

CS11-711 Advanced NLP

Introduction to Natural Language Processing

Graham Neubig



Carnegie Mellon University

Language Technologies Institute

Site

<https://phontron.com/class/anlp2021/>

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



All



Images



Shopping



Maps



News



More

Tools

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



About featured snippets

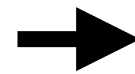


Feedback

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る、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」

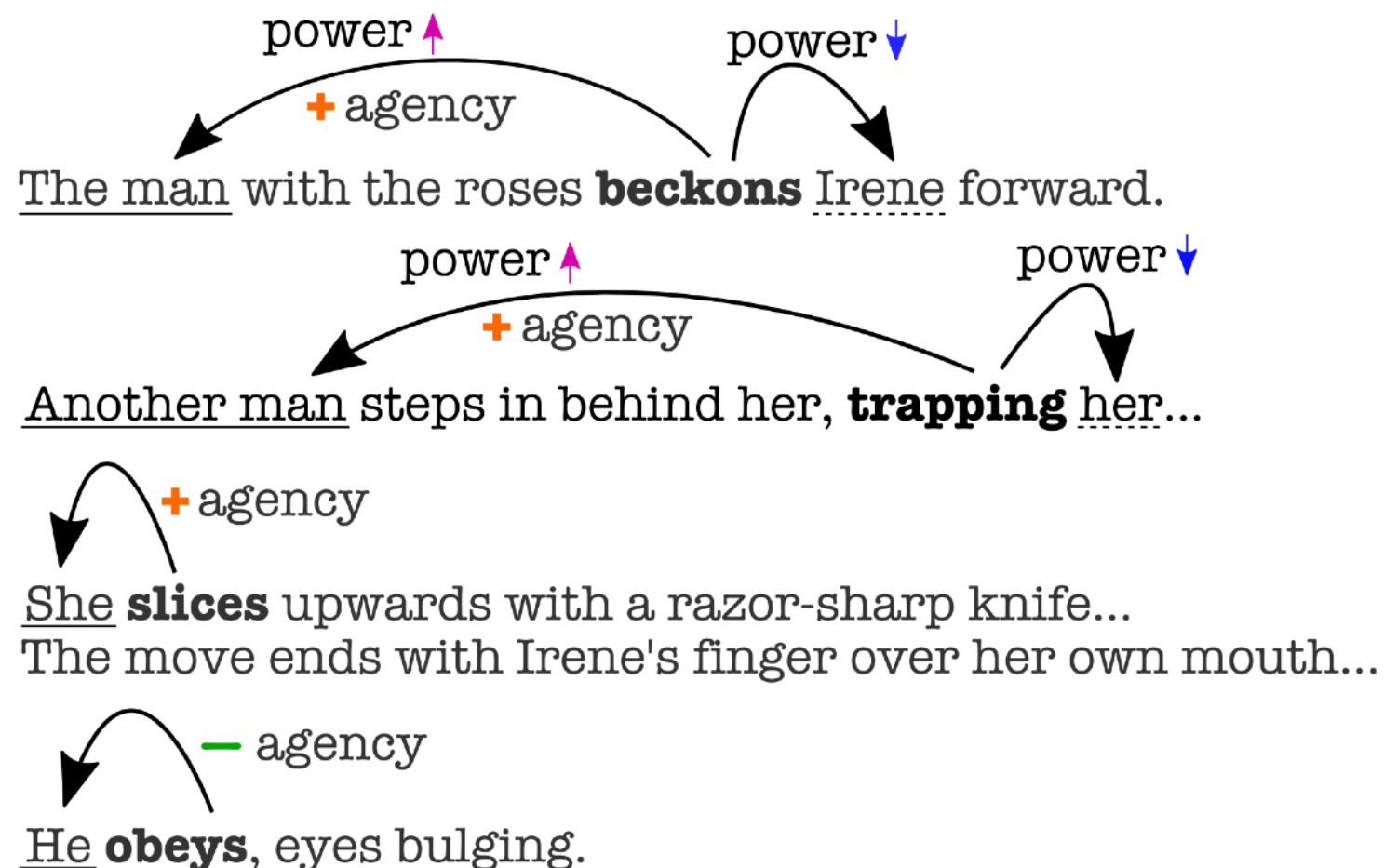


In Osaka Prefecture, which has shifted from a state of emergency to "priority measures such as prevention of spread," the provision of alcoholic beverages at restaurants has been partially lifted. However, it is essential to apply for a "gold sticker" certified by the prefecture to provide it. It is 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]



