

CS769 Advanced NLP

Introduction to Natural Language Processing

Junjie Hu

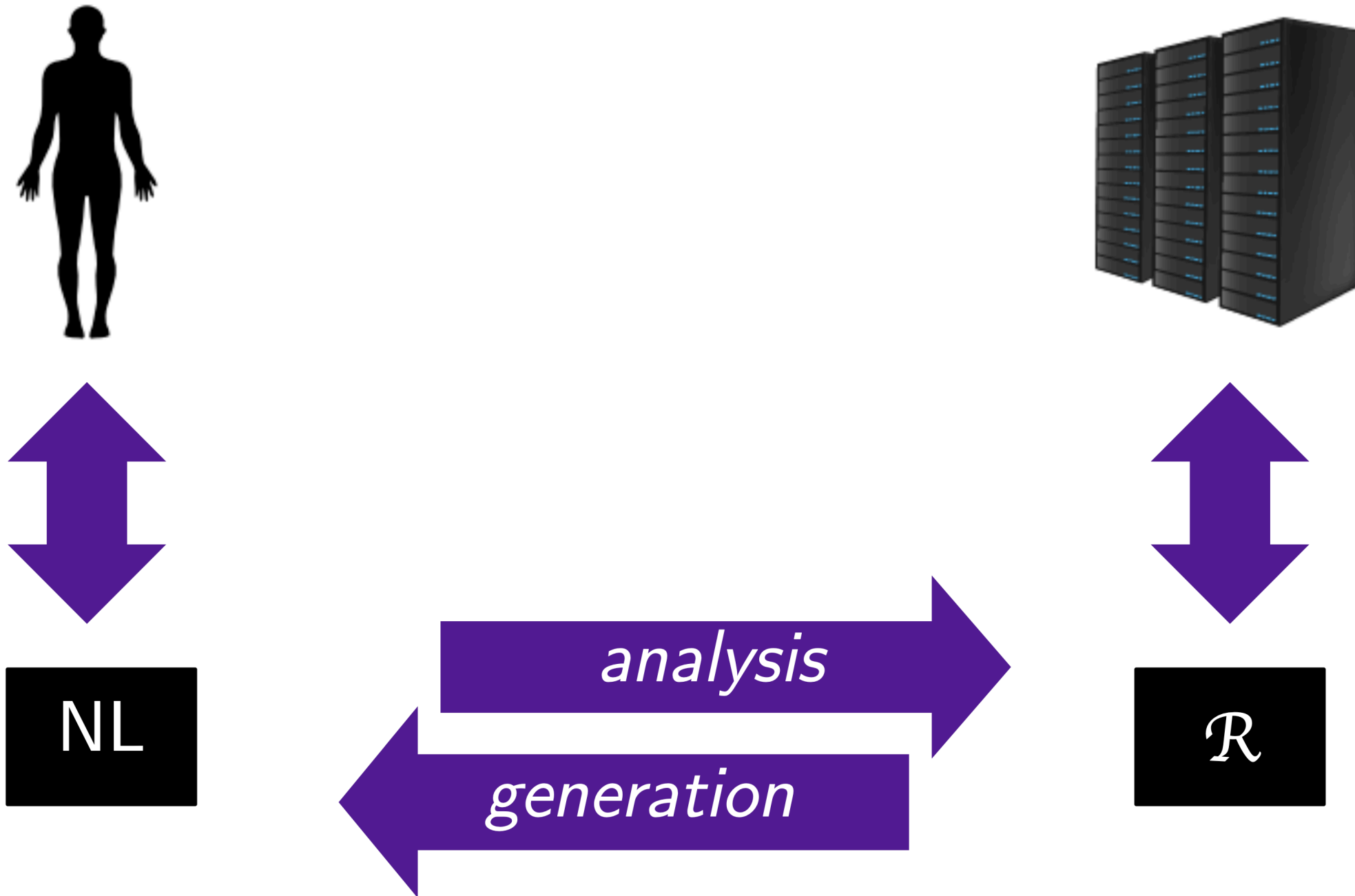


Slides adapted from Noah, Yulia, Graham
<https://junjiehu.github.io/cs769-spring22/>

What is NLP?

- $NL \in \{\text{Chinese, English, Spanish, Hindi, ...}\}$
- R : *intermediate meaning representations*
- Automation of:
 - **Analysis** or Interpretation of what a text means ($NL \rightarrow R$)
 - **Generation** of fluent, meaningful text
 - **Acquisition** of these capabilities from knowledge and data

What is NLP?

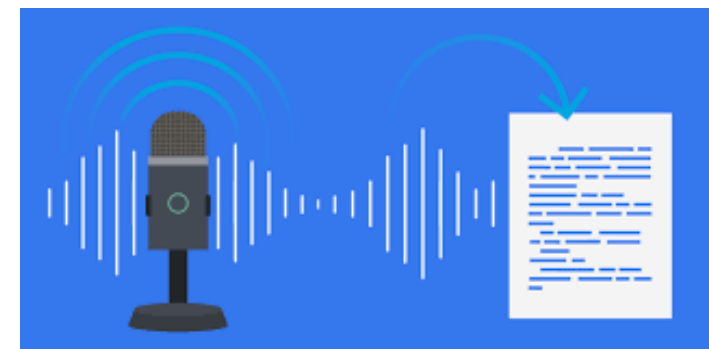
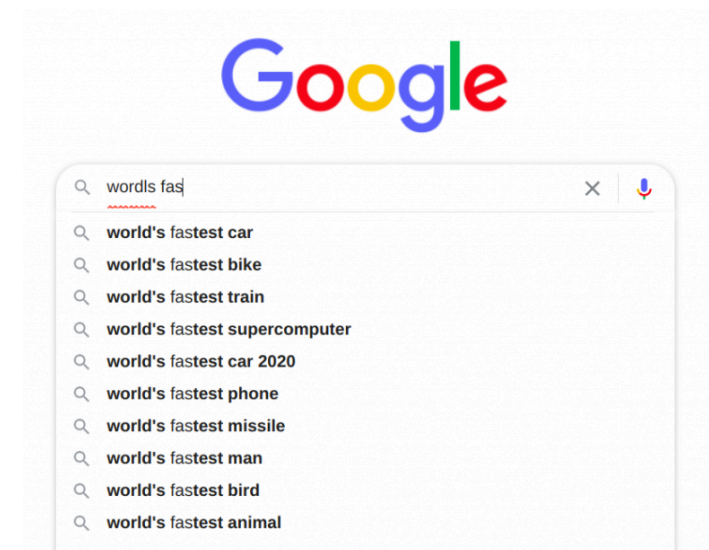
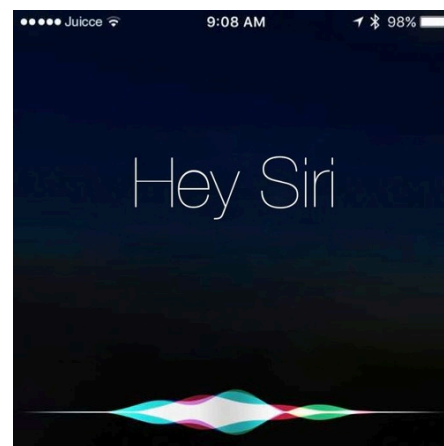
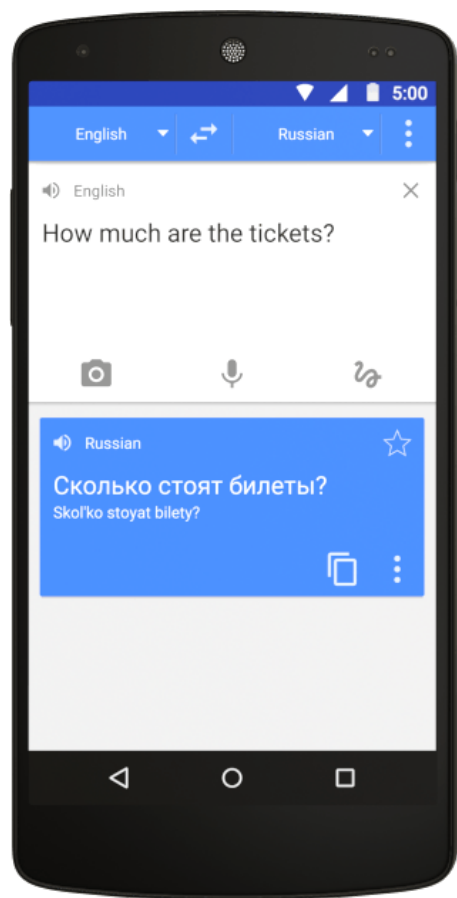


What is NLP?




