CS11-711 Advanced NLP

Introduction to Natural Language Processing

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Site
https://phontron.com/class/anlp2022/

What is NLP Anyway?

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

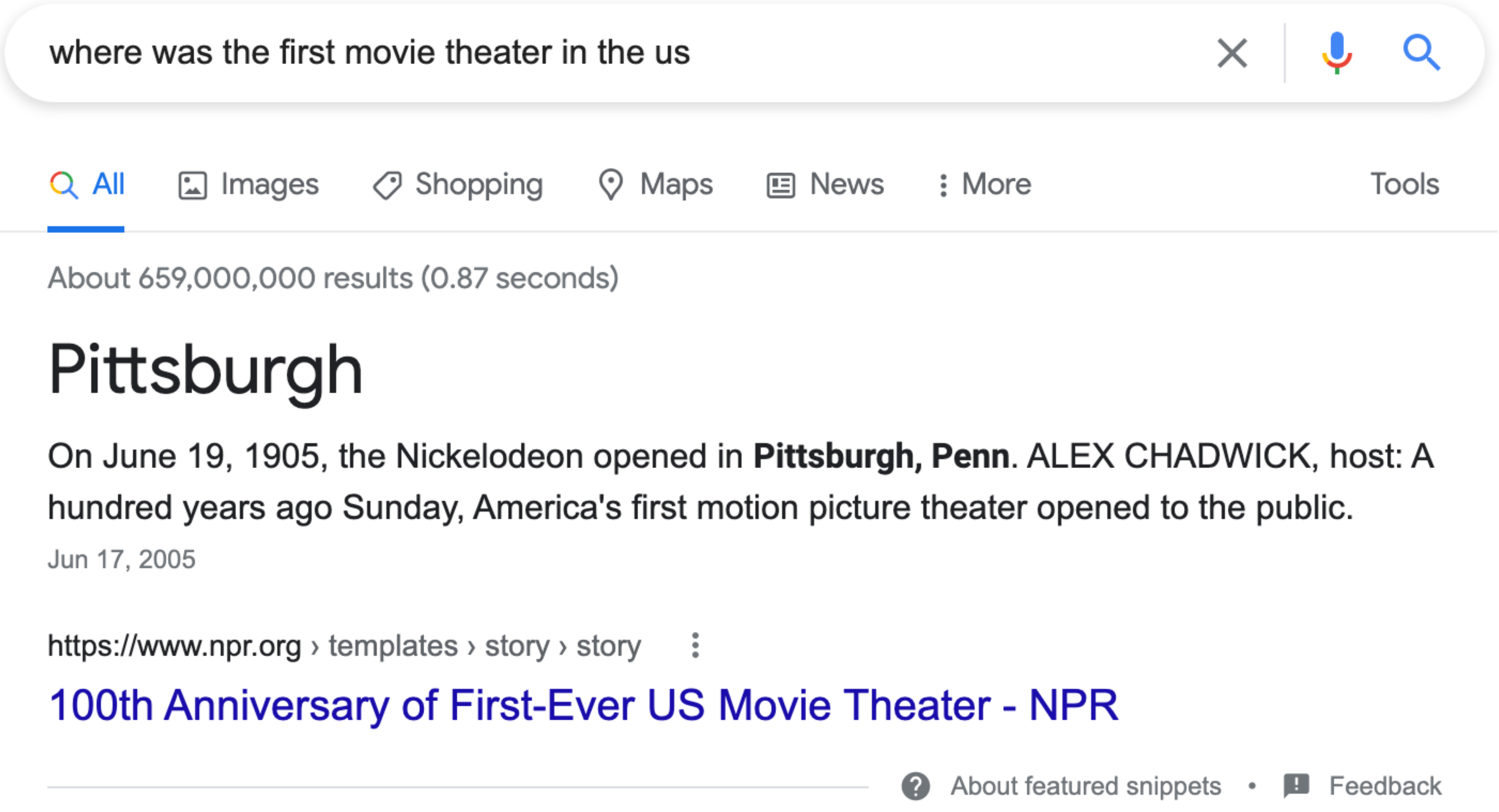
- Aid **human-human communication** (e.g. machine translation)

- 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



where was the first movie theater in the us

About 659,000,000 results (0.87 seconds)

Pittsburgh

On June 19, 1905, the Nickelodeon opened in **Pittsburgh, Penn**. ALEX CHADWICK, host: A hundred years ago Sunday, America's first motion picture theater opened to the public.