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

NLP cannot Answer our Questions

who won the 2021 Pittsburgh mayor democratic primary



All

News

Maps

Images

Shopping

More

Tools

About 2,210,000 results (0.94 seconds)

[https://en.wikipedia.org › wiki › 2021_Pittsburgh_may...](https://en.wikipedia.org/wiki/2021_Pittsburgh_mayoral_election)

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

The image shows a Google search interface. At the top, a search bar contains the text "who invented neural machine translation". Below the search bar, navigation links for "All", "News", "Images", "Videos", "Shopping", and "More" are visible, along with a "Tools" link. The search results indicate "About 725,000 results (0.63 seconds)". A Google Translate widget is highlighted with an orange border. The widget shows the source language as "English - detected" and the target language as "French". The English input text is "who invented neural machine", and the French output text is "qui a inventé la machine neuronale". The widget also includes icons for voice input/output and a link to "Open in Google Translate".

who invented neural machine translation

All News Images Videos Shopping More Tools

About 725,000 results (0.63 seconds)

English - detected French

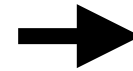
who invented neural machine × qui a inventé la machine neuronale

Open in Google Translate • Feedback

Retrieved Aug. 29, 2021

NLP cannot Translate Text

“၃၇၊ ၃၈ မနီလေးမီဂရု ၈၄ မိနူးလမှူးမဆက်ပေါ့၊ စုတုနူးပဲ
ရွှံသေးတယုပေါ့နော့၊ ကြံ့မူအကြံ့၊ အဲဒါကို သူတို့ကြံ့ တယု
ကေန ဆက်တင့် သတင်းရလာတယု မသိဘူး၊ ခံတွင်း ရောက္ခလာမ
ပီးတော့ အဟုကမှူးမက ဝင်ရောက္ခမီခြင်းတာပေါ့။ အတိအကော်
တာ ကြံ့နော့တာတို့လဲ မသိရသေးဘူး။ ၄ ယောက္ခိဩားတယု
လို့လဲ ချေဟတယု။ ၆ ယောက္ခိဩားတယုလဲ ချေဟတယု။ ဘ
ယုလောက္ခိဩားလဲဆိုတာ ခုအခိန္တိ အတိအကော် မသိရသေးဘူး။ တ
ခါကြံ့ဆို ရောက္ခတာဝင် မရောက္ခကသေးဘူး။ စုတုကတို့မက
အဟုကမှူးမက မီခြင်းလိုကတော့ လက္ခိတော့ အမ္မတ ၆ ထဲမှာ မ
မှူးခံထားရတယု၊ အဲလောက္ခိ သိရသေးတယု။”



"37," he said. 38. 84 Main Road of
Mandalay. I'm still collecting. I don't know
where they got the information in advance.
It arrived immediately and was violently
suppressed. We do not know exactly. He
said four people were involved. He also
said that six people were involved. It is
unknown at this time what he will do after
leaving the post. Some have not even
arrived. He is currently being held in No. 6
after a violent crackdown by the military
junta.

Front page news from Voice of America Burmese, translated by Google Jun 25., 2021

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.

CoreNLP annotations: CAUSE_OF_DEATH (Hurricane), STATE_OR_PROVINCE (Louisiana), DATE (2021-08-29), NUMBER (4.0), CAUSE_OF_DEATH (storm), ORGANIZATION (Baton Rouge), CITY (New Orleans), NUMBER (1.0), CAUSE_OF_DEATH (storms), 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:

- 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

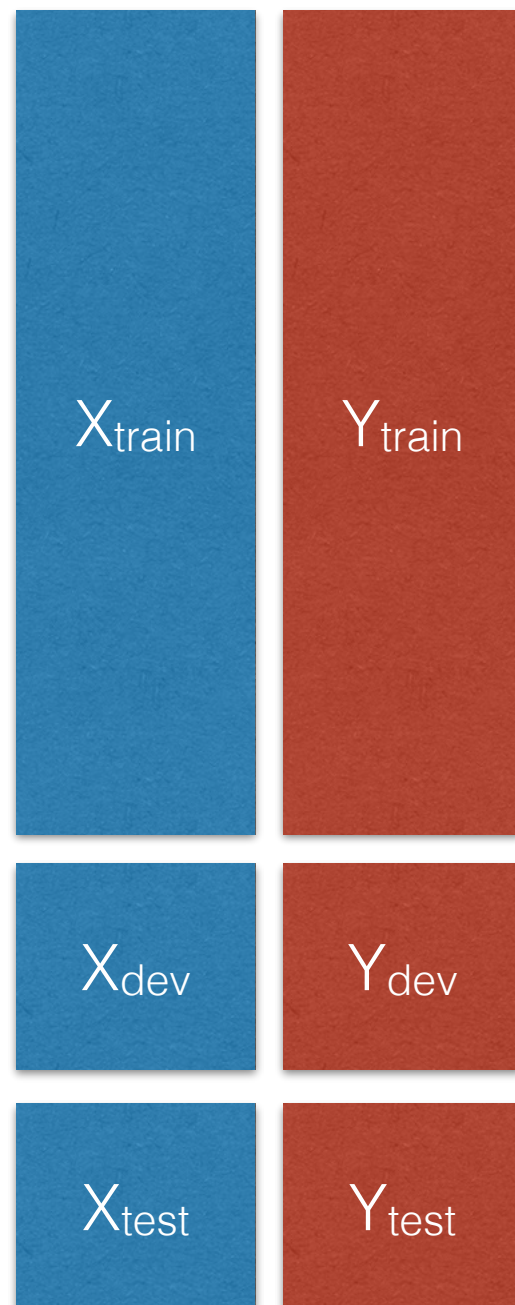
- Formally, create a function to map an **input X (language)** into an **output Y** . Examples:

<u>Input X</u>	<u>Output Y</u>	<u>Task</u>
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 $\langle X, Y \rangle$

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)

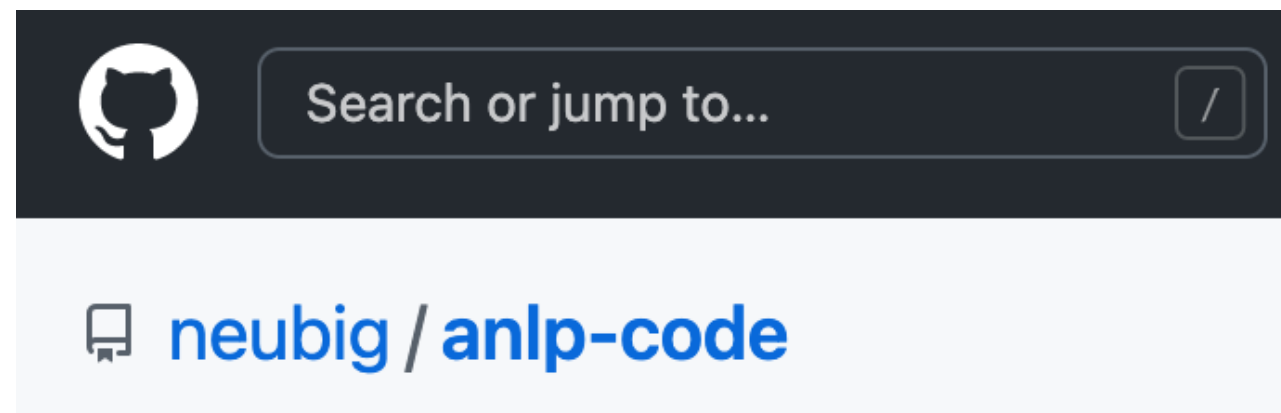
I hate this movie →
positive
neutral
negative

I love this movie →
positive
neutral
negative

I saw this movie →
positive
neutral
negative

Let's Look at Data

<https://github.com/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
 $s = \mathbf{w} \cdot \mathbf{h}$ $\mathbf{s} = W\mathbf{h}$
- **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