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

Language Technologies

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



NLP can Answer our Questions

how many lakes does Madison have   


 All  News  Images  Maps  Shopping  More Tools

About 64,200,000 results (0.84 seconds)

five lakes

Lake Kegonsa

From fishing to watersports, runs, bike rides, or simply nature watching, each of the **five** lakes around Madison provides a different experience for tourists and locals. Nov 29, 2018

<https://www.smarttoyota.com> › [blogs](#) › [madison-wi-lakes](#) 

[Madison Lakes | Explore the Incredible Lakes of Madison, WI](#)

 About featured snippets •  Feedback

Retrieved Jan. 25, 2022

NLP can Answer our Questions

- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?



IBM Watson won Jeopardy! Champions in early 2011

NLP can Translate Text

The screenshot shows the Google Translate web interface. At the top, there are two tabs: 'Text' (selected) and 'Documents'. Below the tabs, there are language selection menus. The source language is set to 'ENGLISH' and the target language is 'CHINESE (SIMPLIFIED)'. The main content area is split into two columns. The left column contains the English text: 'Madison is the capital of the U.S. state of Wisconsin and the seat of Dane County. As of the 2020 census the population was 269,840 which made it the second-largest city in Wisconsin by population, after Milwaukee, and the 80th-largest in the United States. The city forms the core of the Madison Metropolitan Area which includes Dane County and neighboring Iowa, Green, and Columbia counties for a population of 680,796. Madison is named for American Founding Father and President James Madison.' The right column contains the Chinese translation: '麦迪逊是美国威斯康星州的首府，也是戴恩县的所在地。截至 2020 年人口普查，人口为 269,840，使其成为威斯康星州人口第二大城市，仅次于密尔沃基，在美国排名第 80。该市是麦迪逊都会区的核心，包括戴恩县和邻近的爱荷华县、格林县和哥伦比亚县，人口达 680,796 人。麦迪逊以美国国父和总统詹姆斯麦迪逊的名字命名。' Below the Chinese text is the pinyin: 'Màidí xùn shì měiguó wēisīkāngxīng zhōu de shǒufǔ, yěshì dài ēn xiàn de suǒzàidì. Jiézhì 2020 nián rénkǒu pǔchá, rénkǒu wèi 269,840, shǐ qí chéngwéi wēisīkāngxīng zhōu'. At the bottom of the interface, there are icons for voice input/output, a character count '496 / 5,000', and a 'Send feedback' link.

English Wikipedia, translated by Google Jan. 25, 2022

NLP can Translate Text

The collage illustrates the application of NLP in translation. It features three overlapping screenshots of the City of Madison website, each translated into a different language: Chinese (Simplified), Japanese, and Portuguese. The Chinese and Japanese versions show the website's navigation and content translated into their respective languages. The Portuguese version shows a sign that reads "ESTRADA DESMORONADA" and "DESVIO" being translated by a smartphone app into "ROAD COLLAPSED" and "DIVERSION 2km".

City Highlights

[Remember: Make a Work Order for Large Pickup](#)

During most weeks, residents should see week opportunities to set out large items when fillin work order.

[Services to Undocumented Immigrants](#)

The City of Madison is seeking to fund proposa 2022 that support community partners in their delivering direct assistance to undocumented individuals and families to help them cope wit consequences of the pandemic.

[Community Resources](#)

Find resources for food, housing, utilities, and

城市亮点

[记住：为大件物品提货下一个工单](#)

在大多数情况下，居民在填写工作订单时应该每周都有机会放置大件物品。

[无证移民服务 RFP](#)

麦迪逊市正在寻求资助 2022 年的提案，以支持社区合作伙伴的工作，为无证个人和家庭提供直接援助，以帮助他们应对大流行的后果。

[社区资源](#)

查找食物、住房、公用事业等资源。

シティハイライ

[覚えておいてください](#)

ほとんどの週の間、居住るときに大きなアイテムを置く機会を見ることがあります。

[文書化されていない](#)

マディソン市は、パンデミックのために、文書化されていない提案を提供するコミュニティパートナーを探しています。

[コミュニティリソース](#)

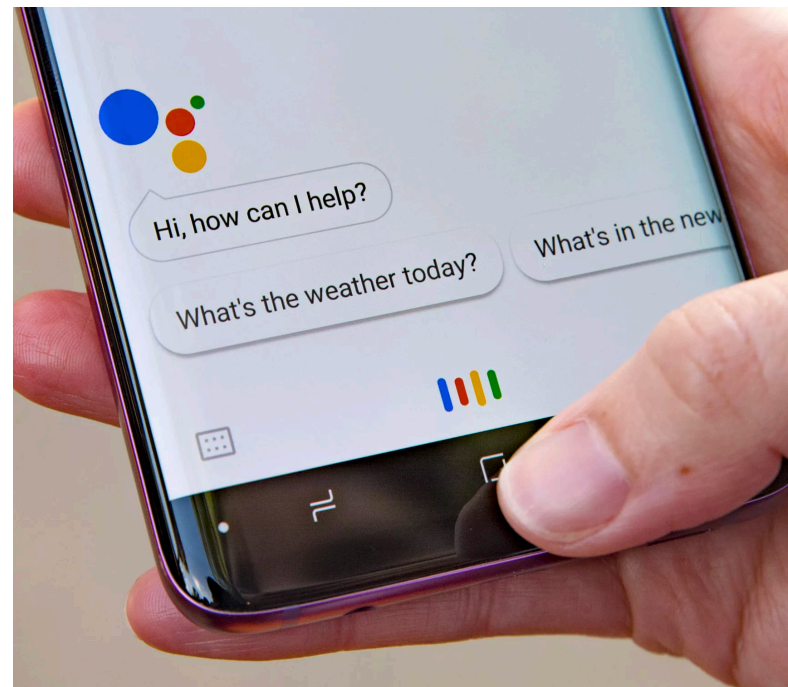
食料、住宅、ユーティリティリソースを探索します。

Translate webpages
or even images!!