Jun 17, 2005

https://www.npr.org › templates › story › story

100th Anniversary of First-Ever US Movie Theater - NPR

Retrieved Aug. 29, 2021

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」

→

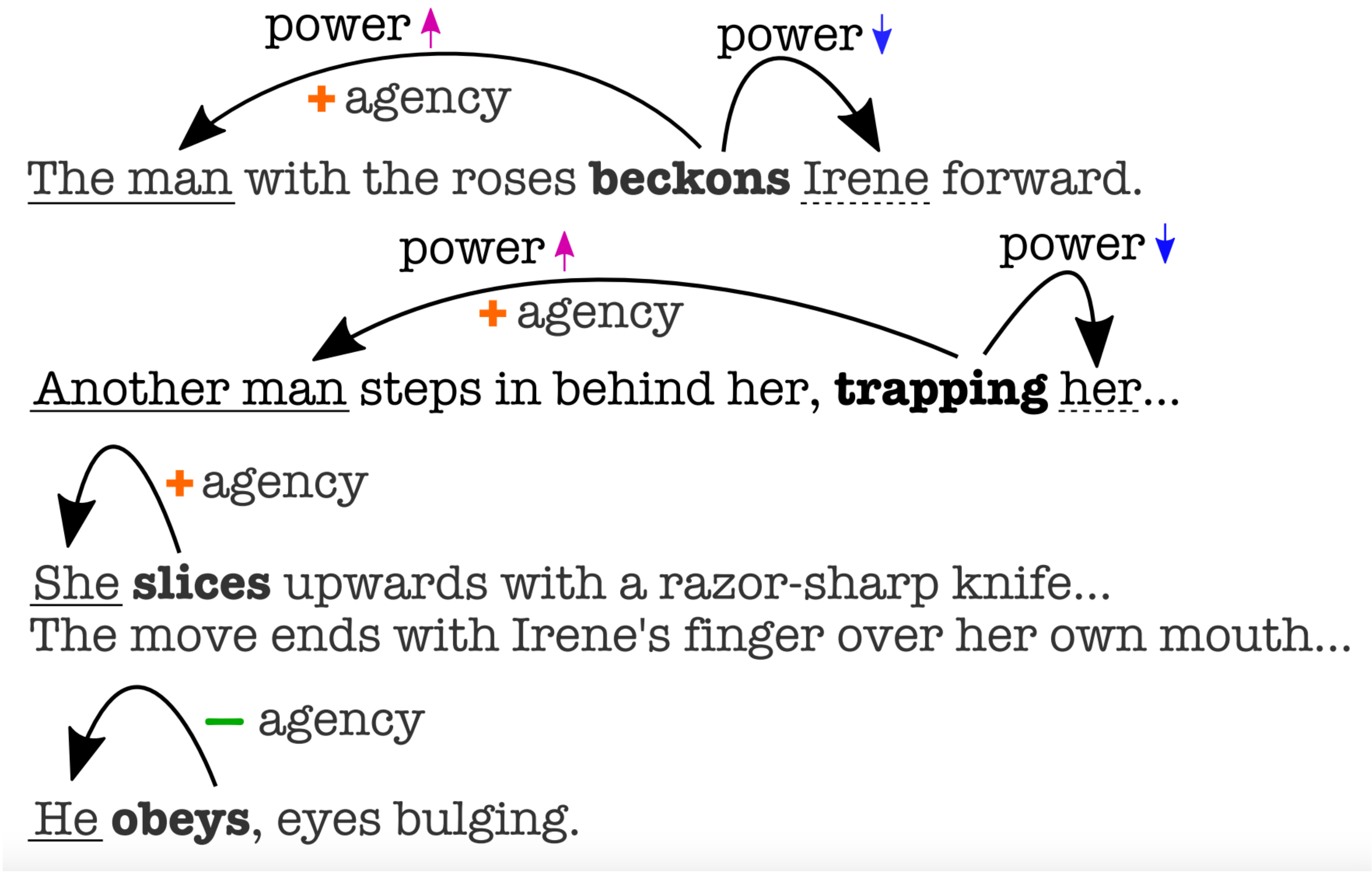
In Osaka Prefecture, which has shifted from a state of emergency to "priority measures such as prevention of spread," the provision of alcoholic beverages at restaurants has been partially lifted. However, it is essential to apply for a "gold sticker" certified by the prefecture to provide it. It is necessary to clear 43 items of infection control in the application, and restaurants are screaming that the hurdle is too high. "There are more than 40 items, too many, and it is difficult to complete the procedure online. It's really complicated."

Front page news from Asahi Shimbun, translated by Google Jun 25., 2021

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]



The man with the roses **beckons** Irene forward.

Another man steps in behind her, **trapping** her...

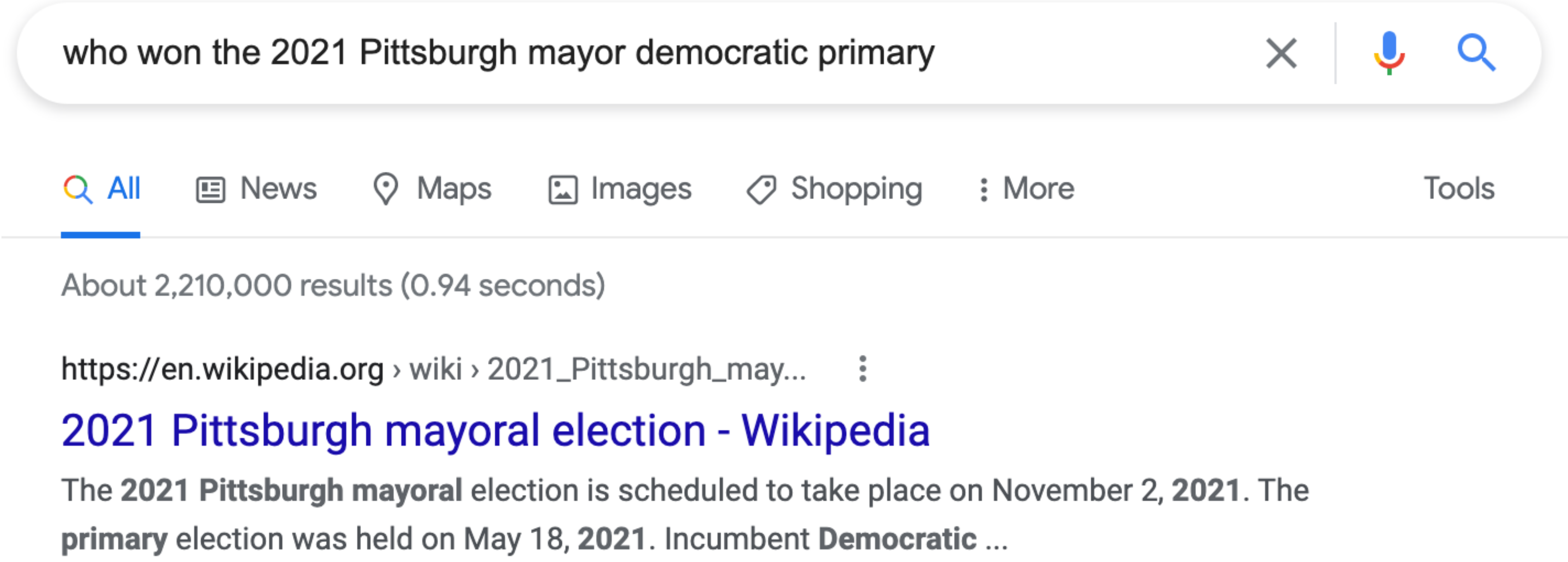
She **slices** upwards with a razor-sharp knife...
The move ends with Irene's finger over her own mouth...

He **obeys**, eyes bulging.

Frame	β	**gender**
agency(AG)=+	−0.951	**M****
power(AG>TH)	−0.468	**M****
agency(AG)=−	0.277	**F****
power(AG<TH)	*not sig.*	

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

NLP cannot Answer our Questions



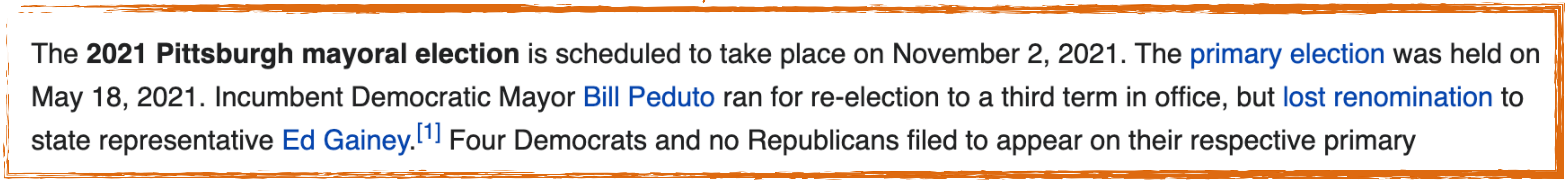
who won the 2021 Pittsburgh mayor democratic primary

About 2,210,000 results (0.94 seconds)

https://en.wikipedia.org › wiki › 2021_Pittsburgh_may...