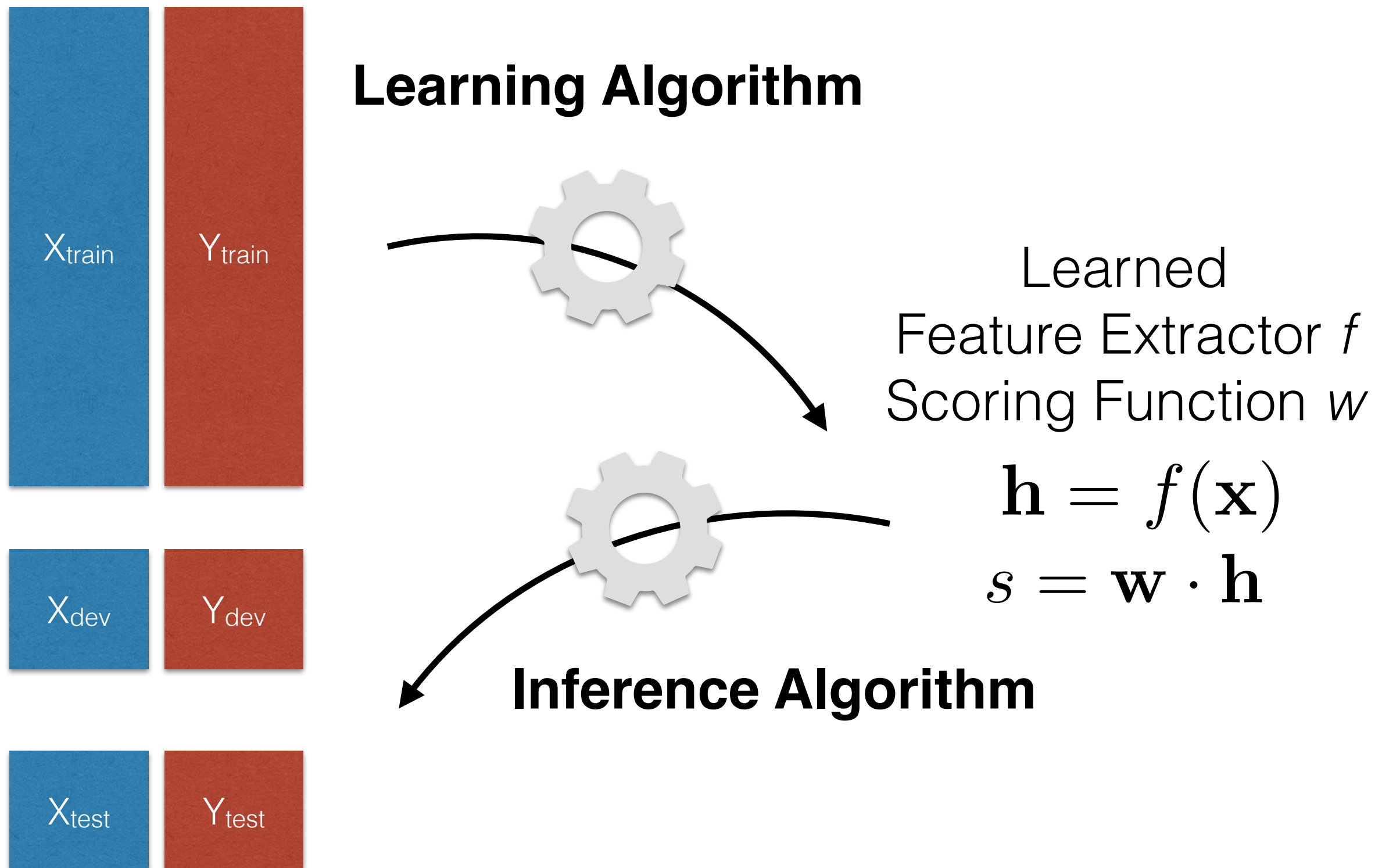
モンハンの名前がついてるからとりあえずモンハン要素を
ちょこちょこ入れればいいだろ感が凄い。

negative

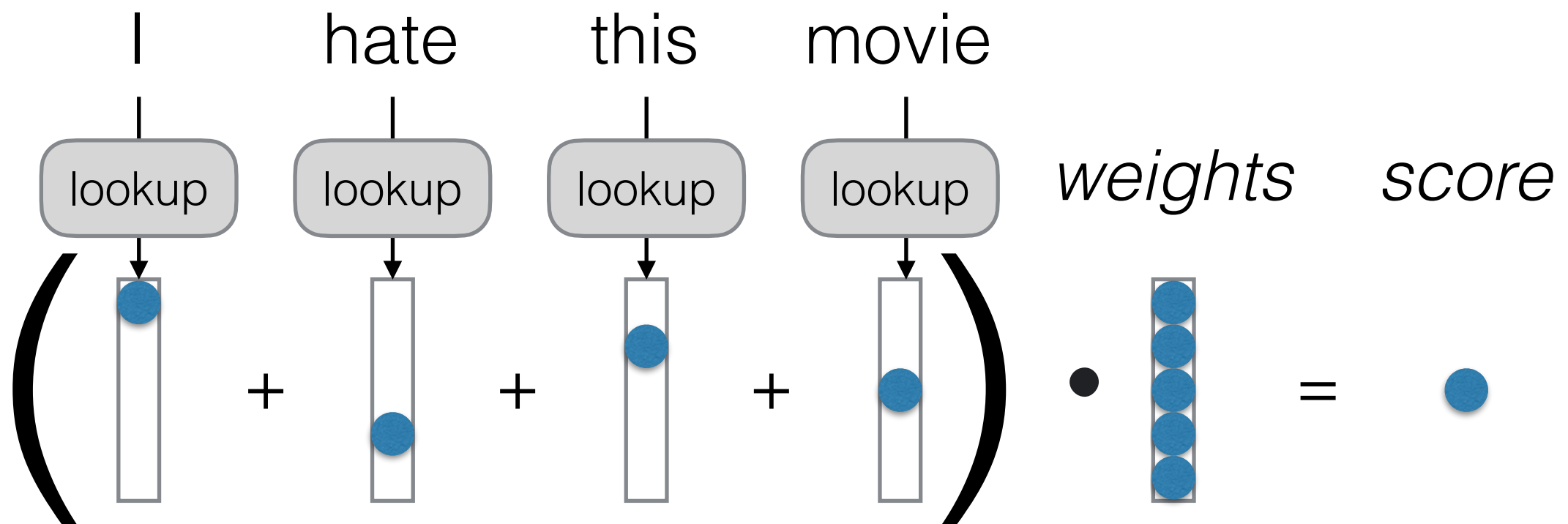
Solution?: Learn Japanese?