NLP can Chat with You

Conversational agents:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech




works with the
Google Assistant



I just try to be the best me I can be

am I smart

You're as smart as Grace Hopper. She invented the first ever computer 🖥️



NLP

- **Applications**

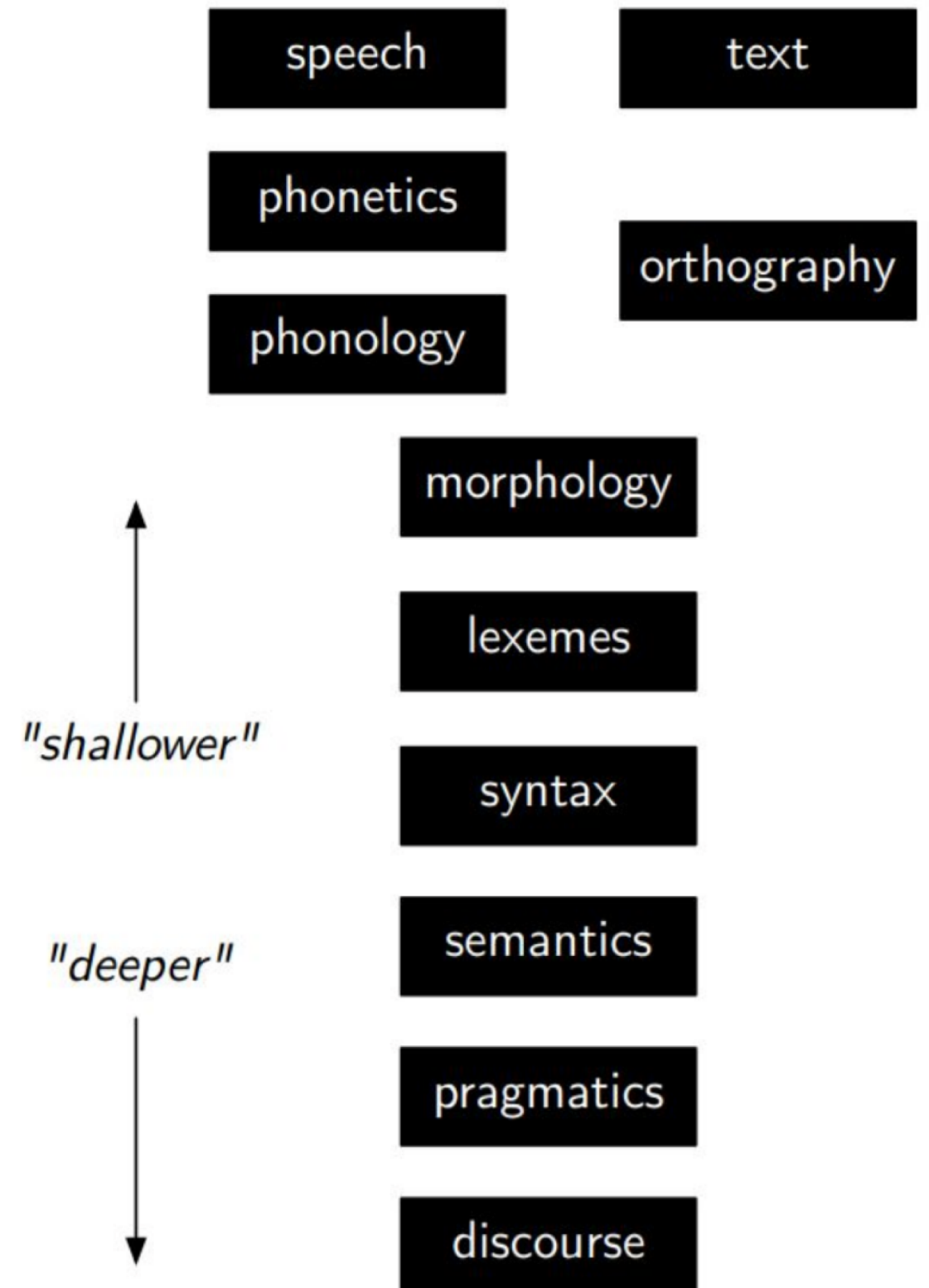
- Machine translation
- Information retrieval
- Question answering
- Dialogue systems
- Information extraction
- Summarization
- Sentiment analysis
- ...

- **Core technologies**

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic role labelling
- ...

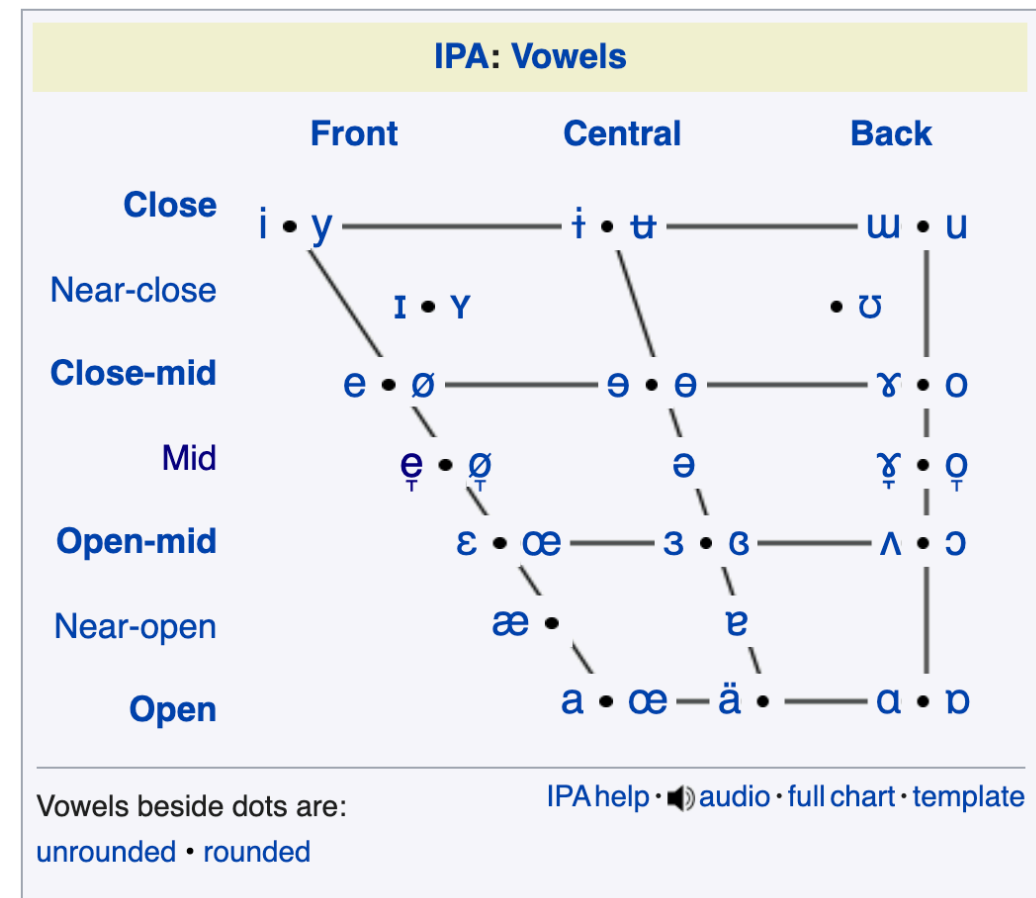
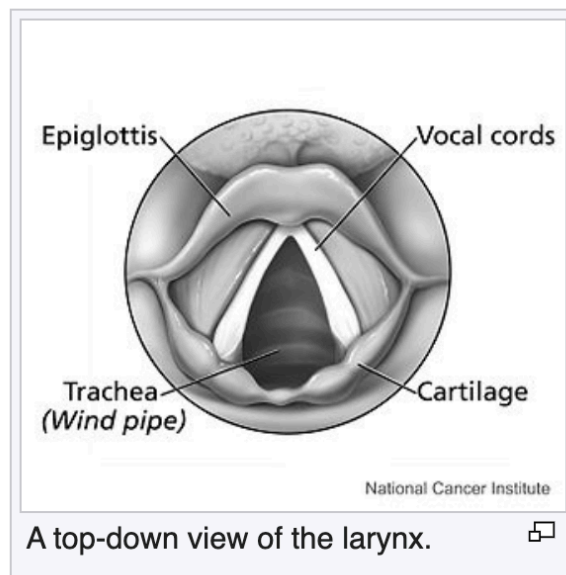
Levels of Linguistic Knowledge

- What dose an NLP system need to “know” a language?



