casual fans could correct in their sleep .

negative

Solution?: Keep working till we get all of them? Incorporate external resources such as sentiment dictionaries?

Conjugation

An operatic , sprawling picture that 's **entertainingly** acted , **magnificently** shot and gripping enough to sustain most of its 170-minute length .

positive

It 's basically an **overlong** episode of Tales from the Crypt .

negative

Solution?: Use the root form and POS of word?

Note: Would require morphological analysis.

Negation

This one is not nearly as dreadful as expected .

positive

Serving Sara does n't serve up a whole lot of laughs .

negative

Solution?: If a negation modifies a word, disregard it.

Note: Would probably need to do syntactic analysis.

Metaphor, Analogy

Puts a human face on a land most Westerners are unfamiliar with.

positive

Green might want to hang onto that ski mask , as robbery may be the only way to pay for his next project .

negative

Has all the depth of a wading pool .

negative

Solution?: ???

Other Languages

見事に視聴者の心を掴む作品でした。

positive

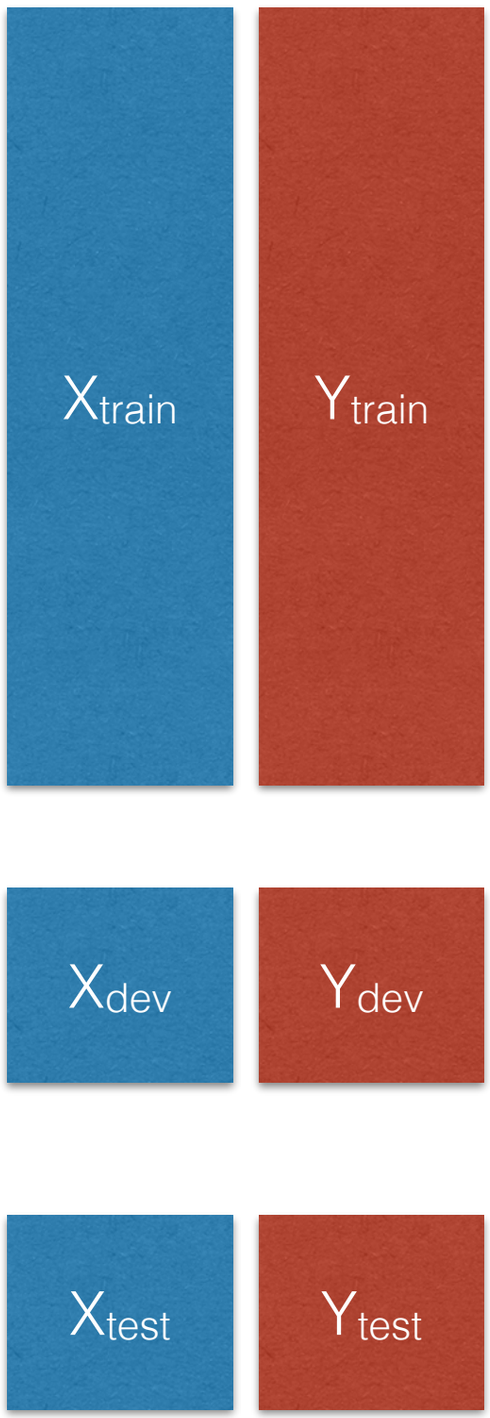
モンハンの名前がついてるからとりあえずモンハン要素をちょこちょこ入れればいいだろ感が凄い。