Machine Learning Based NLP

Machine Learning



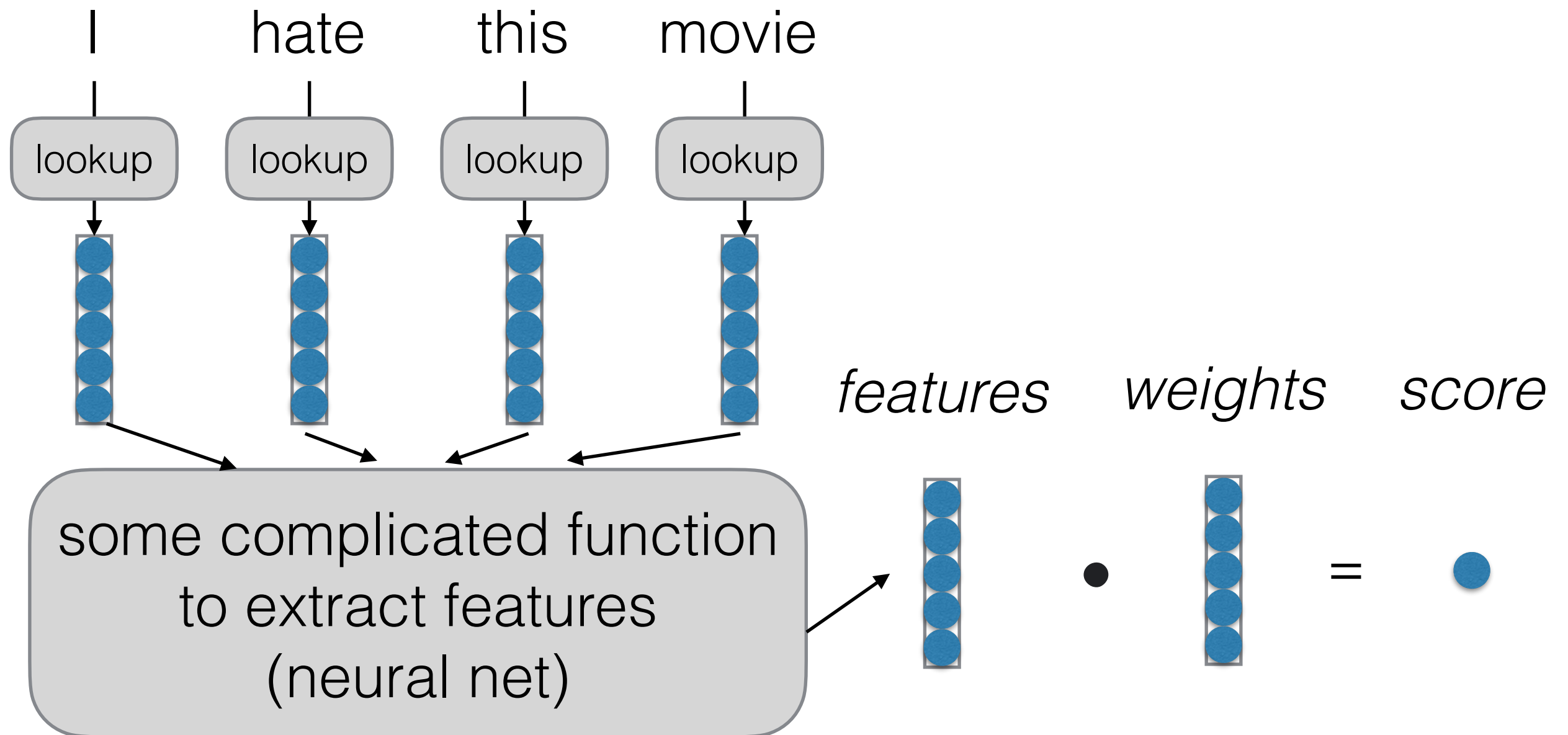
A First Attempt: Bag of Words (BOW)