Phonetics, Phonology

- Study how humans produce and perceive **sounds**, or in the case of **sign languages**, the equivalent aspects of **sign**



Orthography (Writing Systems)

- Thai script:
 - ลูกศิษย์วัดกระตังยั้งยื่นปิดถนนทางขึ้นไปนมัสการพระบาทเจ้าศิษย์คุณหวัดปะทะกับเจ้าถิ่นที่ออกมาเผชิญหน้าเพราะเดือดร้อนสัญจรไม่ได้ ผวจ.เร่งทุกฝ่ายเจรจา ก่อนที่ชื่อเสียงของจังหวัดจะเสียหายไปมากกว่านี้ พร้อมเสนอหยุดจัดงาน 15 วัน....
- Latin script:
 - The Latin script, also known as Roman script, is an alphabetic writing system based on the letters of the classical Latin alphabet.
- Arabic script:
 - لم تعترف منظمة الأمم المتحدة باللغة العربية رسمياً إلا في 18 ديسمبر عام 1973، بعد محاولات مضنية من قادة الدول العربية للاعتراف باللغة العربية داخل المنظمة الأممية الكبيرة منذ تأسيسها عام 1945 وحتى تاريخ الاعتراف طيلة السنوات التي لم تكن الأمم المتحدة اعترفت باللغة العربية رسمياً، كان رؤساء الدول العربية يتحدثون اللغة العربية مع حضور مترجم، وكان أول رئيس يقوم بإلقاء خطاب سياسي قبل قرار الاعتراف هو رئيس جمهورية مصر العربية جمال عبد الناصر

.. المزيد على دنيا الوطن

Morphology (Assembly of Words)

- Study of how words are formed: such as stems, root words, prefixes, suffixes
 - [Turkish]: uygarla_,stiramadıklarımızdanmı_,ssinizcasına ← **agglutinative** language
 - [English]: “(behaving) as if you are among those whom we could not civilize”
 - [English]: unfriend → un + friend , Obamacare → Obama + care

- **WORDS** This is a simple sentence
MORPHOLOGY be
3sg
present

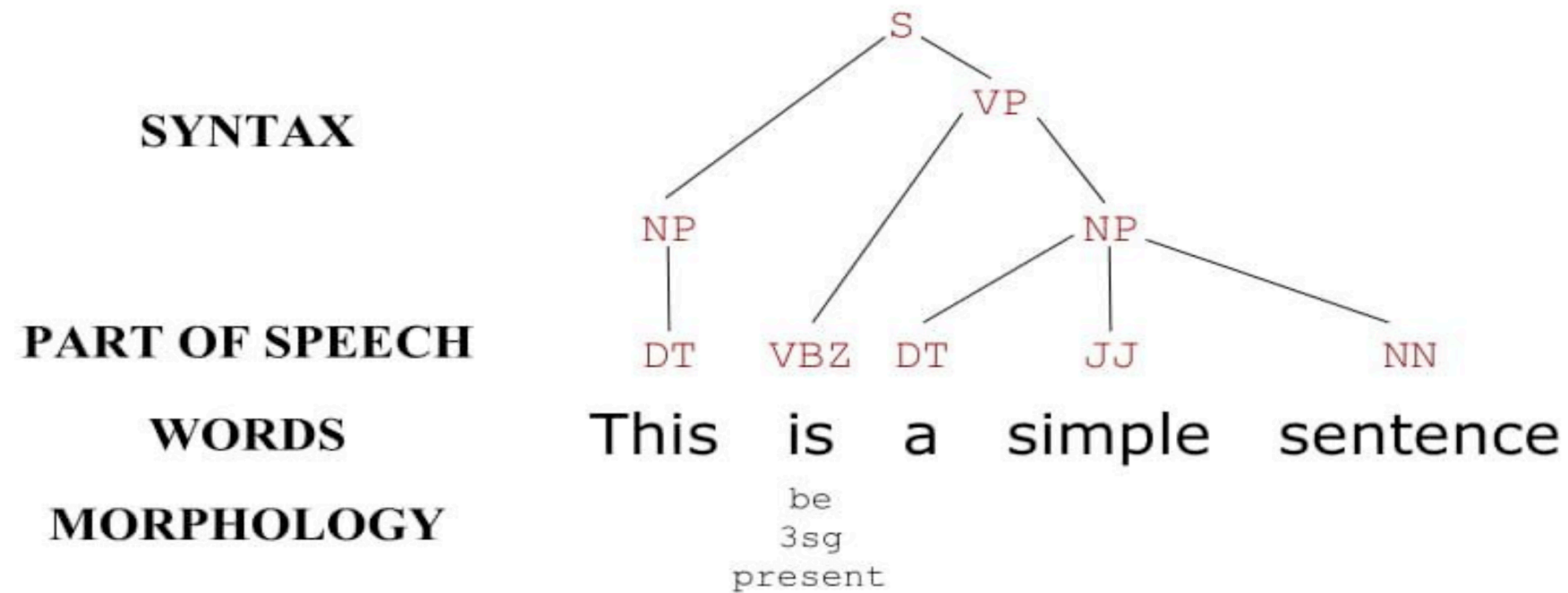
Part-of-Speech

- Predict which category a word is assigned to in accordance with its syntactic functions.

PART OF SPEECH	DT	VBZ	DT	JJ	NN
WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

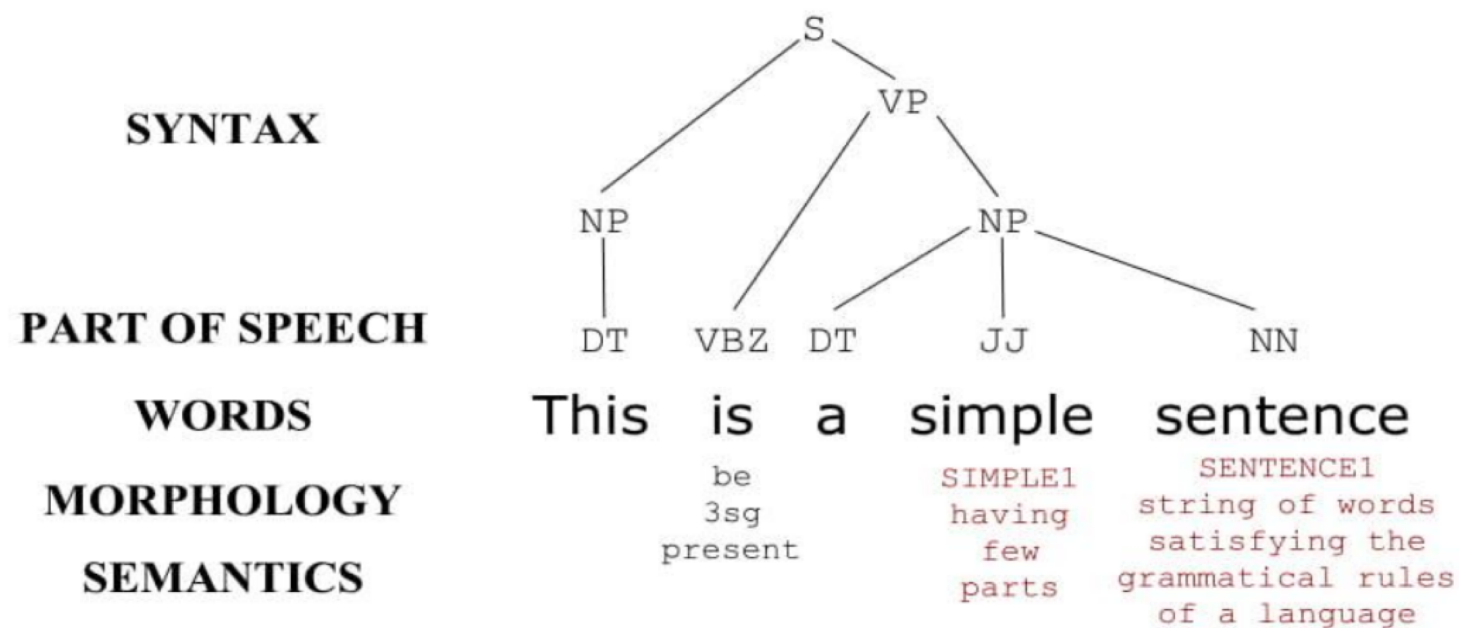
Syntax

- Study of how words and **morphemes** combine to form larger units such as **phrases** and **sentences**.
 - Constituency Grammars
 - Dependency Grammars