2021 Pittsburgh mayoral election - Wikipedia

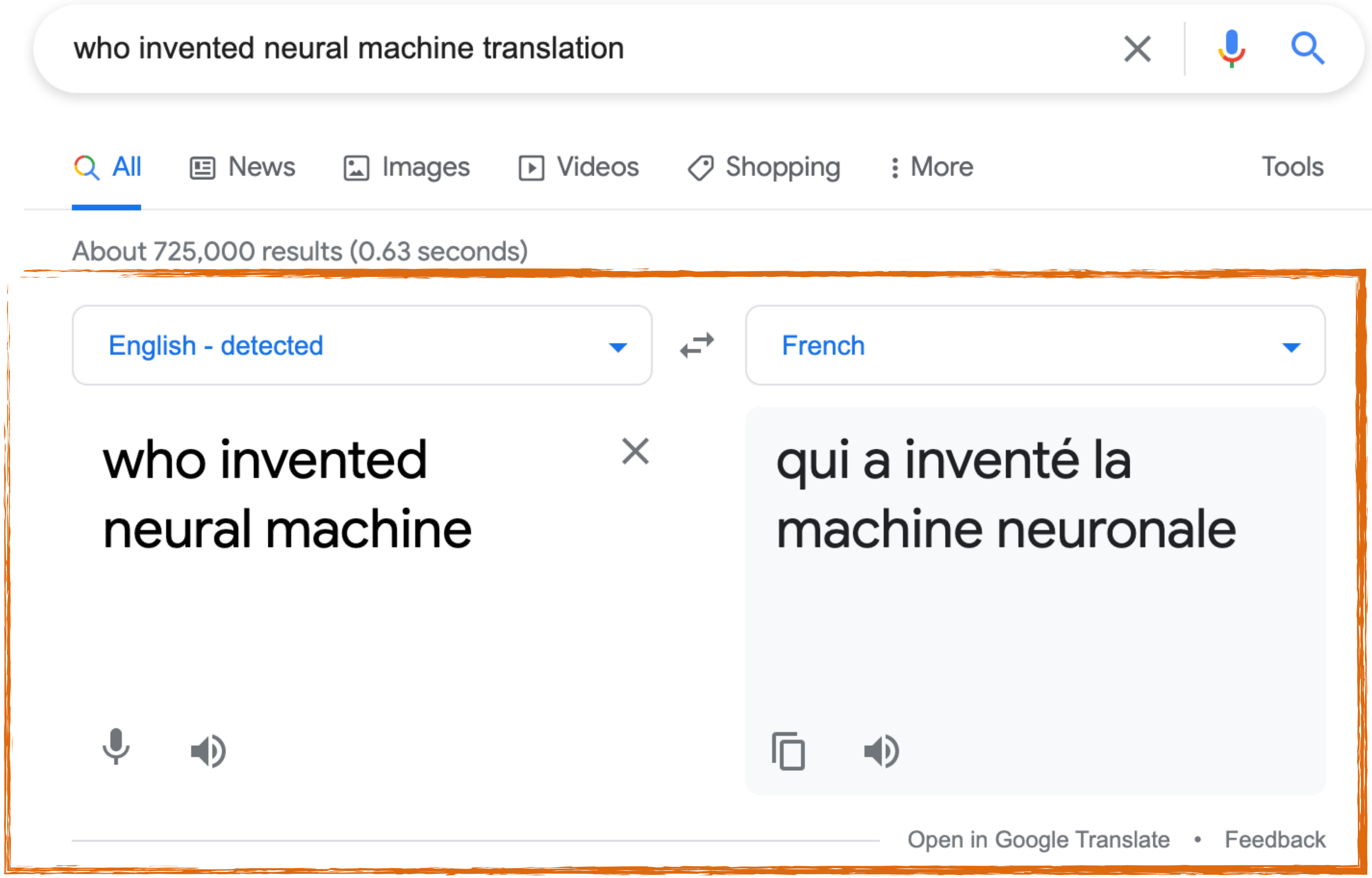
The **2021 Pittsburgh mayoral** election is scheduled to take place on November 2, **2021**. The **primary** election was held on May 18, **2021**. Incumbent **Democratic** ...



The **2021 Pittsburgh mayoral election** is scheduled to take place on November 2, 2021. The primary election was held on May 18, 2021. Incumbent Democratic Mayor Bill Peduto ran for re-election to a third term in office, but lost renomination to state representative Ed Gainey.[1] Four Democrats and no Republicans filed to appear on their respective primary

Retrieved Aug. 29, 2021

NLP cannot Answer our Questions



who invented neural machine translation

About 725,000 results (0.63 seconds)

English - detected

French

who invented neural machine

qui a inventé la machine neuronale

Checked Aug. 7, 2022

NLP cannot Translate Text

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

کە بە واتای "پاشای مارمێلکەی دڕندە" دێت، T. rex جگە لە جۆری
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.

بە واتای "ئیمپراتۆری مارمێلکەی دڕندە دێت T. imperator

T. imperator means "emperor of the savage lizard

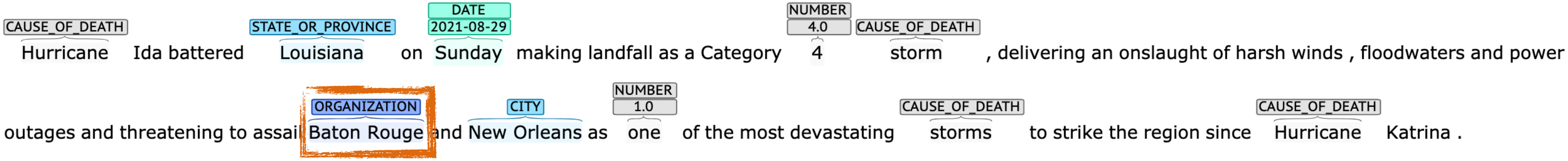
بە واتای "شاژنی مارمێلکەی دڕندە T. regina.

