CS11-711 Advanced NLP

Introduction to Natural Language Processing

Graham Neubig

Carnegie Mellon University

Language Technologies Institute

https://phontron.com/class/anlp2024/

What is NLP Anyway?

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

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

- Aid **human-human communication** (e.g. machine translation, spell checking, assisted writing)

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

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

NLP can Answer Questions

GN **You**

Who is the current president of Carnegie Mellon University?

ChatGPT

I did a quick search for more information and here's what I found.

The current president of Carnegie Mellon University is Farnam Jahanian. He has been serving in this role since July 1, 2017, first as an interim president and then, from March 8, 2018, as the president [,,].

Retrieved Jan 5, 2024

NLP cannot Answer Questions

GN **You**

How many layers are included in gpt-3.5-turbo's architecture?

ChatGPT

GPT-3.5-Turbo, which is an optimized version of GPT-3.5 for faster responses and lower computational requirements, doesn't have a specific "layer" structure like the traditional GPT-3 models. The "Turbo" variant is designed for interactive applications, providing faster response times by optimizing certain aspects of the model's architecture and operation. However, it's important to note that the underlying model, GPT-3.5, has 175 billion parameters. In the context of GPT models, "layers" typically refer to transformer layers used in the architecture, but for GPT-3.5-Turbo, the focus is more on operational efficiency than on a specific layer count.

Retrieved Jan 5, 2024

NLP can Translate Text

緊急事態宣言から「まん延防止等重点措置」に移行した大阪府では、飲食店での酒類提供が一部解禁された。ただ、提供には府が認証する「ゴールドステッカー」の申請が必須。申請には43項目にのぼる感染対策をクリアする必要があり、飲食店からは「ハードルが高すぎる」との悲鳴が上がっている。「項目が40個以上もあって多すぎるし、ネットでの手続きも難しい。本当に、何から何までややこしい」

→

In Osaka Prefecture, which has transitioned from a state of emergency to ``priority measures to prevent the spread of the disease,'' the ban on serving alcoholic beverages at restaurants has been partially lifted. However, in order to provide the service, it is necessary to apply for a ``gold sticker'' certified by the prefecture. To apply, it is necessary to clear 43 infection control measures, and restaurants have complained that the hurdles are too high. "There are over 40 items, which is too many, and it's difficult to complete the procedures online. It's really complicated in every way."

Front page news from Asahi Shimbun, translated by Google Jan 5, 2024

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 T. rex سالّی وەک سێ 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.

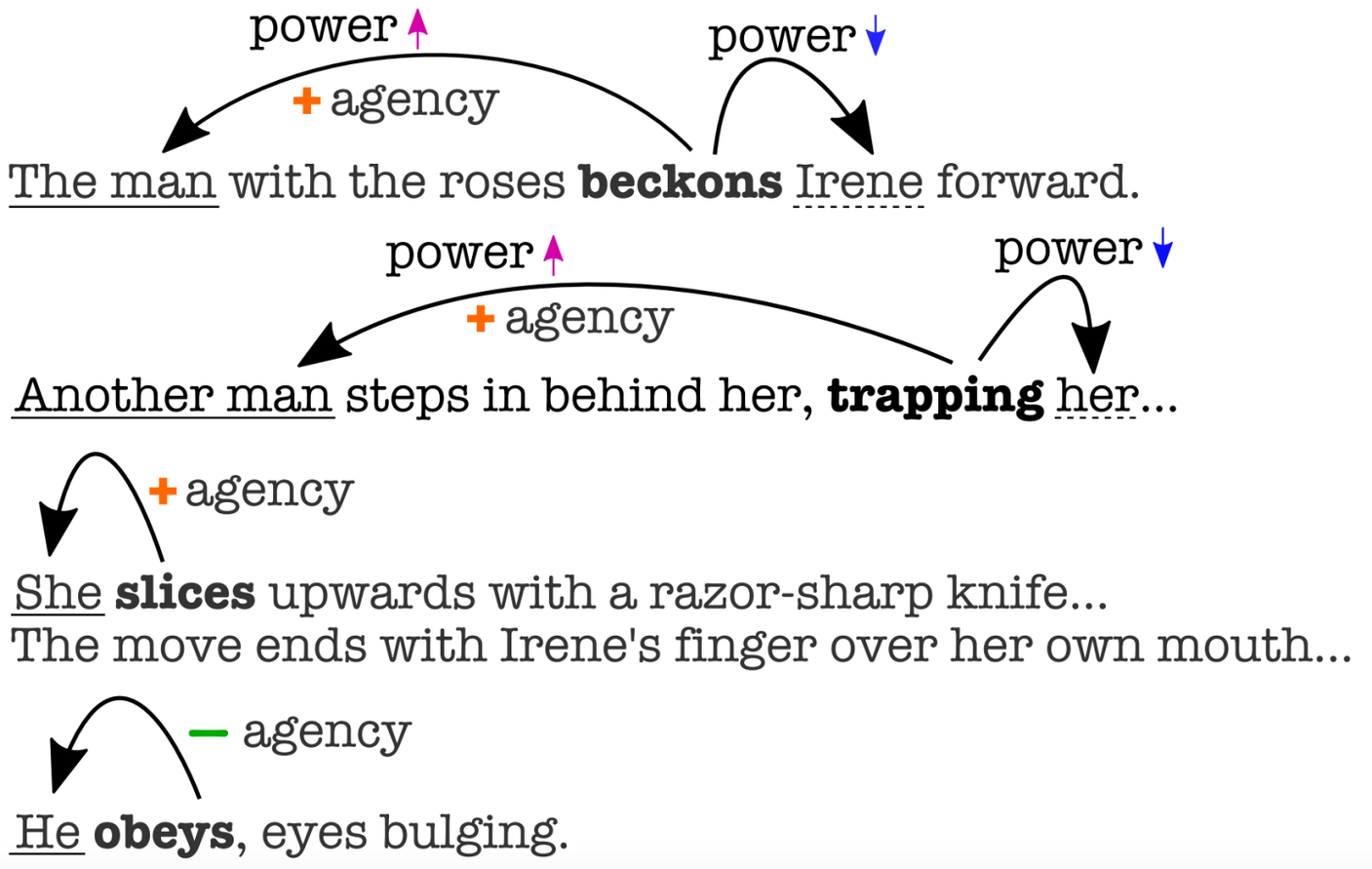
T. regina means "Queen of the ferocious snail.