Semantics

- Study meaning of words, phrases, sentences, or larger units (w/ discourse)
 - Named entity recognition
 - Word sense disambiguation
 - Semantic role labeling



Discourse

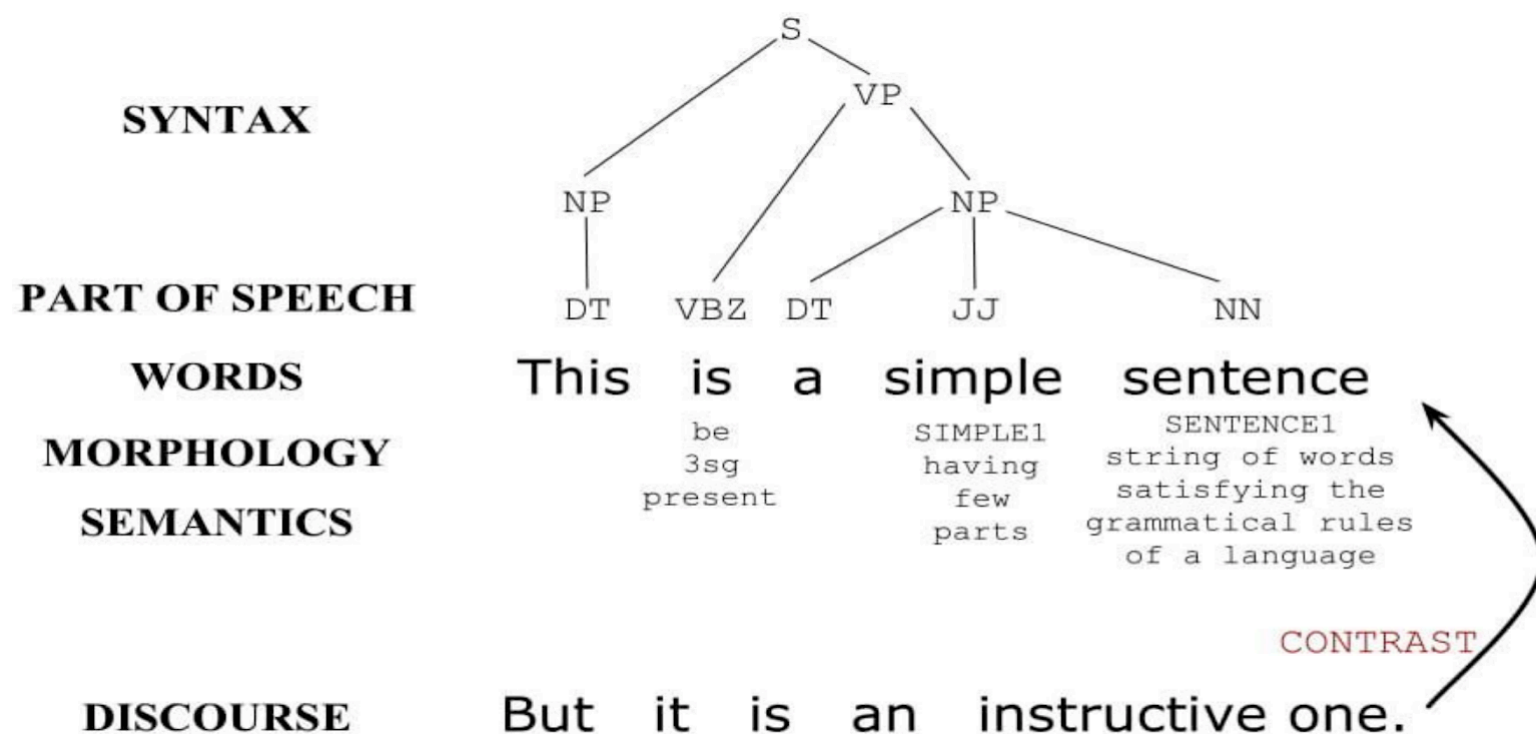
- Analysis of language “beyond the sentence”

<> analysis of sounds (phonetics)

<> analysis of words (morphology)

<> analysis of meaning (semantics)

<> analysis of word order (syntax)



Where are we now for NLP
research?

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...

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

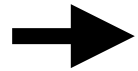
NLP cannot Answer our Questions

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 are navigation tabs for "All", "News", "Images", "Videos", "Shopping", and "More", 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 text "who invented neural machine" in English (detected) and the translated text "qui a inventé la machine neuronale" in French. The widget also includes a swap button, a close button, and icons for voice input and output. At the bottom of the widget, there are links for "Open in Google Translate" and "Feedback".

Retrieved Aug. 29, 2021

NLP cannot Translate Text

“၃၇၊ ၃၈ မဒ်လေးမိၼ်ရဲၼ် ၈၄ မိနူးလမူးမဆကီးပေါ့၊ စုတုနူးပဲ
ရွံသေးတယုပေါ့နော့ ထြကိၼ်အတြကု၊ အဲဒါကို သူတိုၼ်တြ ဘယု
ကေန ဆကီတင သတငူးရလာတယု မသိဘူး၊ ခ်က္ခင်း ရောက္ခလာမ
ပီးတော့ အဟုကမူးဖကု ဝငရောက္ခမိခြင်းတာပေါ့။ အတိအက်
တာ ကြးနော့တာတိုၼ်လဲ မသိရေသးဘူး။ ၄ ဝေယာက္ခိဩားတယု
လိုၼ်လဲ ဝေယာက္ခိဩားတယု။ ၆ ဝေယာက္ခိဩားတယုလဲဝေယာက္ခိဩားတယု။ ဘ
ယုလောက္ခိဩားလဲဆိုတာ ခုအခိန္တိ အတိအက် မသိရေသးဘူး။ တ
ခိဩားတြဆို ရောက္ခတာဝင မေရာက္ခကေသးဘူး။ စစောက္ခိဩား
အဟုကမူးဖကု ဖမိခြင်းလိုက္ခတော့ လက္ခိဩားတော့ အမ္မတု ၆ ထဲမှာ ဖ
မူးခိဩားရတယုၼ် အဲလောက္ခိဩား သိရေသးတယု။”



"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.

Annotations from Stanford CoreNLP: 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?**

Why NLP is Hard?