negative

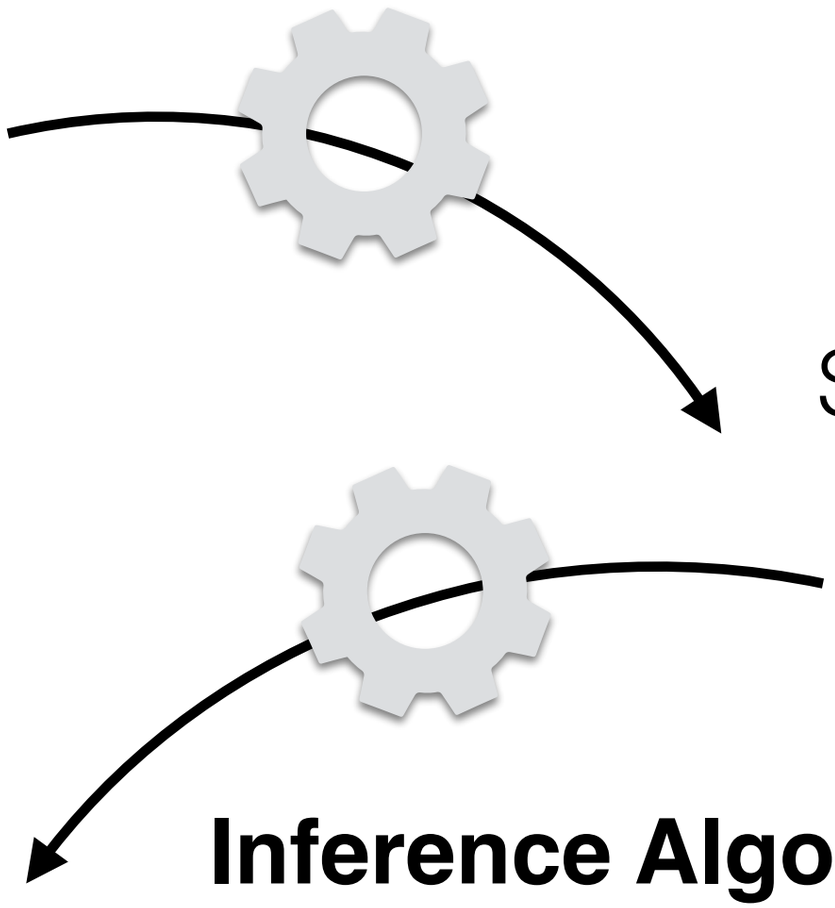
Solution?: Learn Japanese?

Machine Learning Based NLP

Machine Learning



Learning Algorithm



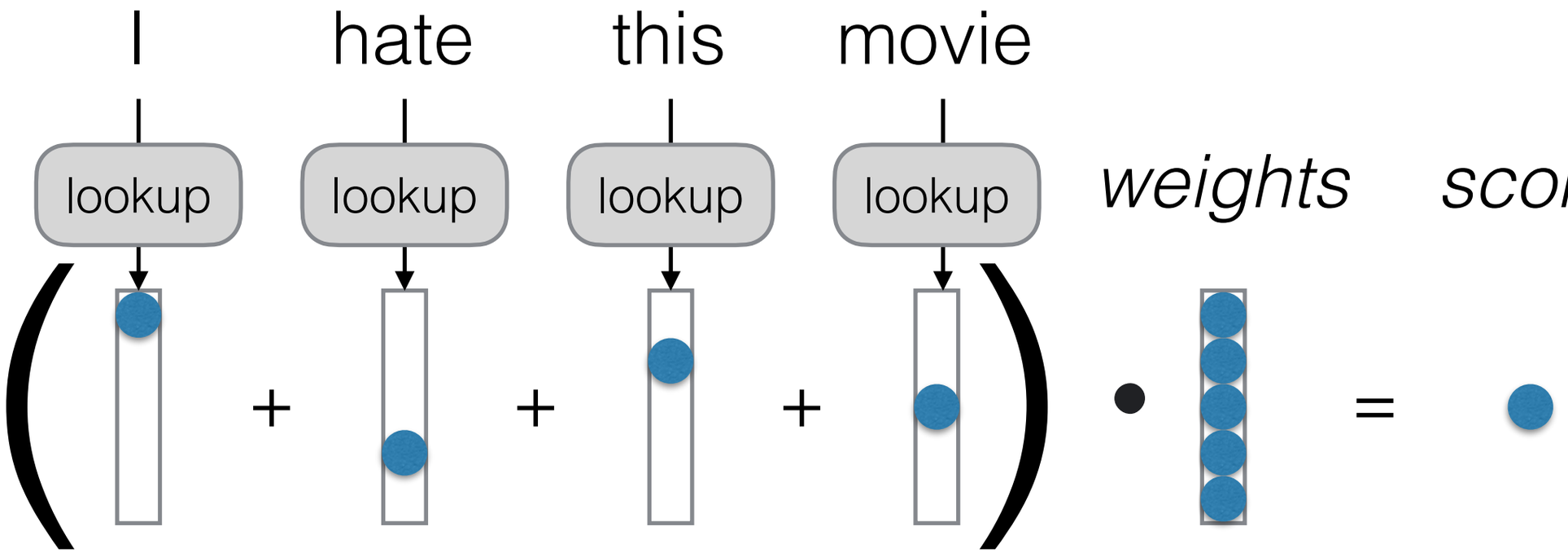
Learned
Feature Extractor *f*
Scoring Function *w*