Front page news from Voice of America Kurdish, translated by Google Jan 5, 2024

Language Analysis 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.

Language Analysis Fails at Basic Tasks

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

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

- What goes into building the state-of-the-art NLP systems that **work uncannily well at some tasks**?

- Where and why do current state-of-the-art NLP systems still **fail**?

- How can we **make appropriate improvements** and **achieve whatever we want to do with NLP**?

NLP System Building

Overview

A General Framework for NLP Systems

- Create a function to map an input *X* into an output *Y*, where *X* and/or *Y* involve language.

Input *X*	Output *Y*	Task
Text	Continuing Text	Language Modeling
Text	Text in Other Language	Translation
Text	Label	Text Classification
Text	Linguistic Structure	Language Analysis
Image	Text	Image Captioning

Methods for Creating NLP Systems

- **Rules:** Manual creation of rules

```
def classify(x: str) -> str:
    sports_keywords = ["baseball", "soccer", "football", "tennis"]
    if any(keyword in x for keyword in sports_keywords):
        return "sports"
    else:
        return "other"
```

- **Prompting:** Prompting a language model w/o training

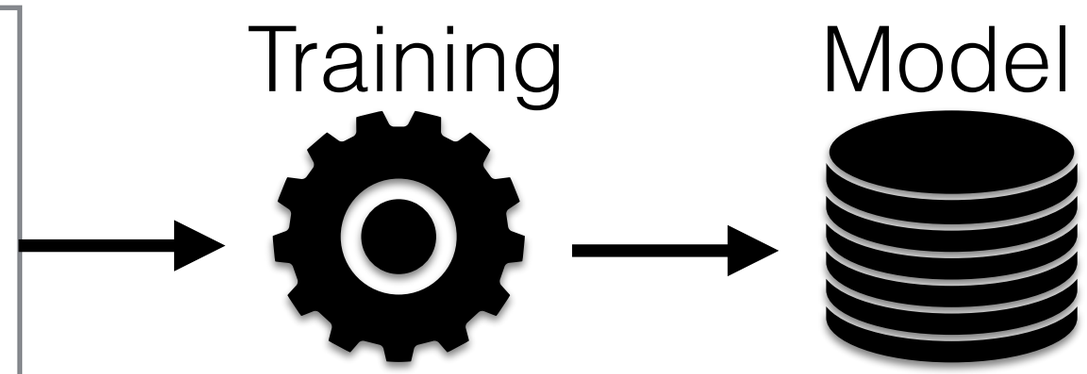
If the following sentences is about “sports” reply “sports”. Otherwise reply “other”.

{X}

→ LM

- **Fine-tuning:** Machine learning from paired data <*X*, *Y*>

I love to play baseball.	sports
The stock price is going up.	other
He got a hat-trick yesterday.	sports
He is wearing tennis shoes.	other