T. regina means "Queen of the ferocious snail.

Front page news from Voice of America Kurdish, translated by Google Aug 7., 2022

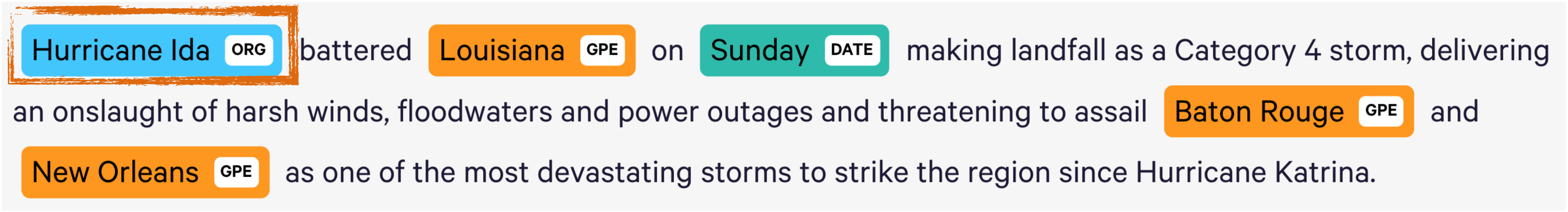
NLP Fails at Even Basic Tasks

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



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

- 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

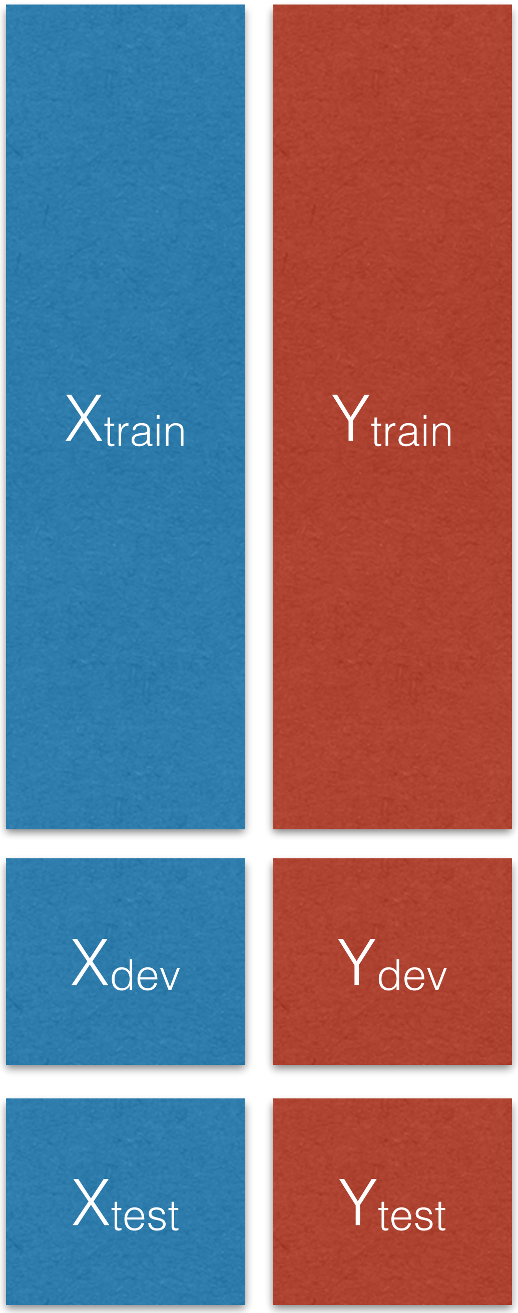
- Formally, create a function to map an input *X (language)* into an output *Y*. Examples:

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.

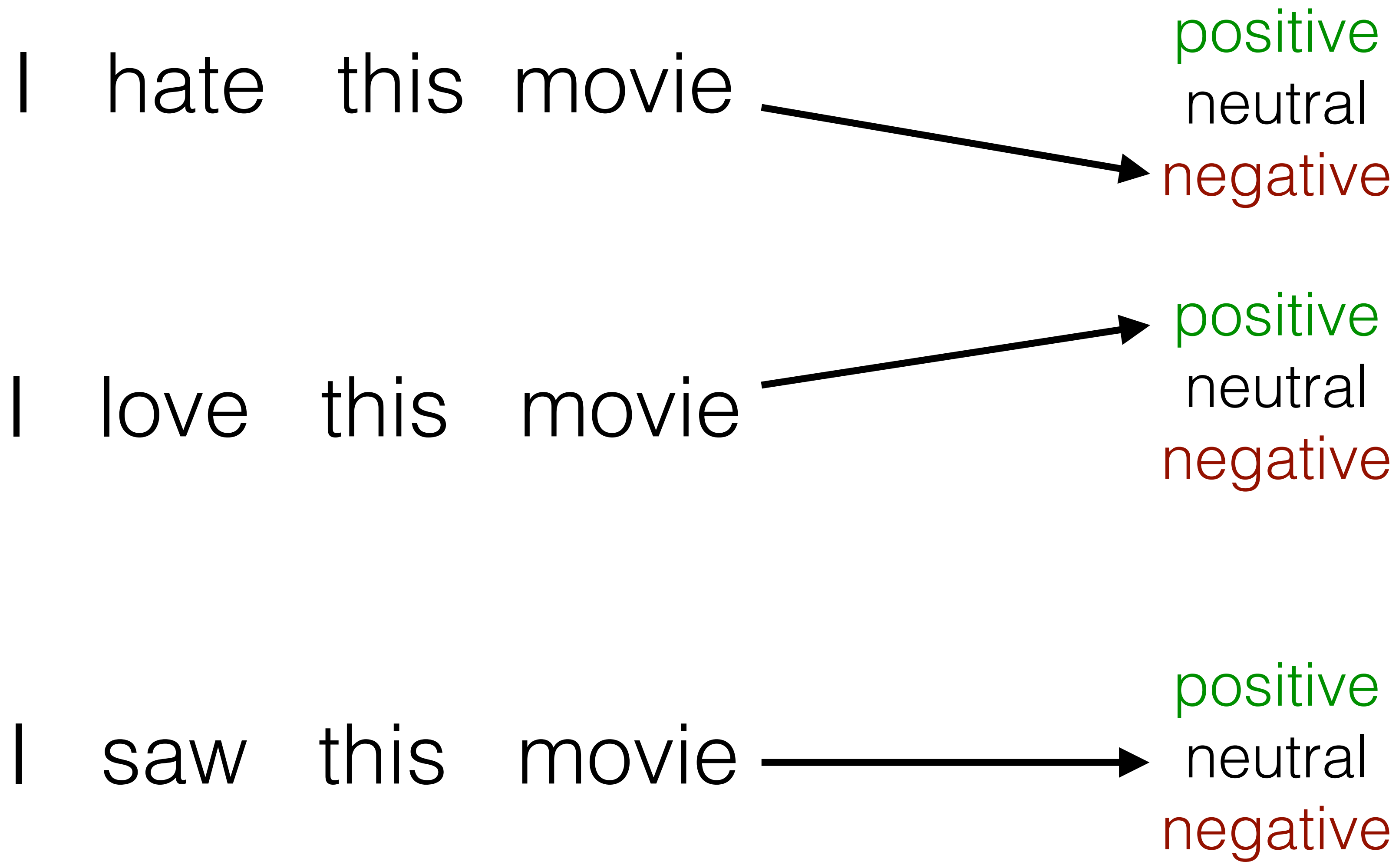
Development ("dev", "validation") Set: Smaller dataset for testing different design decisions ("hyper-parameters").