$$\mathbf{h} = f(\mathbf{x})$$

$$s = \mathbf{w} \cdot \mathbf{h}$$

Inference Algorithm

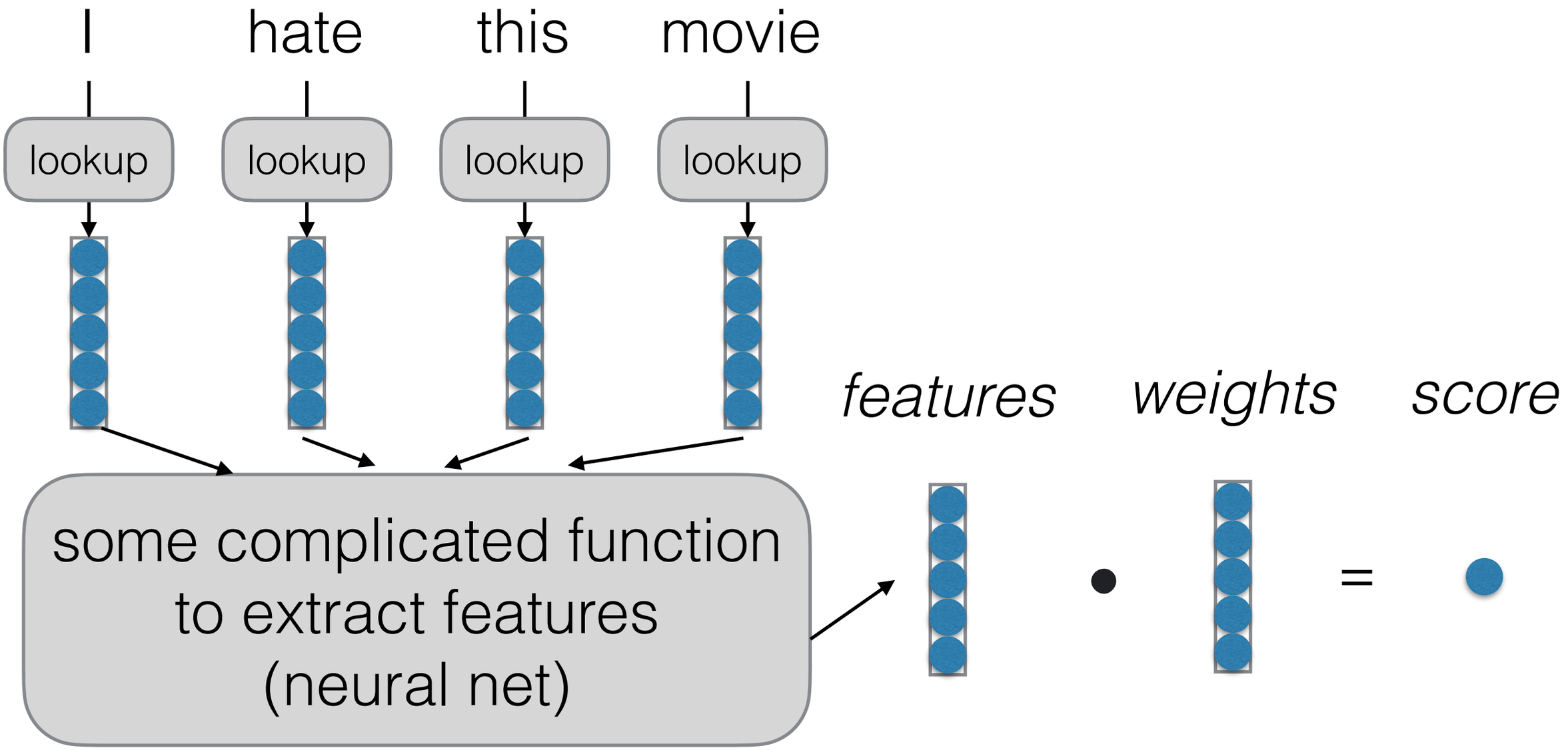
A First Attempt: Bag of Words (BOW)