- Ambiguity
- Scale
- Sparsity
- Variation
- Expressivity
- Unmodeled variables
- Unknown representations \mathbb{R}

Ambiguity

- Ambiguity at multiple levels:
 - Words with multiple meanings: *bank* (finance or river?)
 - Domain-specific meanings: *latex*
 - Part-of-speech: *chair* (noun or verb?)
 - Multiple meanings: *I made her duck.* →
 - I cooked waterfowl for her
 - I cooked waterfowl belonging to her
 - I created the (plaster?) duck she owns
 - I caused her to quickly lower her head or body

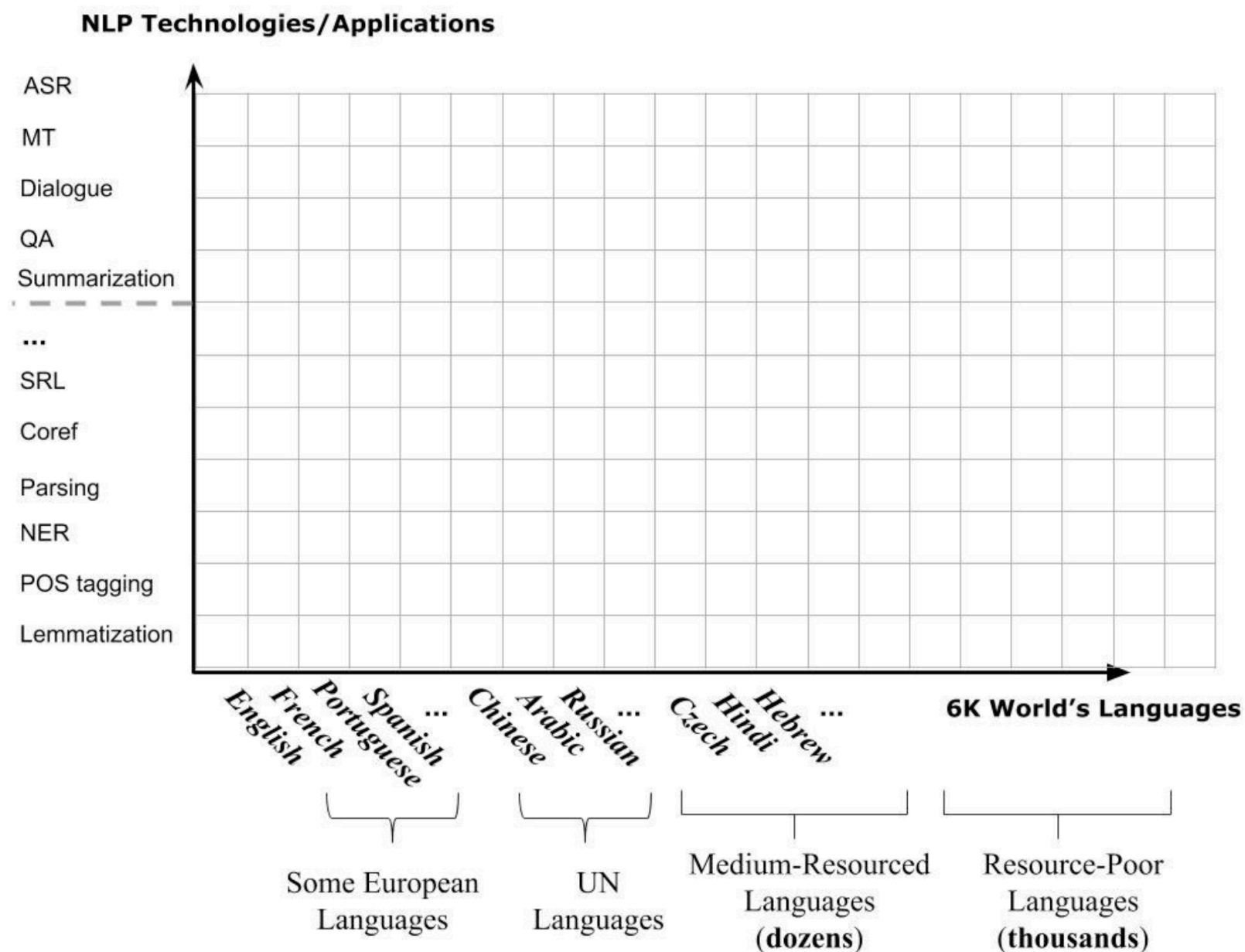


More Challenges of “Words”

- Segmenting text into words (e.g., Thai example)
- Morphological variation (e.g., Turkish example)
- Multiword expressions: *take out, make up*
- New words (e.g., *covid*) and changing meanings (e.g., *Bachelor*: a young knight → an academic degree)

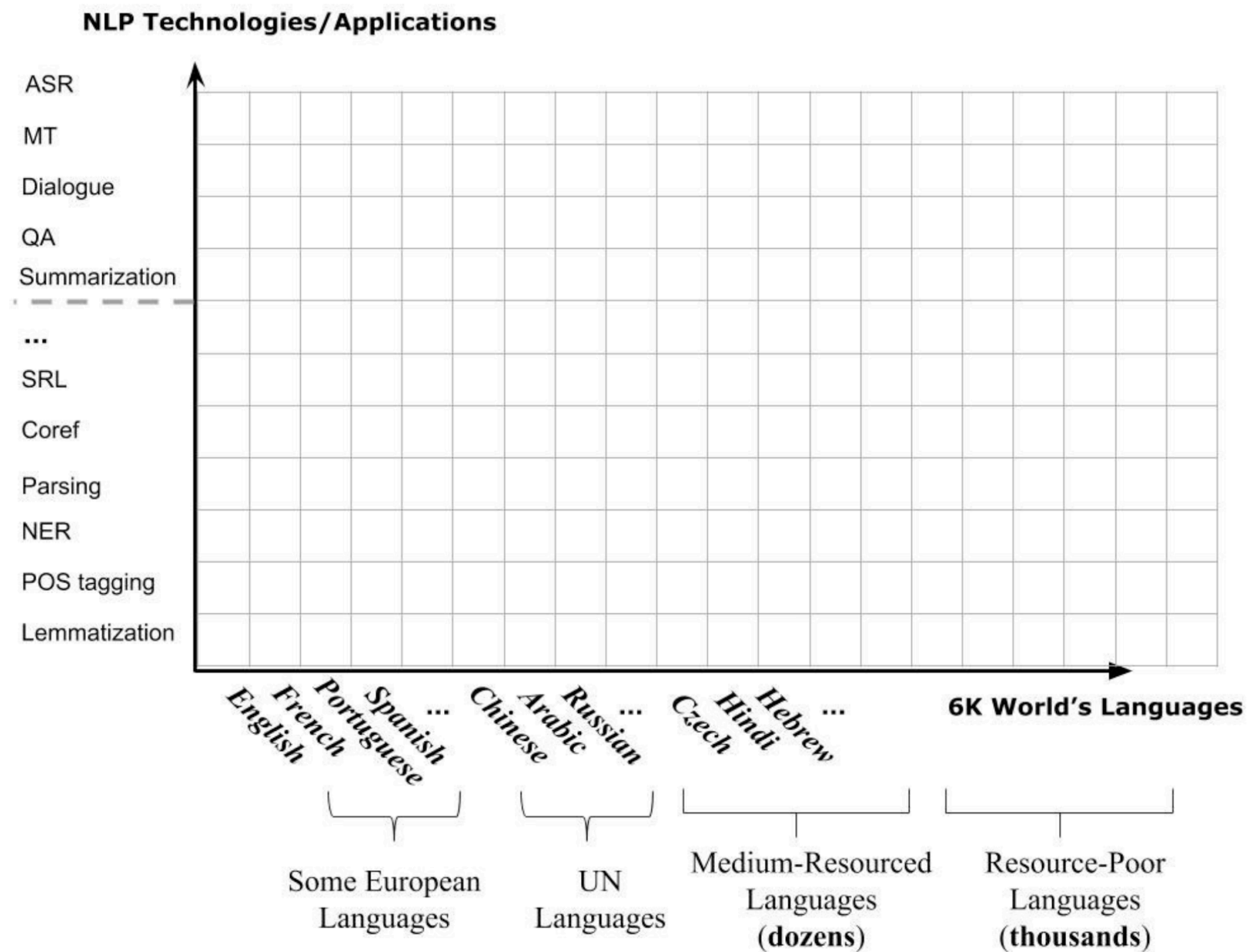
Ambiguity + Scale

- Scale up to different **languages** & **tasks**.