Features f are based on word identity, weights w learned

Which problems mentioned before would this solve?

A Better Attempt: Neural Network Models

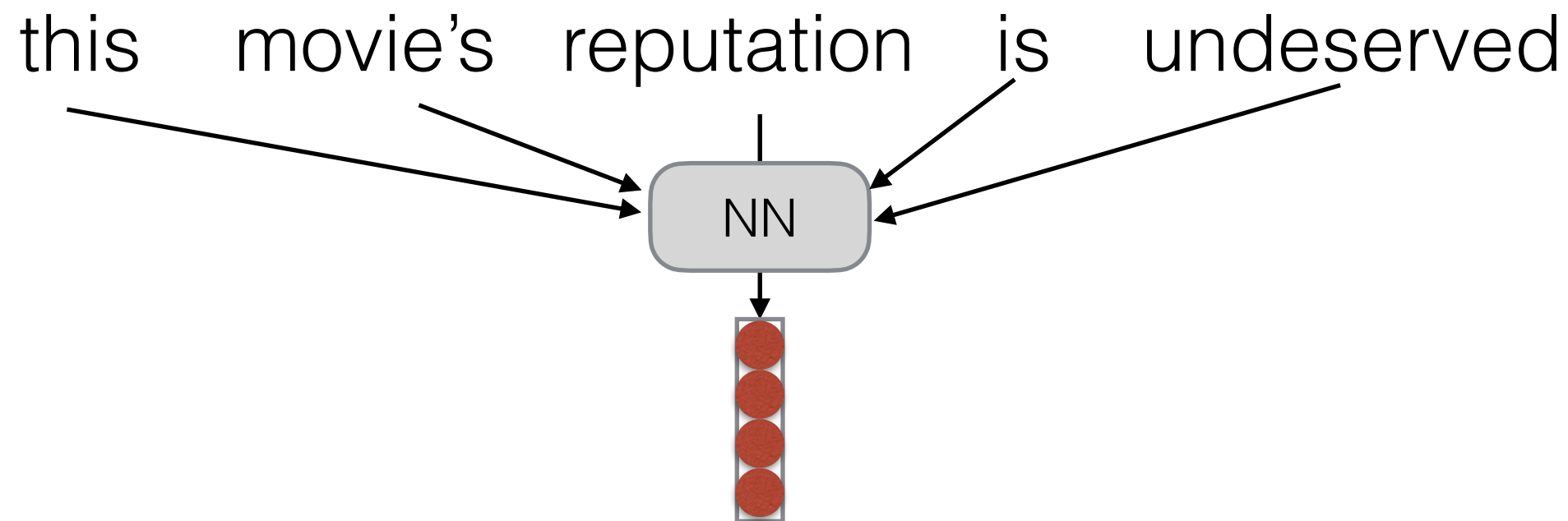


Class Goals

- 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