Features *f* are based on word identity, weights *w* learned

Which problems mentioned before would this solve?

A Better Attempt: Neural Network Models

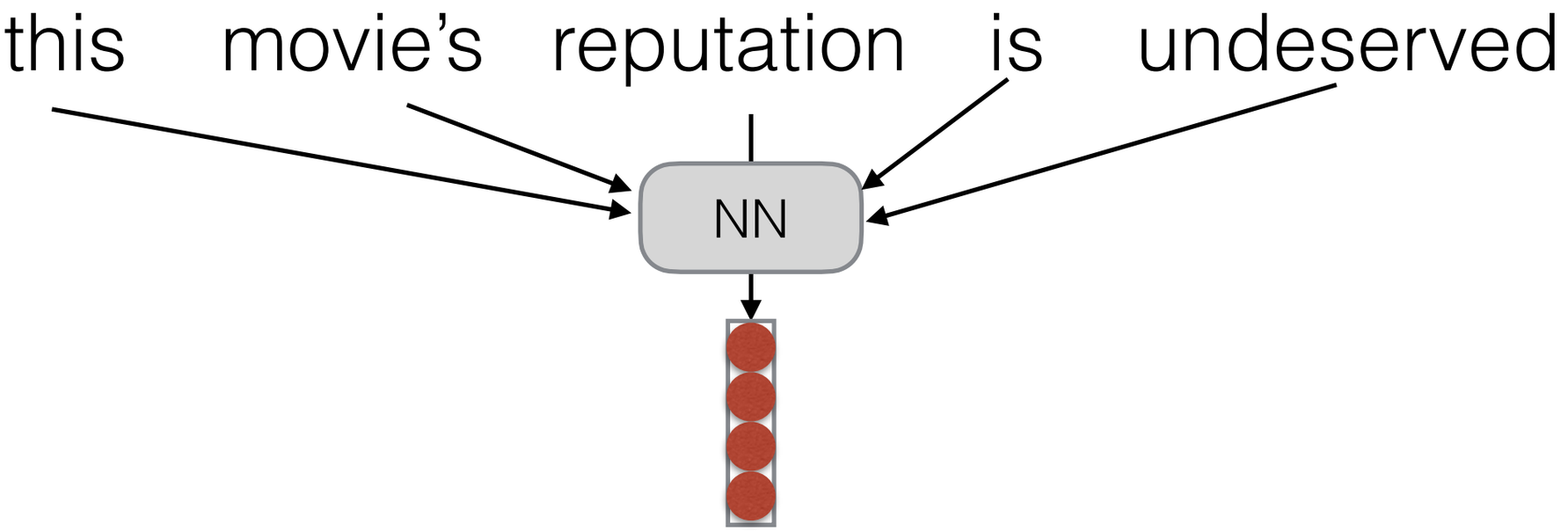


Class Goals

- Learn in detail about **building NLP systems from a research perspective**
- Learn basic and advanced topics in **machine learning and neural network approaches** to NLP
- Learn **basic linguistic knowledge** useful in NLP, and learn methods to **analyze linguistic structure**
- See several case studies of **NLP applications** and learn how to identify unique problems for each
- Learn how to debug **when and where NLP systems fail**, and build improvements based on this