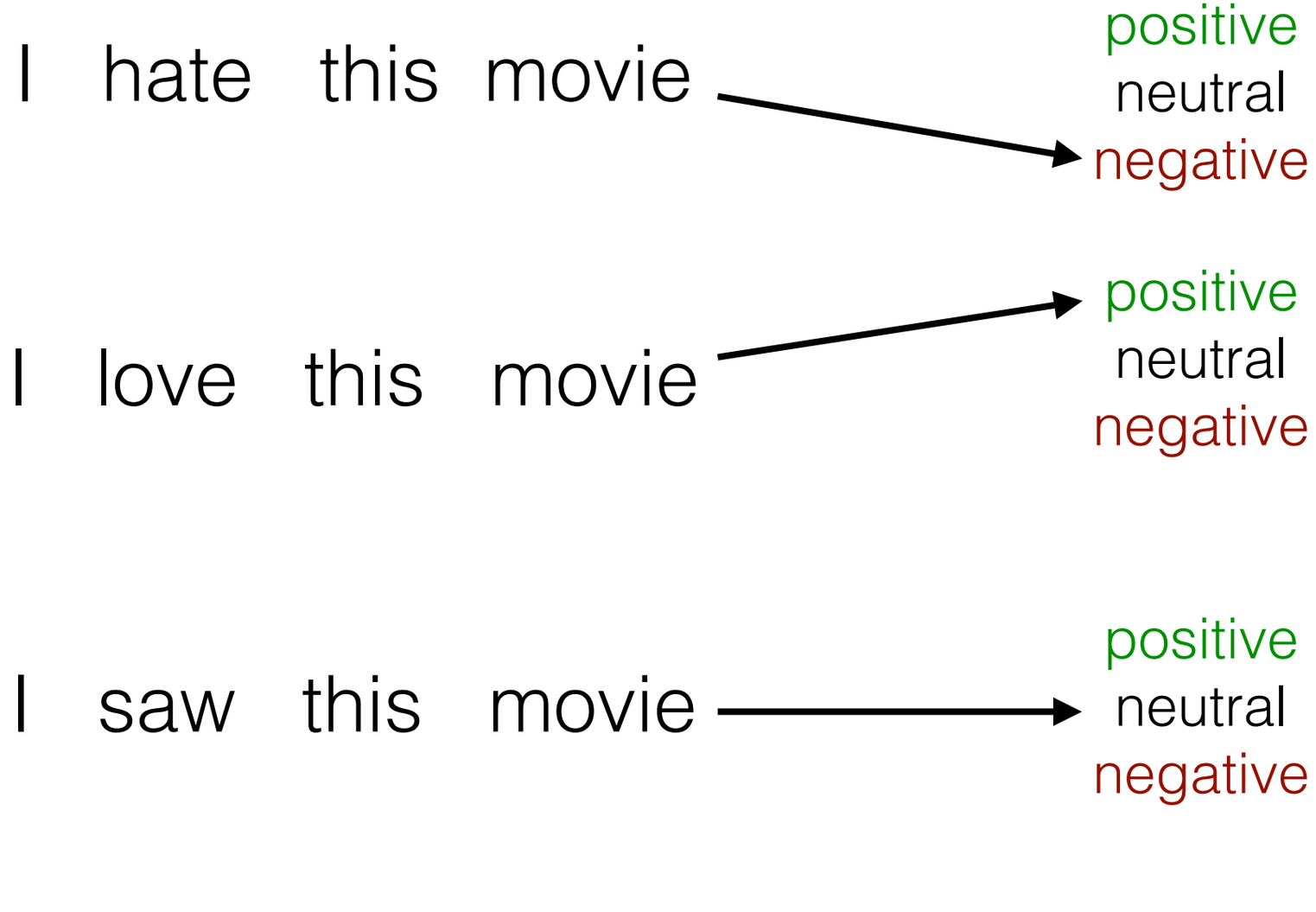
Data Requirements for System Building

- **Rules/prompting based on intuition:**
  No data needed, but also no performance guarantees
- **Rules/prompting based on spot-checks:**
  A small amount of data with input *X* only
- **Rules/prompting with rigorous evaluation:**
  Development set with input *X* and output *Y* (e.g. 200-2000 examples). Additional held-out test set also preferable.
- **Fine-tuning:**
  Additional train set. More is often better — constant accuracy increase when data size doubles.

Let's Try to 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)



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

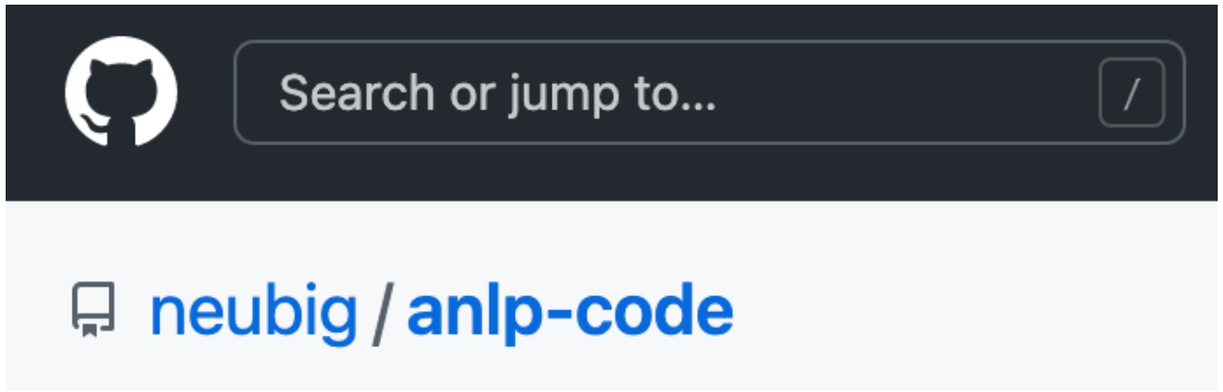
Sentiment Classification Code Walk!

https://github.com/neubig/anlp-code/tree/main/01-simpleclassifier

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

Now Let's Look at Data

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



data/sst-sentiment-text-threeclass

- Remember: don’t look at “test”!

Now Let's Improve!