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)



I hate this movie → negative

positive
neutral
negative

I love this movie → positive

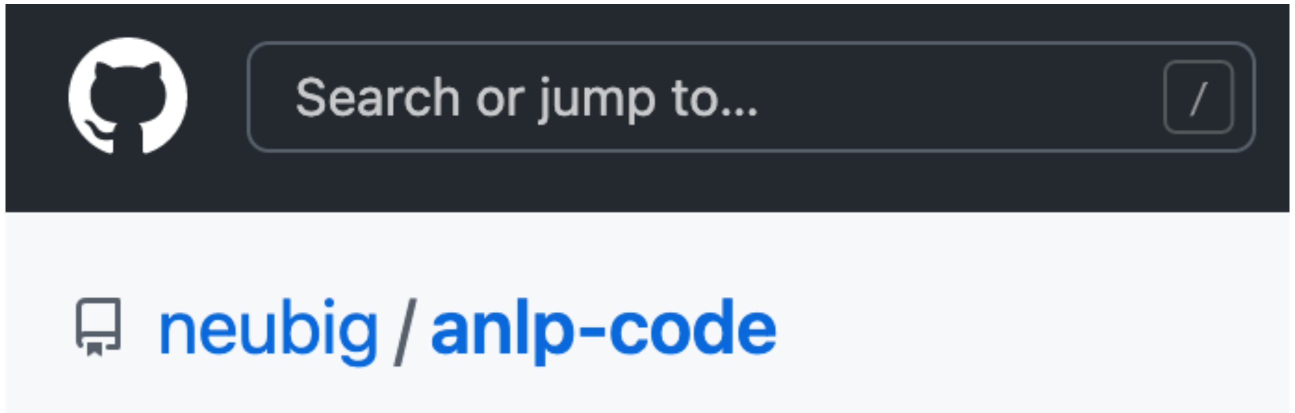
positive
neutral
negative

I saw this movie → neutral

positive
neutral
negative

Let's Look at Data

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



Search or jump to...

neubig / 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

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

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

Sentiment Classification Code Walk!

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

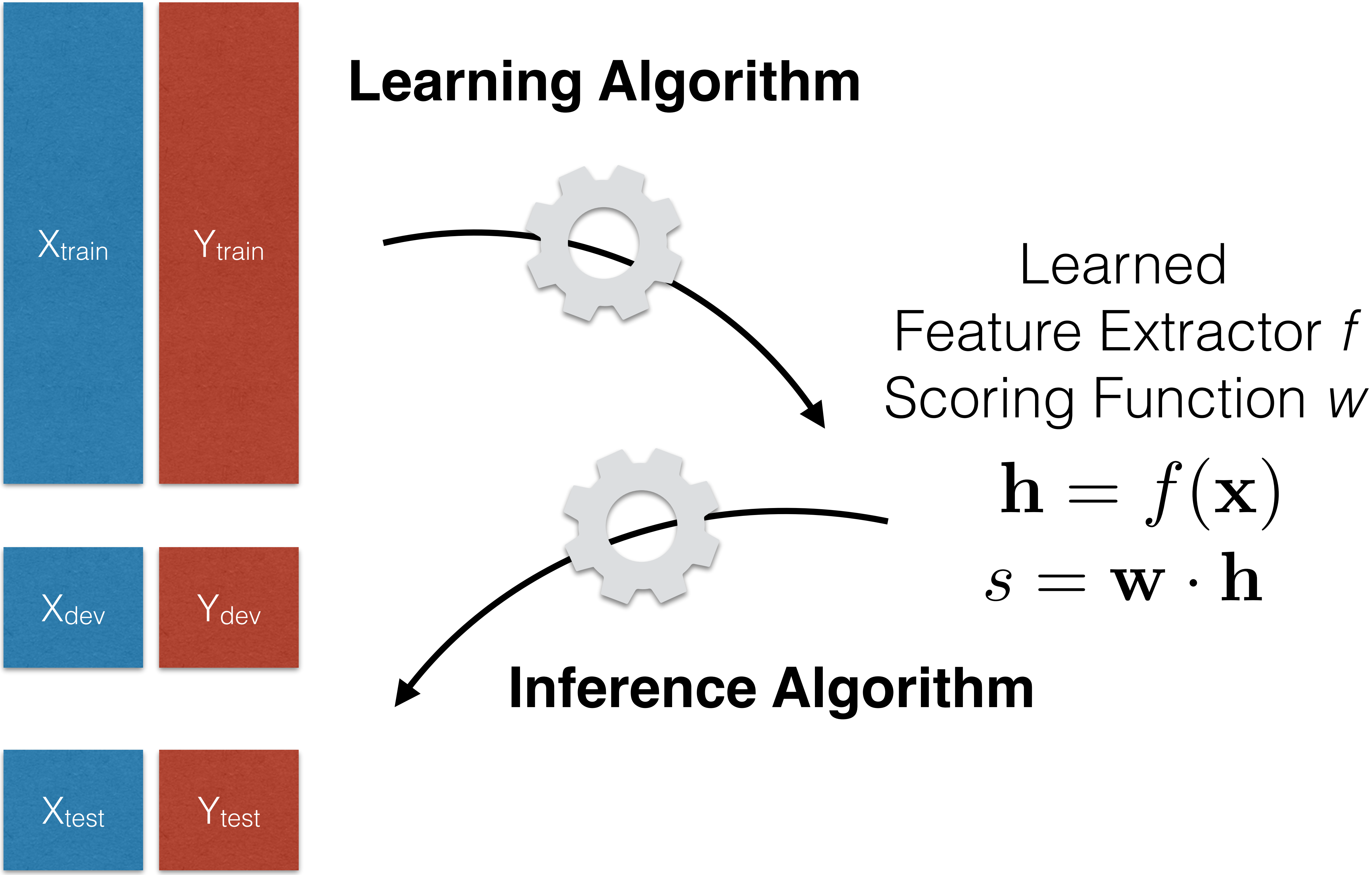
モンハンの名前がついてるからとりあえずモンハン要素を
ちょこちょこ入れればいいだろ感が凄い。

negative

Solution?: Learn Japanese?