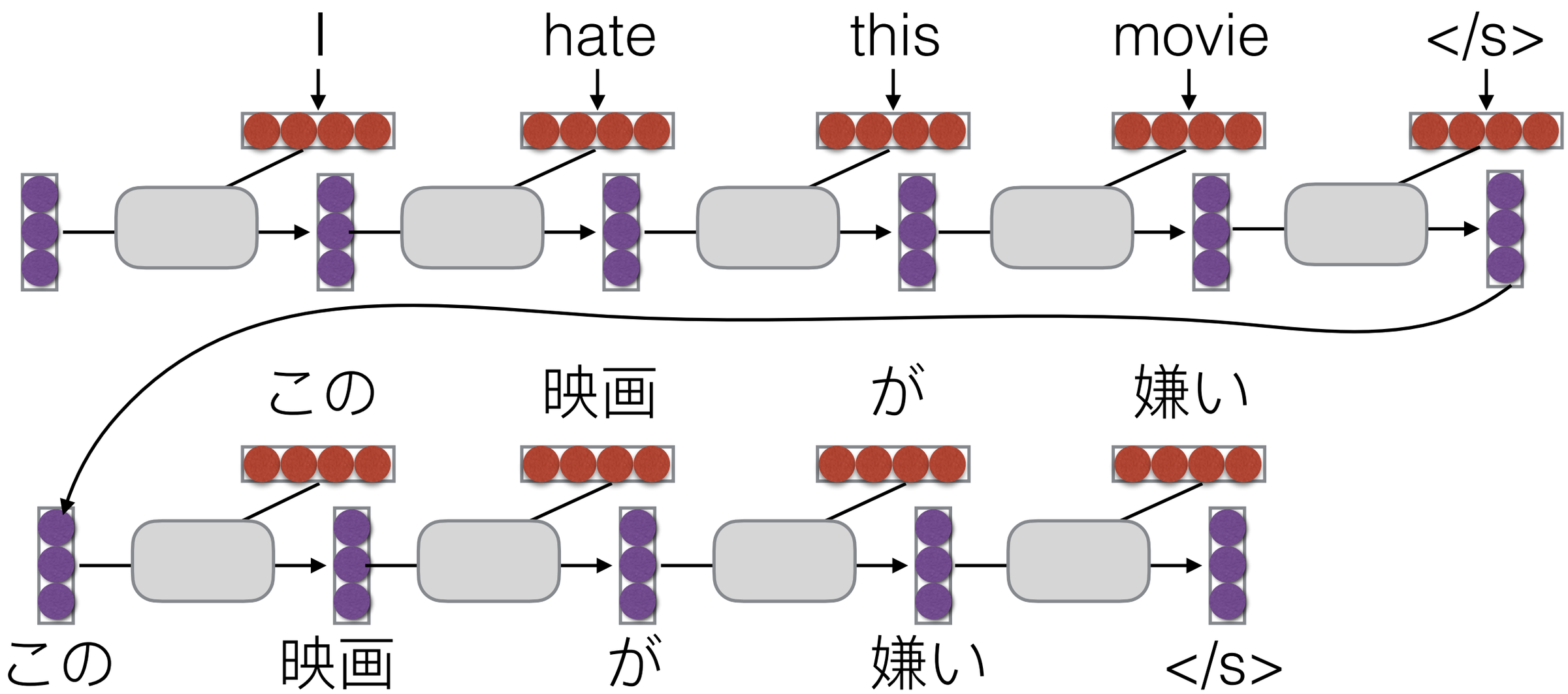
Roadmap Going Forward

Topic 1: Machine Learning and Neural Net Fundamentals



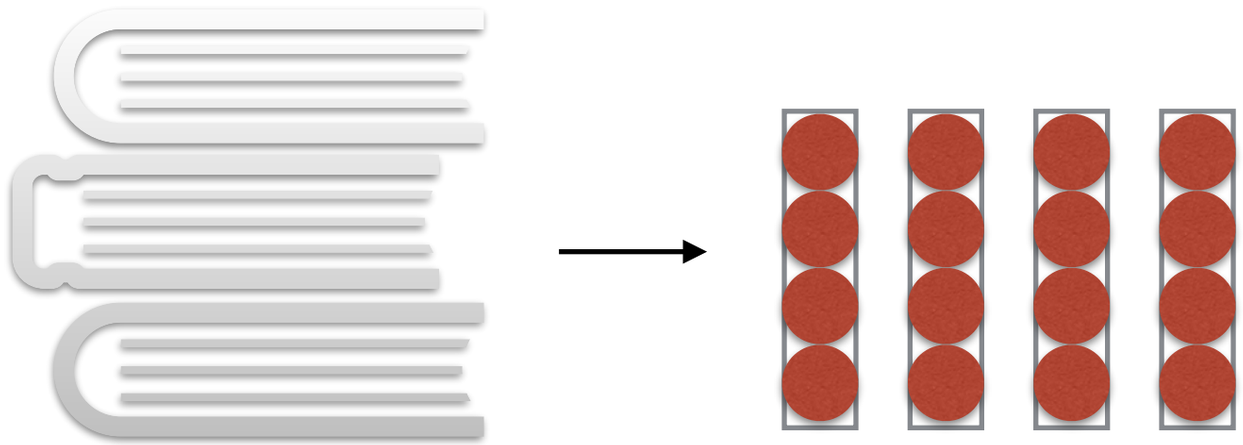
- Text Classification and ML Fundamentals
- Neural Network Basics and Toolkit Construction
- Language Modeling and NN Training Tricks