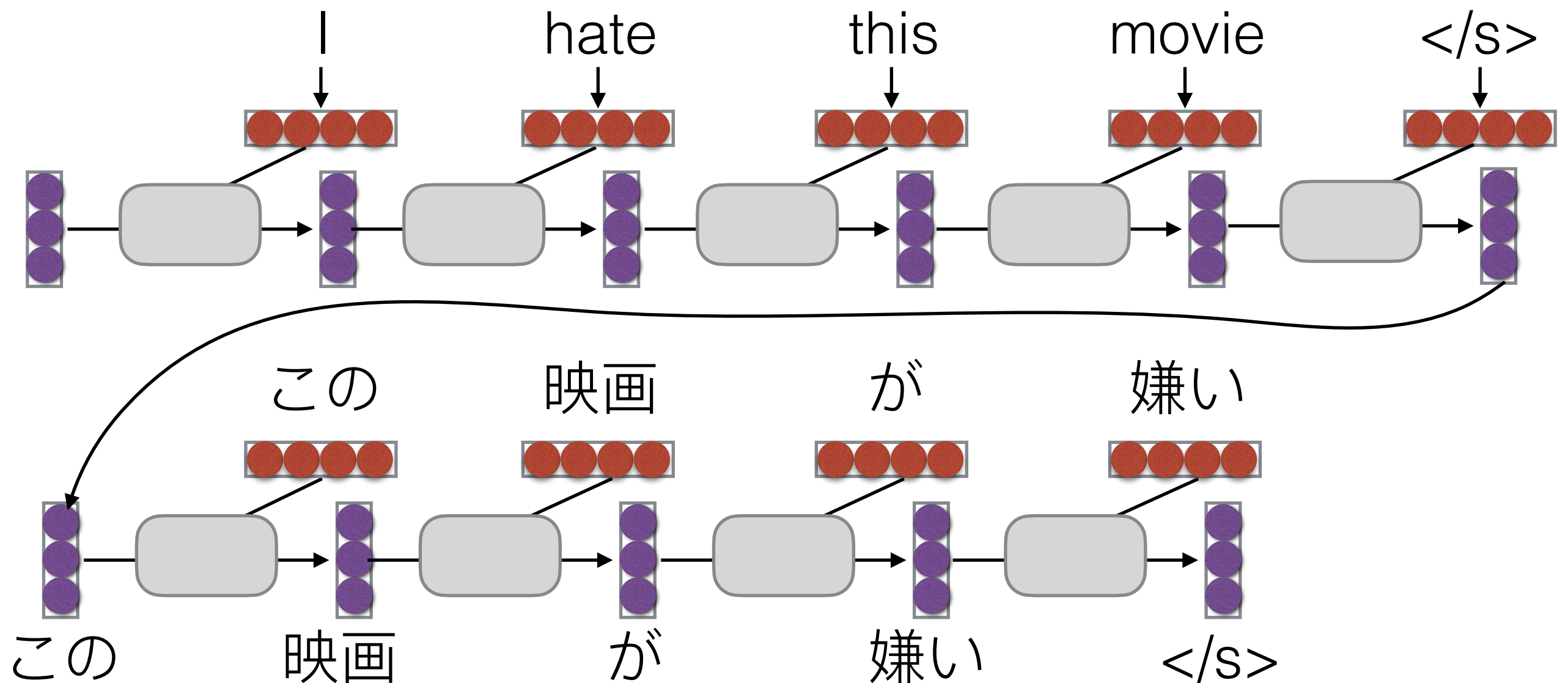
Roadmap Going Forward

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

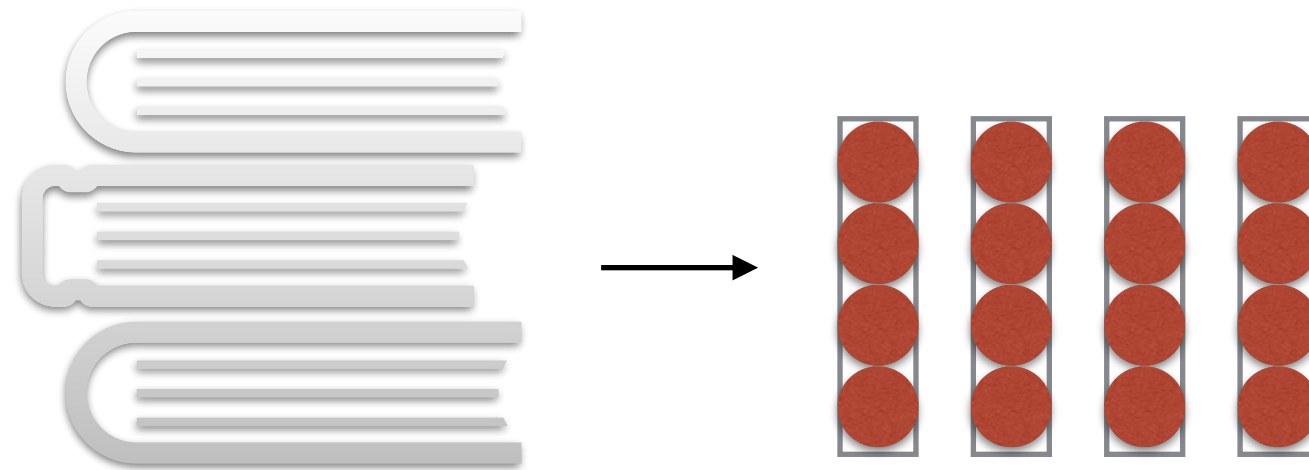
Topic 2: Sequence Models



- Recurrent Networks
- Sequence Labeling
- Conditioned Generation
- Attention

Topic 4:

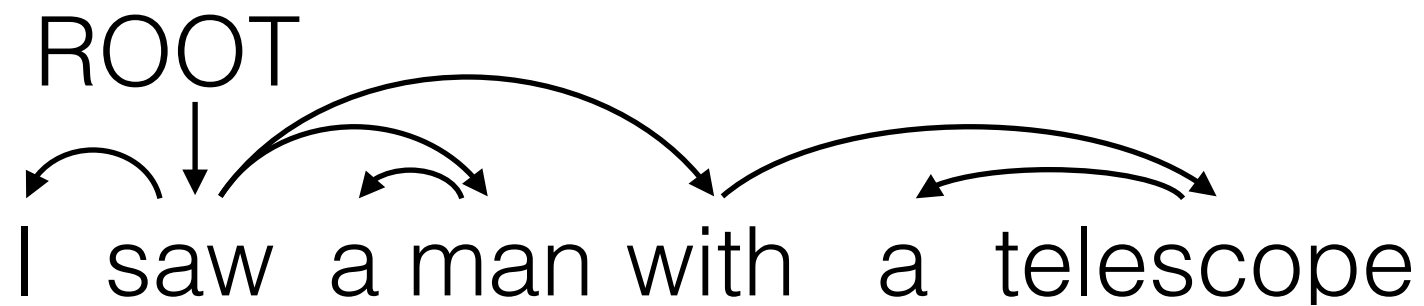
Representation and Pre-training



- Transfer Learning
- Pre-training Methods
- Sequence-to-sequence Pre-training and Prompting
- Interpreting and Debugging NLP Models

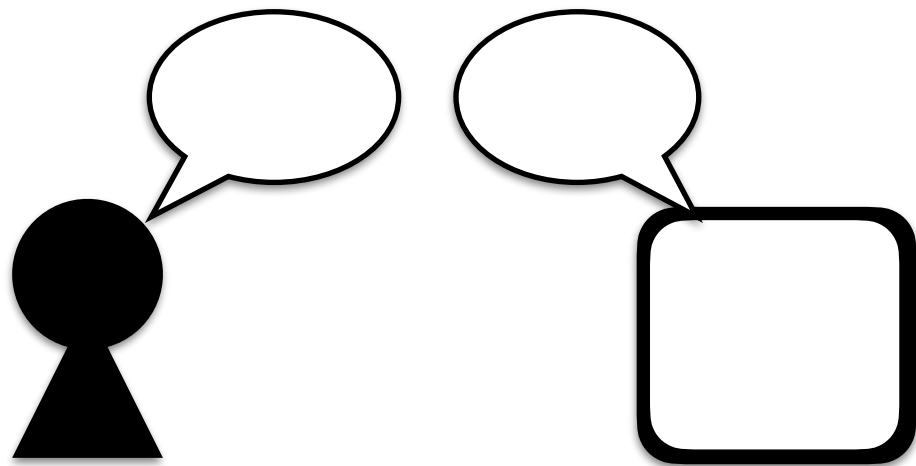
Topic 4:

Natural Language Analysis



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

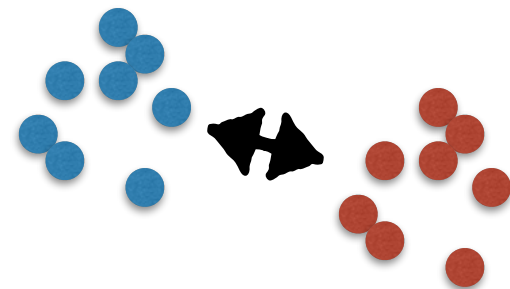
Topic 5: NLP Applications



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

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 if possible
- 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 Neural Network Toolkit:** *Individually* implement some parts of a neural network
- **Assignment 2 - Text Classifier / Questionnaire:** *Individually* implement a text classifier and fill in questionnaire on topics of interest
- **Assignment 3 - SOTA Survey / Re-implementation:** 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:**
 - Hao Zhu (pragmatics)
 - Shuyan Zhou (natural language command and control)
 - Zhisong Zhang (syntax and shallow semantic analysis)
 - Frank F. Xu (natural language to code generation)
 - Shikib Mehri (dialogue)
 - Brendon Boldt (emergent communication)
- **Piazza:** <http://piazza.com/cmu/fall2021/cs11711/home>

Thanks, Any Questions?