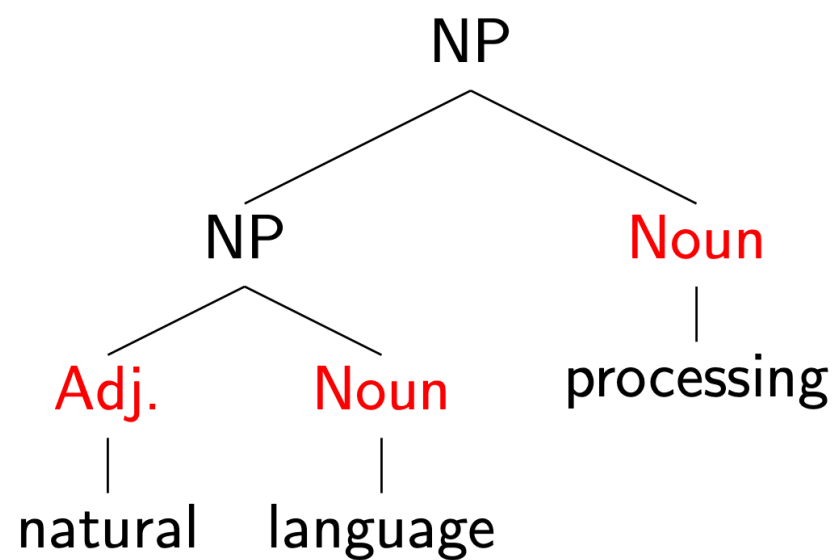


Scale

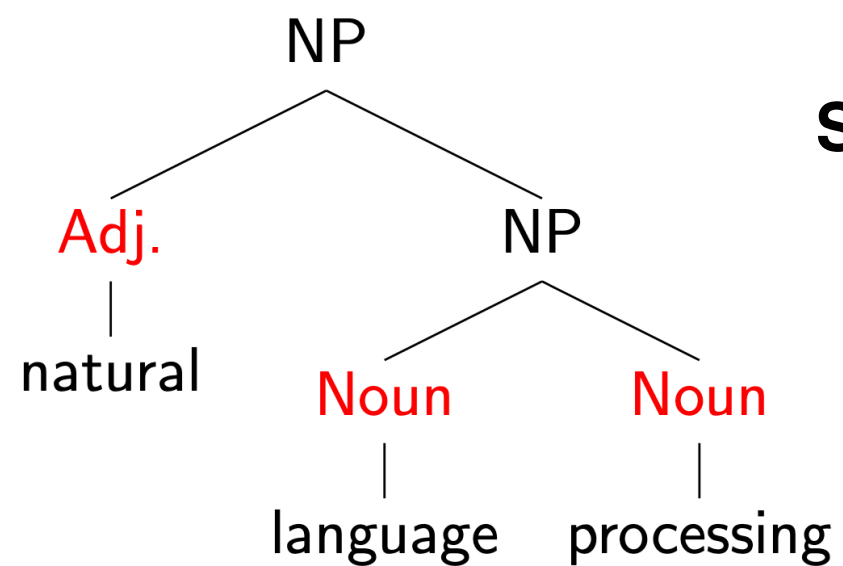
- Scale up to different **languages** & **tasks**.



Syntax Ambiguity



vs.



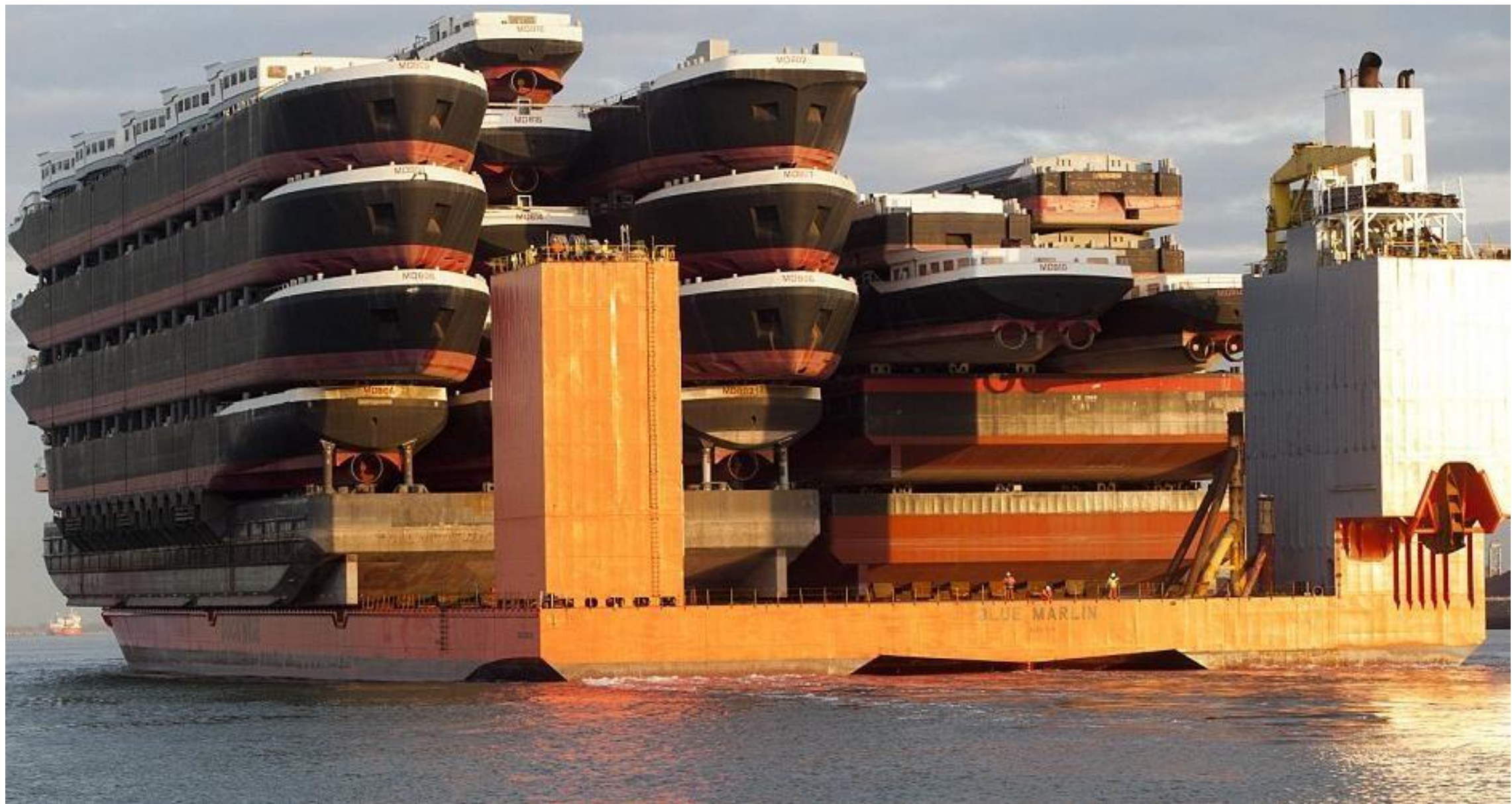
Syntactic parsing

Part-of-Speech

Words

Morphology + Syntax

- A ship-shipping ship, shipping shipping-ships



Syntax + Semantic

We saw the woman with the telescope wrapped in paper.