Machine Learning Based NLP

Machine Learning



X_{train} Y_{train}

X_{dev} Y_{dev}

X_{test} Y_{test}

Learning Algorithm

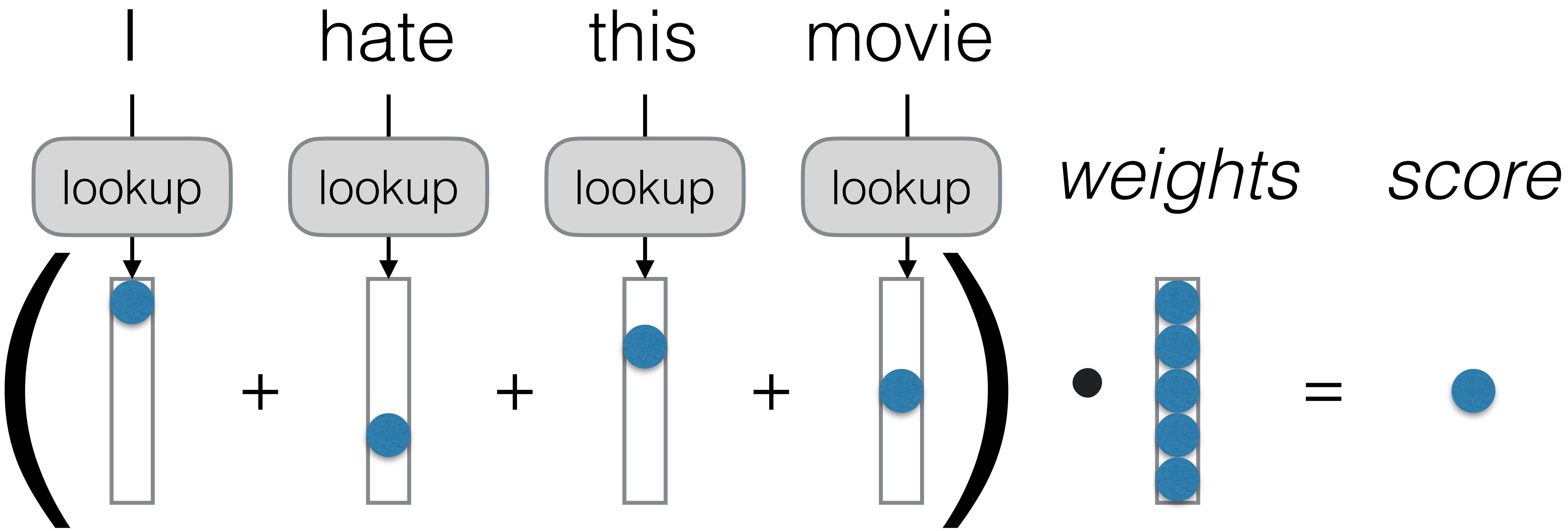
Learned
Feature Extractor *f*
Scoring Function *w*

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

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

Inference Algorithm

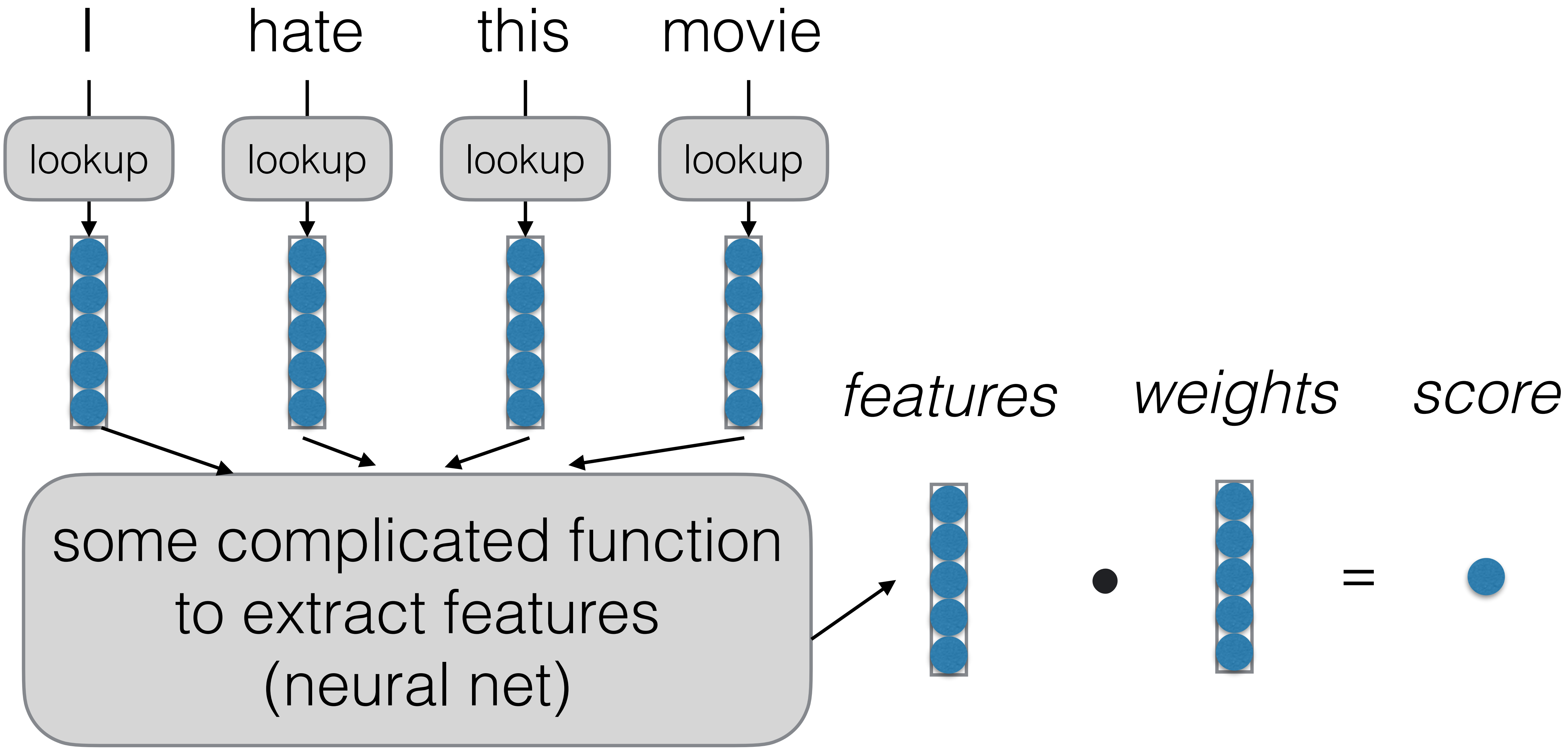
A First Attempt: Bag of Words (BOW)