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

2. Modify the system (featurization, scoring function, etc.)

3. Measure accuracy improvements, accept/reject change

4. Repeat from 1

5. Finally, when satisfied with 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

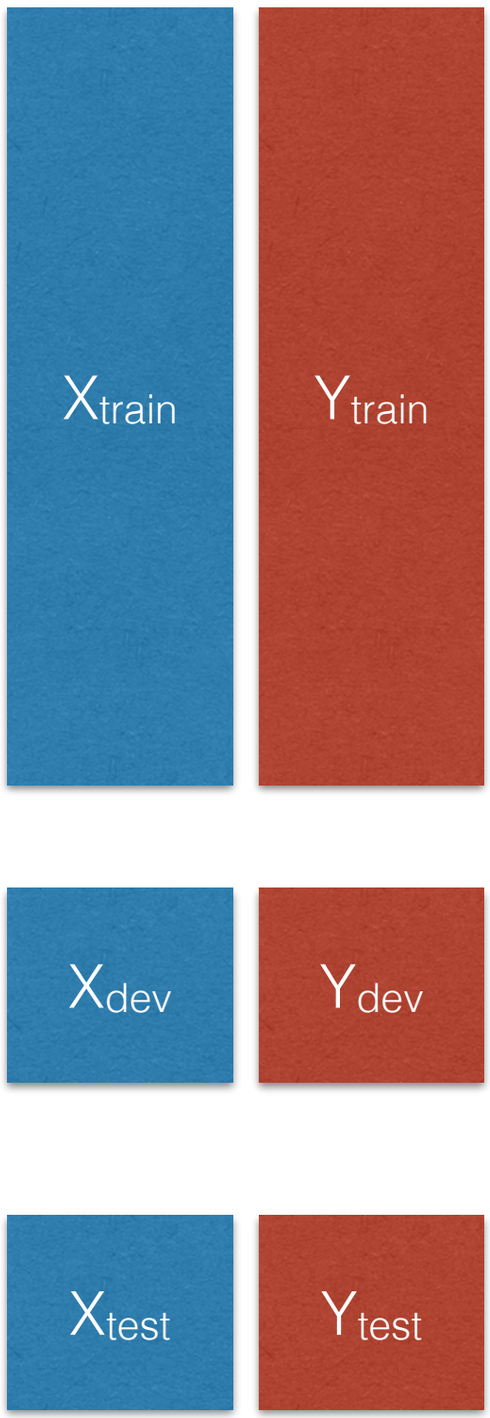
モンハンの名前がついてるからとりあえずモンハン要素をちょこちょこ入れればいいだろ感が凄い。

negative

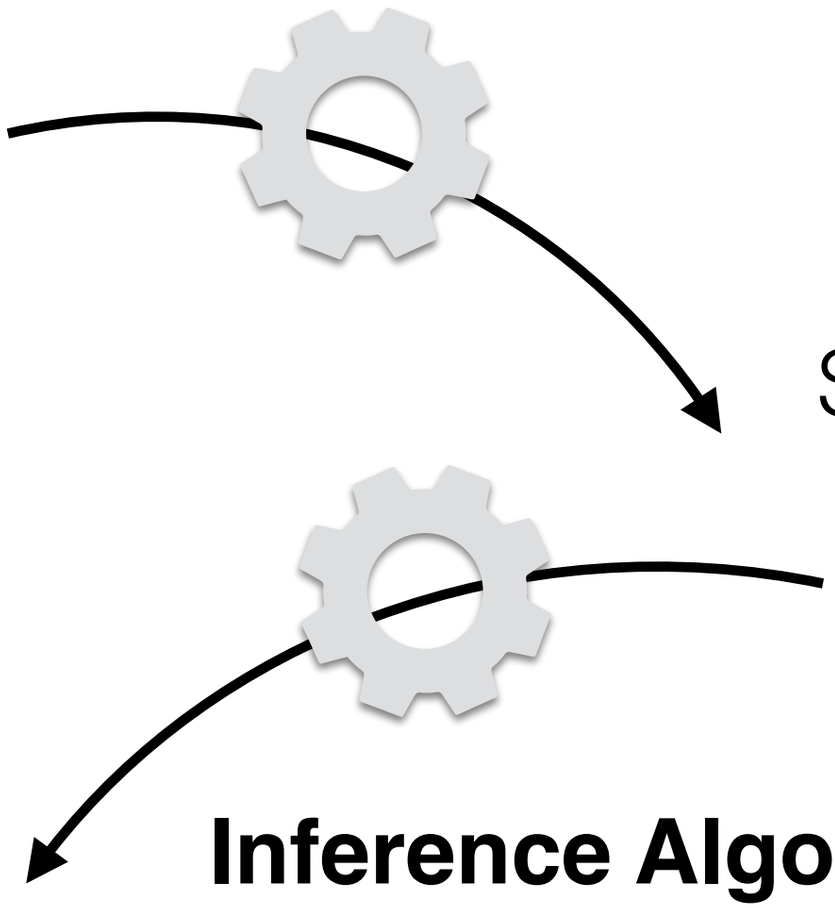
Solution?: Learn Japanese?

Machine Learning Based NLP

Machine Learning



Learning Algorithm



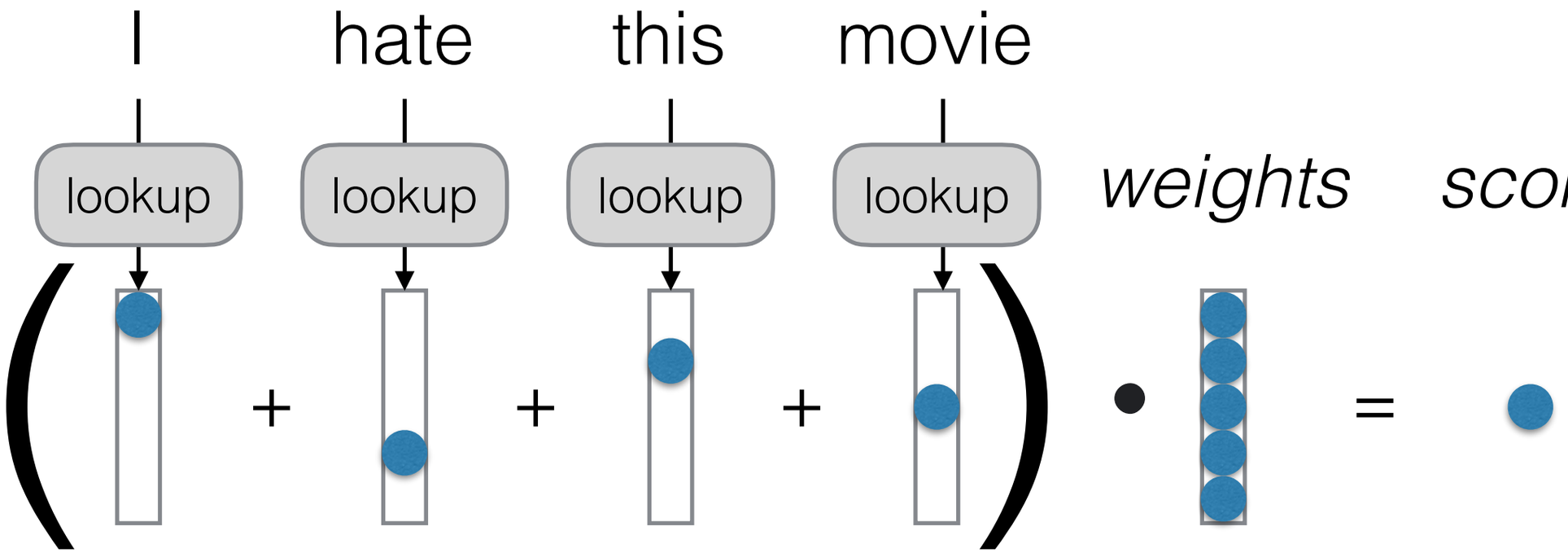
Learned
Feature Extractor *f*
Scoring Function *w*