Topic 2: Sequence Models



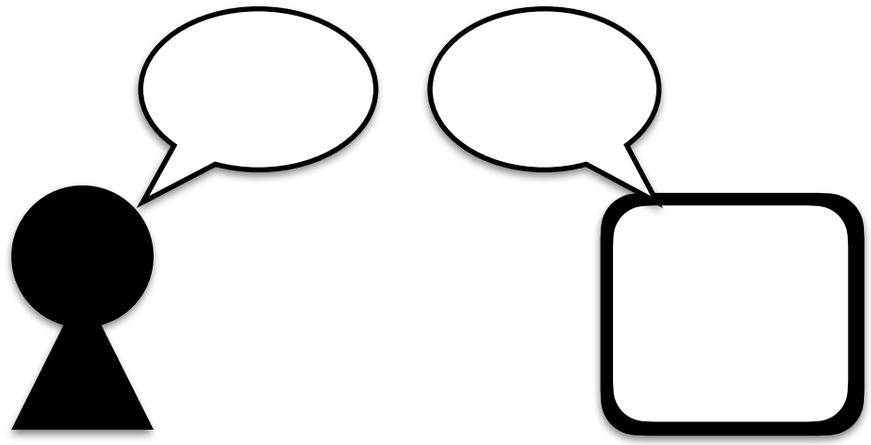
- Recurrent Networks
- Sequence Labeling
- Conditioned Generation
- Attention

Topic 3:
Representation and Pre-training



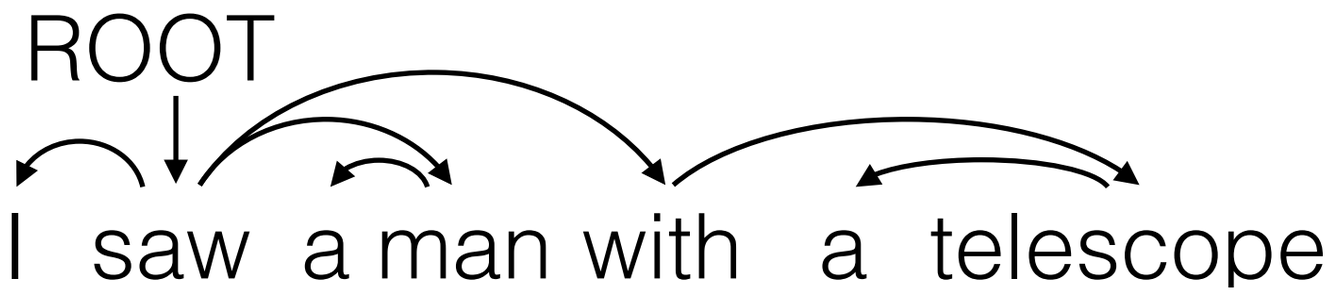
- Pre-training Methods
- Multi-task Learning
- Interpreting and Debugging NLP Models