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

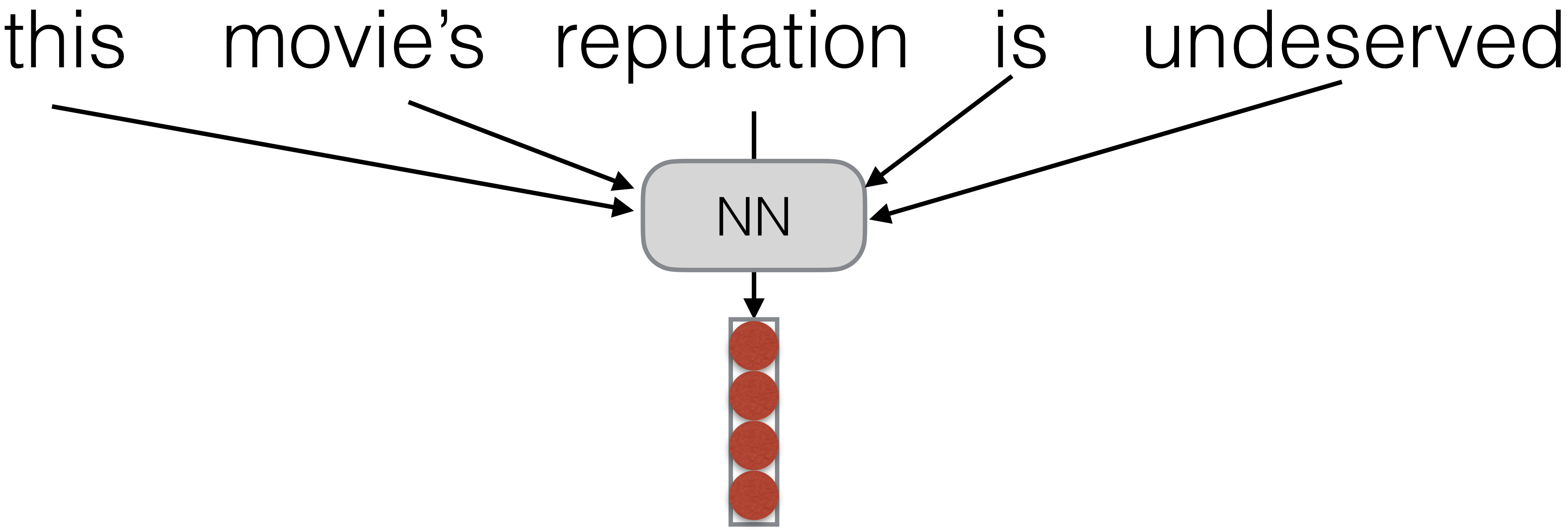
- Learn **basic linguistic knowledge** useful in NLP, and learn methods to **analyze linguistic structure**