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

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

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?

What do Our Vectors Represent?

- **Binary classification:** Each word has a single scalar, positive indicating “yes” and negative indicating “no”

- **Multi-class classification:** Each word has its own 5 elements corresponding to [very good, good, neutral, bad, very bad]

Binary

love	2.4
hate	-3.5
nice	1.2
no	-0.2
dog	-0.3
…	…

Multi-class

	v. positive	positive	neutral	negative	v. negative
love	2.4	1.5	-0.5	-0.8	-1.4
hate	-3.5	-2.0	-1.0	0.4	3.2
nice	1.2	2.1	0.4	-0.1	-0.2
no	-0.2	0.3	-0.1	0.4	0.5
dog	-0.1	0.3	0.6	0.2	-0.2
…			…		

Simple Training of BOW Models

- Use an algorithm called “structured perceptron”

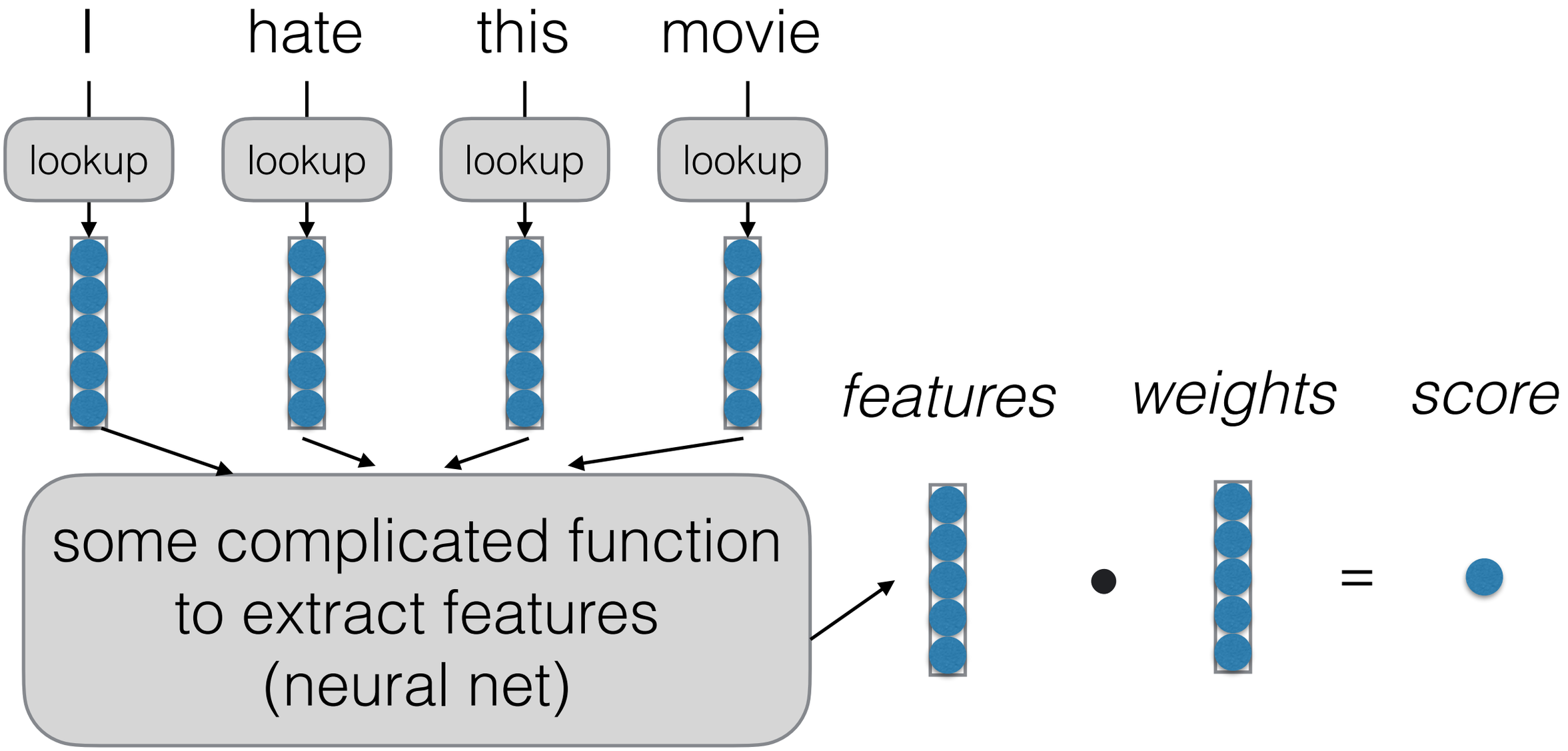
```
feature_weights = {}
for x, y in data:
    # Make a prediction
    features = extract_features(x)
    predicted_y = run_classifier(features)
    # Update the weights if the prediction is wrong
    if predicted_y != y:
        for feature in features:
            feature_weights[feature] = (
                feature_weights.get(feature, 0) +
                y * features[feature]
            )
```