Topic 4: NLP Applications



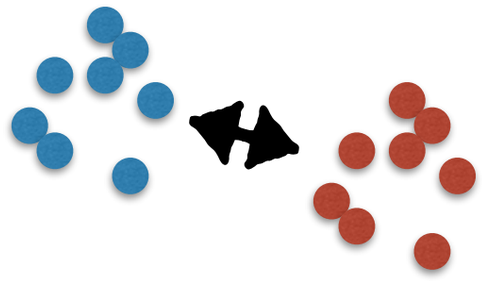
- Question Answering
- Dialog
- Computational Social Science, Bias and Fairness
- Language-to-Code and Language Interfaces

Topic 5:
Natural Language Analysis



- Word Segmentation and Morphology
- Syntactic Parsing
- Semantic Parsing
- Discourse Structure and Analysis

Topic 6:
Advanced Learning Techniques



- Long Sequence Models
- Structured Learning Algorithms
- Latent Variable Models
- Adversarial Methods

Class Format/Structure

Class Delivery Format:
In Person, with Zoom streaming

- In-person preferred (easier to ask and answer questions!)
- We will also stream on Zoom and record. See Piazza for the links.
- Answer a class's quiz within 24 hours of the class (no quiz for this lecture).
- If you're on the waitlist, hang tight — there's a reasonable chance you will get in. You should have gotten an email with information. Please attend on Zoom for now, for capacity reasons.

Class Content Format

- **Before class:** For some classes, do recommended reading
- **During class:**
  - *Lecture/Discussion:* Go through material and discuss
  - *Code/Data Walk:* The TAs (or instructor) will sometimes walk through some demonstration code, data, or model predictions
- **After class:** Do quiz about class or reading material (no quiz for today)

Assignments

- **Assignment 1 - Build-your-own BERT:** *Individually* implement BERT model loading and training
- **Assignment 2 - NLP Task from Scratch:** *In a team,* perform data creation, modeling, and evaluation for a specified task
- **Assignment 3 - SOTA Survey / Re-implementation:** Survey literature, re-implement and reproduce results from a recently published NLP paper
- **Assignment 4 - Final Project:** Perform a unique project that either (1) improves on state-of-the-art, or (2) applies NLP models to a unique task. Present a poster and write a report.

Daniel Fried

Robert Frederking

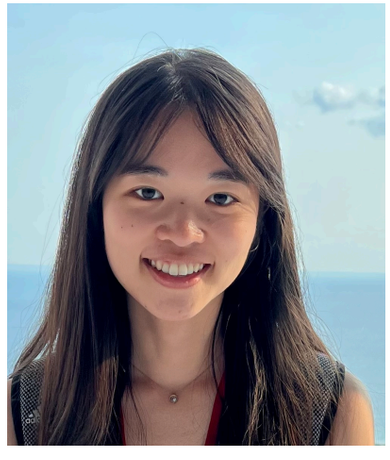
Zora Wang

Saujas Vaduguru

Robert Lo

Sang Choe

Aprameya Bharadwaj



Sophia Chou

Atharva Kulkarni

Bowen Tan

Links

- **Website**: cmu-anlp.github.io
- **Piazza**: https://piazza.com/cmu/fall2023/11711
- **Email**: anlp-fall-2023@mailman.srv.cs.cmu.edu
  - But, Piazza will likely get a faster response for questions about content and assignments.

Thanks, Any Questions?