- 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

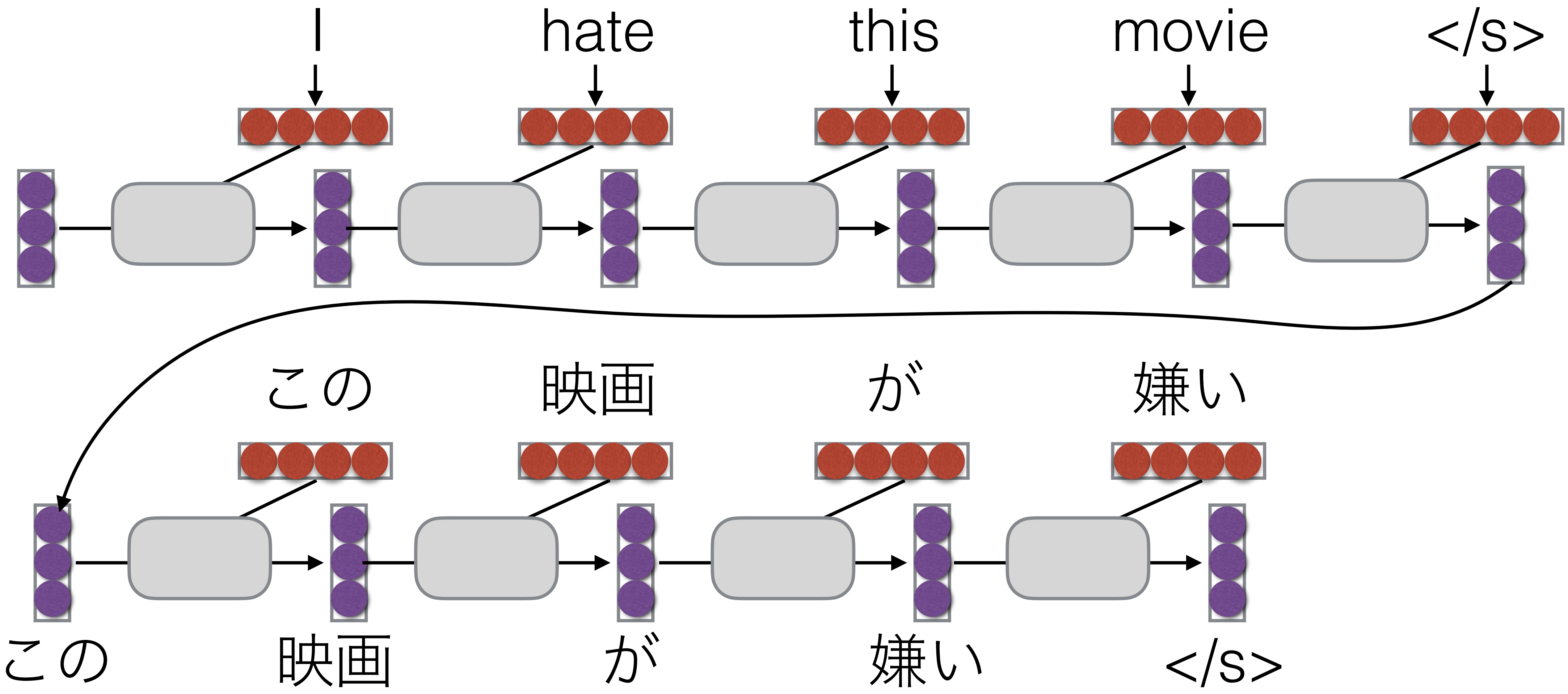
Roadmap Going Forward

Topic 1: Machine Learning and Neural Net Fundamentals



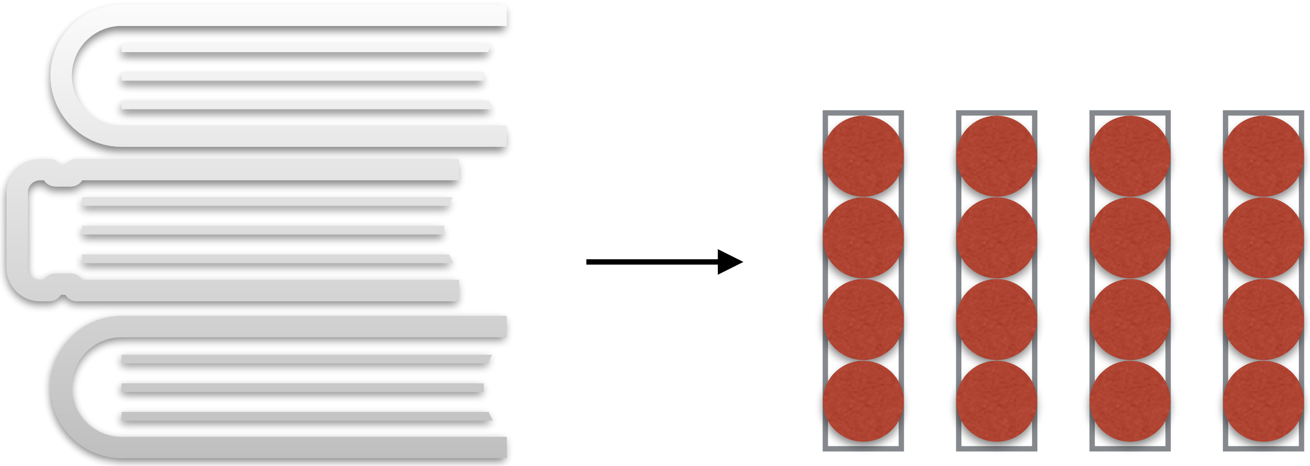
- Text Classification and ML Fundamentals
- Neural Network Basics and Toolkit Construction
- Language Modeling and NN Training Tricks

Topic 2: Sequence Models



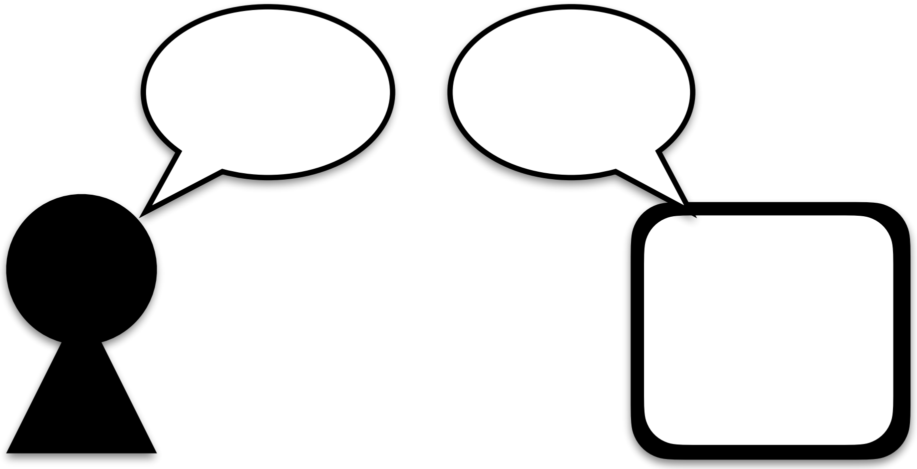
- Recurrent Networks
- Sequence Labeling
- Conditioned Generation
- Attention

Topic 3:
Representation and Pre-training



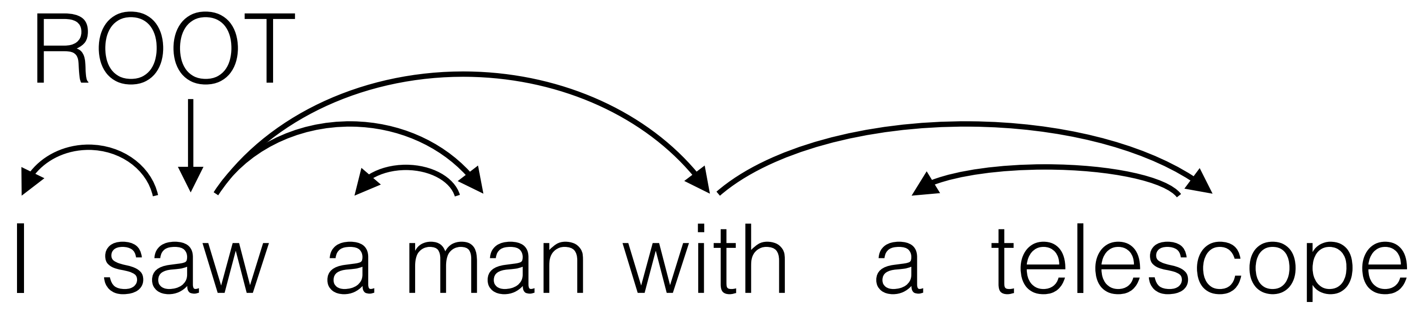
- Pre-training Methods
- Multi-task Learning
- Interpreting and Debugging NLP Models

Topic 4: NLP Applications



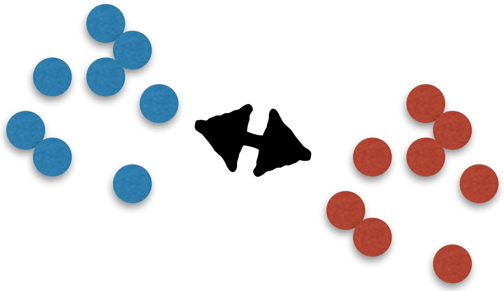
- Machine Reading QA
- Dialog
- Computational Social Science, Bias and Fairness
- Information Extraction and Knowledge-based QA

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 Rotation

- Class split into Tuesday group and Thursday group
- On your day, you are encouraged to come in person
- On the other day, you are encouraged to join synchronously via zoom (but may also come in person if space allows)
- Class will be recorded for review

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

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.

Instructors

- **Instructor:**
  - Graham Neubig (most lectures)
  - Robert Frederking (esp. natural language analysis)
- **TAs:**
  - 10 wonderful TAs, see details on Piazza!
- **Piazza: http://piazza.com/cmu/fall2022/cs11711/home**

Thanks, Any Questions?