Full Example:
https://github.com/neubig/anlp-code/tree/main/01-simpleclassifier

What's Missing in BOW?

- Handling of *conjugated or compound words*
  - I **love** this move -> I **loved** this movie
- Handling of *word similarity*
  - I **love** this move -> I **adore** this movie
- Handling of *combination features*
  - I **love** this movie -> I **don't love** this movie
  - I **hate** this movie -> I **don't hate** this movie
- Handling of *sentence structure*
  - It has an interesting story, **but** is boring overall

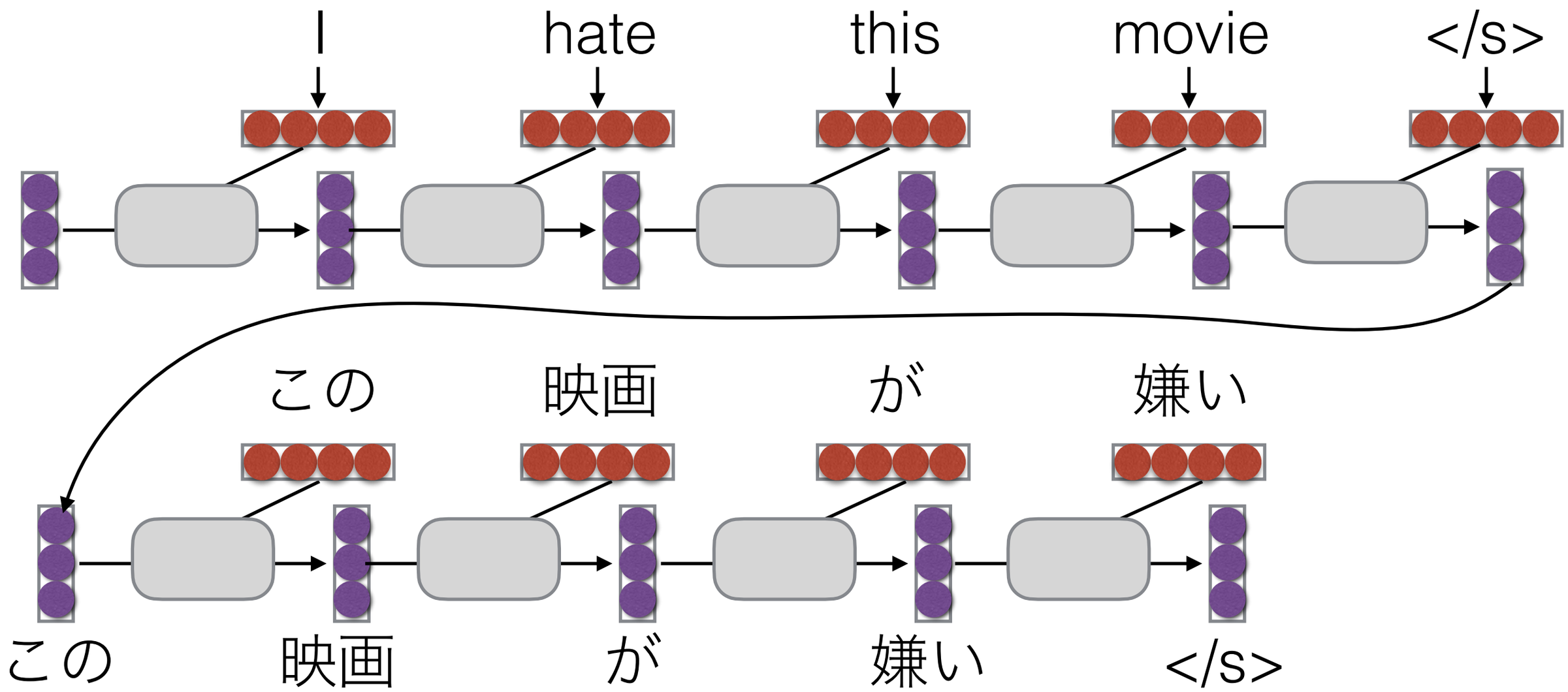
A Better Attempt: Neural Network Models



Powerful enough to perform classification, LM, any task!

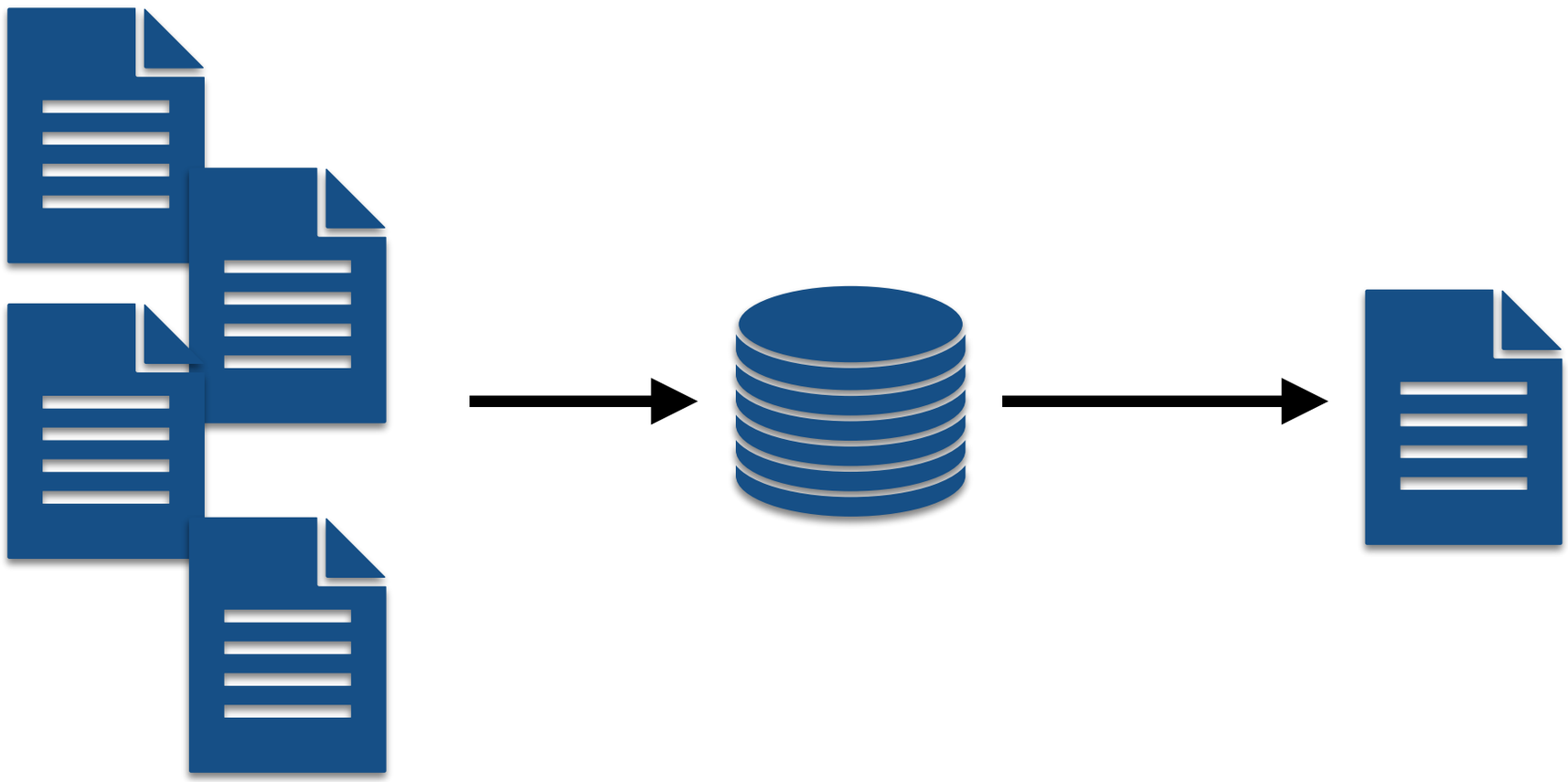
Roadmap Going Forward

Topic 1: Language Modeling Fundamentals



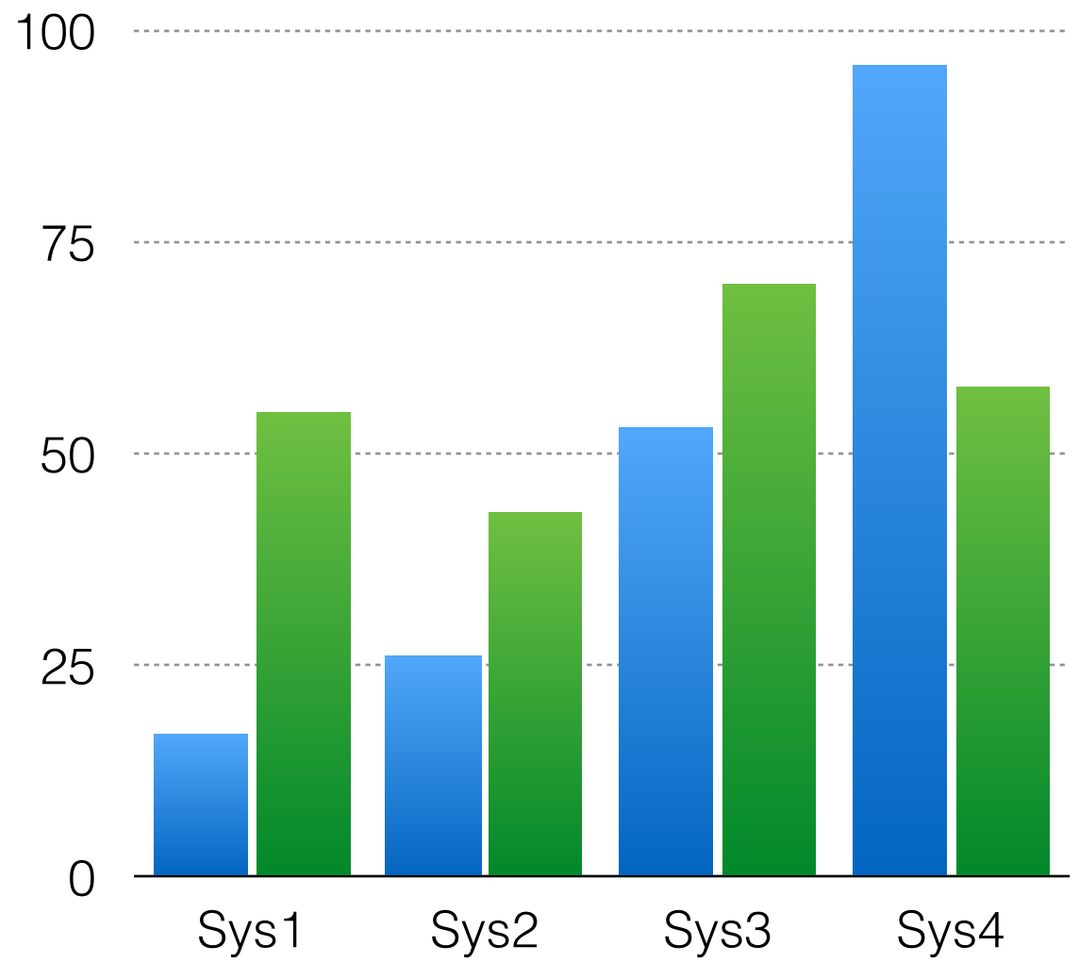
- Representing words
- Language modeling
- Sequence encoding
- Transformers

Topic 2: Training and Inference Methods



- Generation Algorithms
- Prompting
- Instruction Tuning
- Reinforcement Learning

Topic 3: Experimental Design and Evaluation

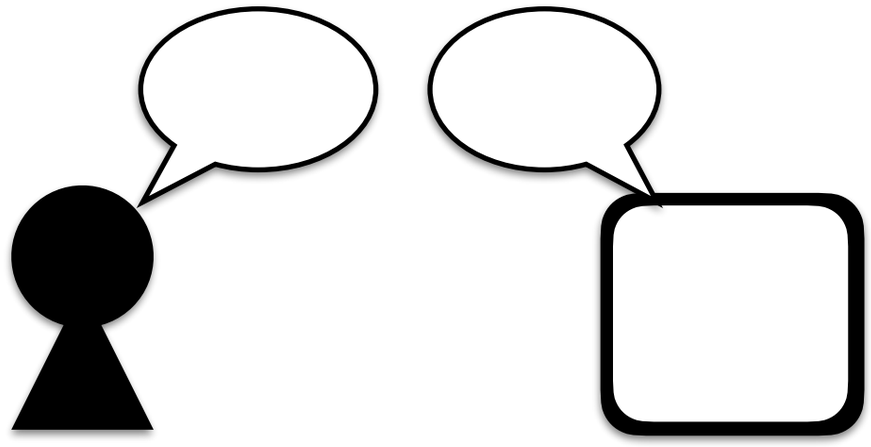


- Experimental Design
- Human Annotation
- Debugging/Interpretation Techniques
- Bias and Fairness Considerations

Topic 4:
Advanced Training and Architectures

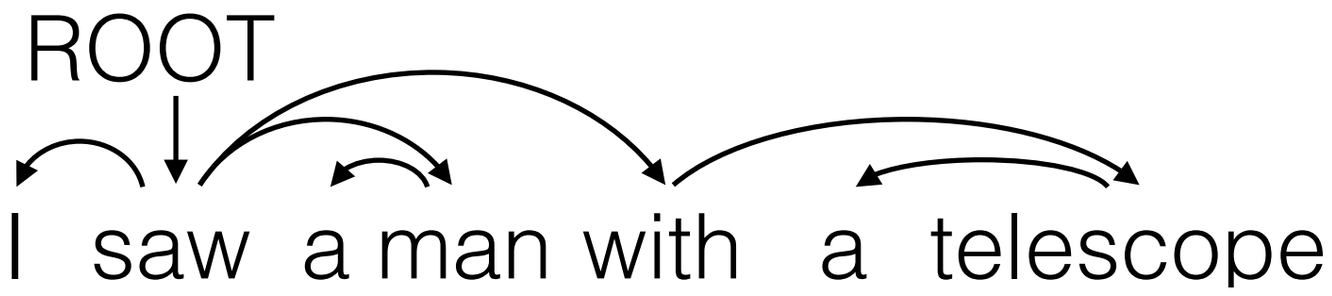
- Distillation and Quantization
- Ensembling and Mixture of Experts
- Retrieval and Retrieval-augmented Generation
- Long Sequence Models

Topic 5: NLP Applications



- Complex Reasoning Tasks
- Code Generation
- Language Agents
- Knowledge-based QA and Information Extraction

Topic 6:
Linguistics and Multilinguality



- Linguistics and Computational Linguistics
- Multilingual NLP

Class Format/Structure

Class Content

- Learn in detail about **building NLP systems from a research perspective**
- Learn basic and advanced topics in **machine learning approaches** to NLP and language models
- Learn **basic linguistic knowledge** useful in NLP
- 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

Class 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

Assignments

- **Assignment 1 - Build-your-own LLaMa:** *Individually* implement LLaMa 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.

Teaching Team

- **Instructor:**
  - Graham Neubig
- **TAs:**
  - Mihir Bansal, Akshay Goindani, Adithya Pratapa, Vishwa Shah, Nishant Subramani, Vijay Visawanathan, Akhila Yerukola
- **Piazza:** https://piazza.com/cmu/spring2024/11711

Thanks, Any Questions?