- Who has the telescope?
- Who or what is wrapped in paper?
- An event of perception, or an assault?

Semantic Ambiguity

- Every fifteen minutes a woman in this country gives birth.

Semantic Ambiguity

- Every fifteen minutes a woman in this country gives birth. Our job is to find this woman, and stop her!

- Groucho Marx

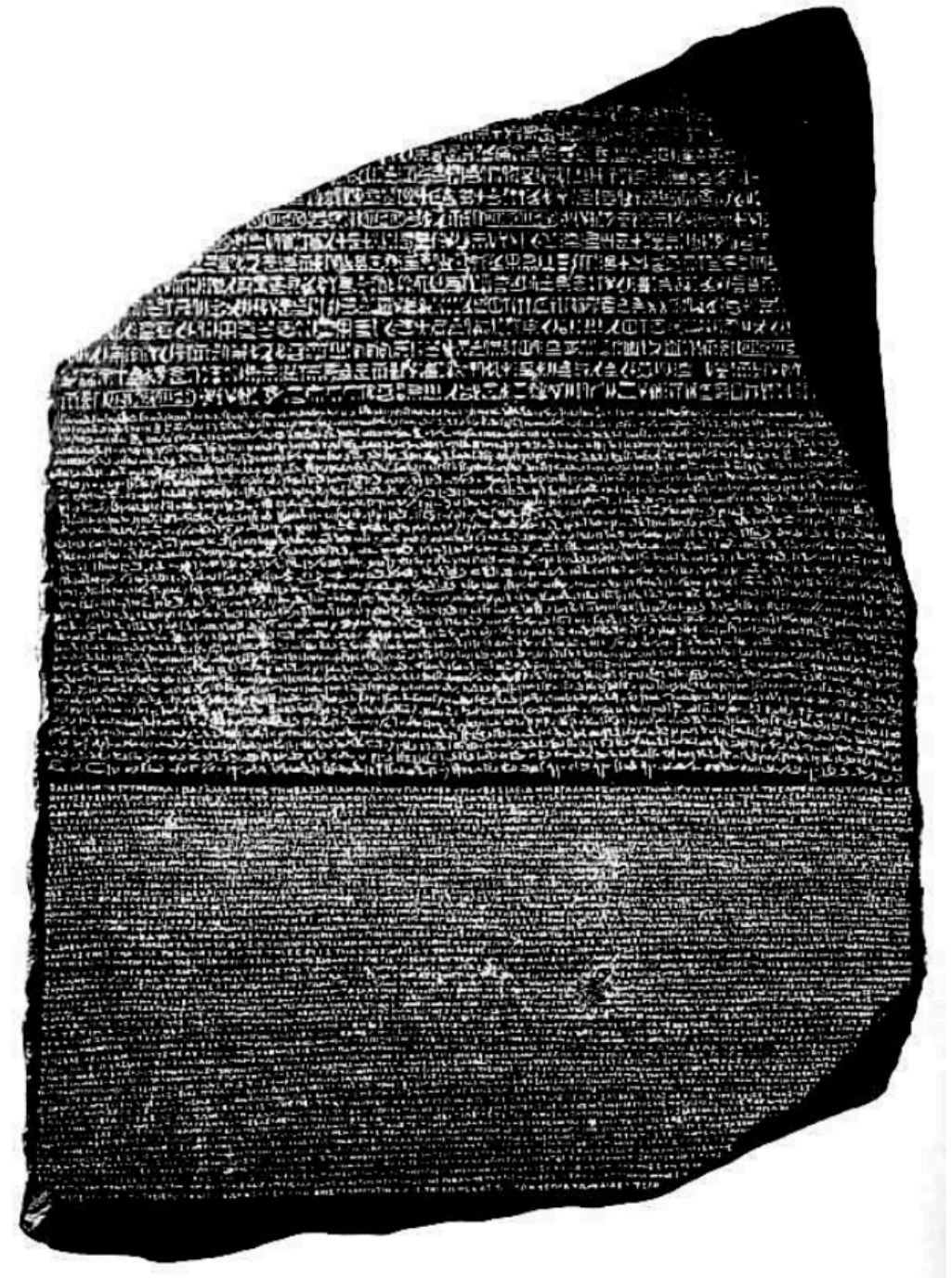


Dealing with Ambiguity

- How can we **model ambiguity** and **choose the correct analysis** in context?
 - Non-probabilistic methods (Finite-state machines for morphology, CKY parsers for syntax) return *all possible analyses*.
 - Probabilistic models (HMMs for POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return *the best possible analysis*
- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?

Corpora

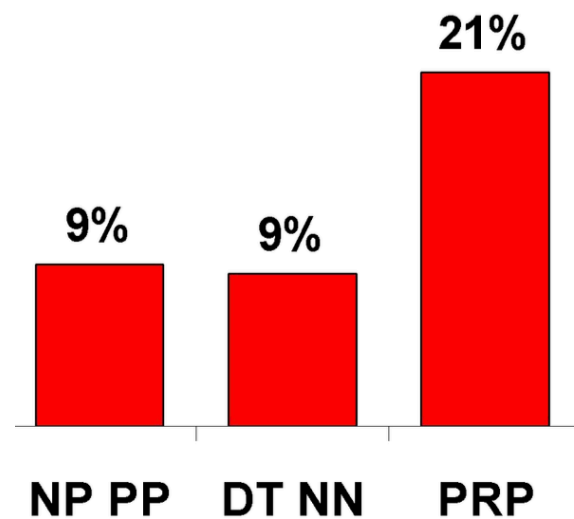
- **A corpus is a collection of text**
 - Often annotated in some way
 - Sometimes just lots of text
- **Examples**
 - Penn Treebank: 1M words of parsed WSJ
 - Canadian Hansards: 10M+ words of aligned French/English sentences
 - Web: billions of words
 - Amazon reviews



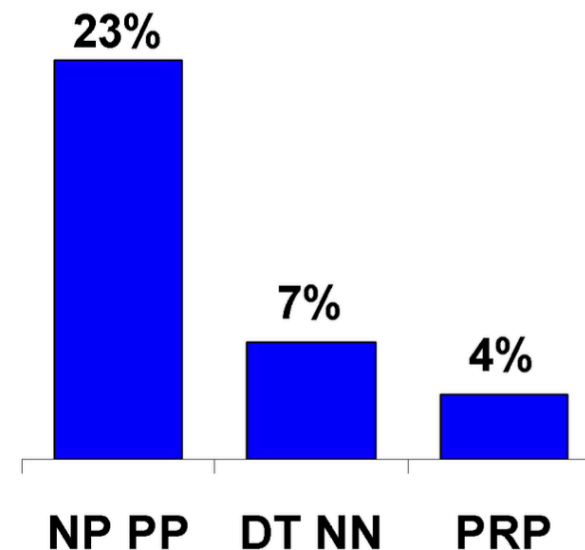
Corpus-based Methods

- Give us statistical information by counting
 - Example: Syntax parsing

NPs under S



NPs under VP



Statistical NLP

- Like most other parts of AI, NLP is dominated by statistical methods
 - Typically more robust than earlier rule-based methods
 - Relevant statistics/probabilities are *learned from data*
 - Normally requires lots of data about any particular phenomenon

Statistical NLP

- Sparse data due to *Zipf's Law*
 - To illustrate, let's look at the frequencies of different words in a large text corpus
 - Assume “word” is a string of letters separated by spaces

Statistical NLP

- Most frequent words in the English Europarl corpus (out of 24m word tokens)

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

Word Counts: Raw Words

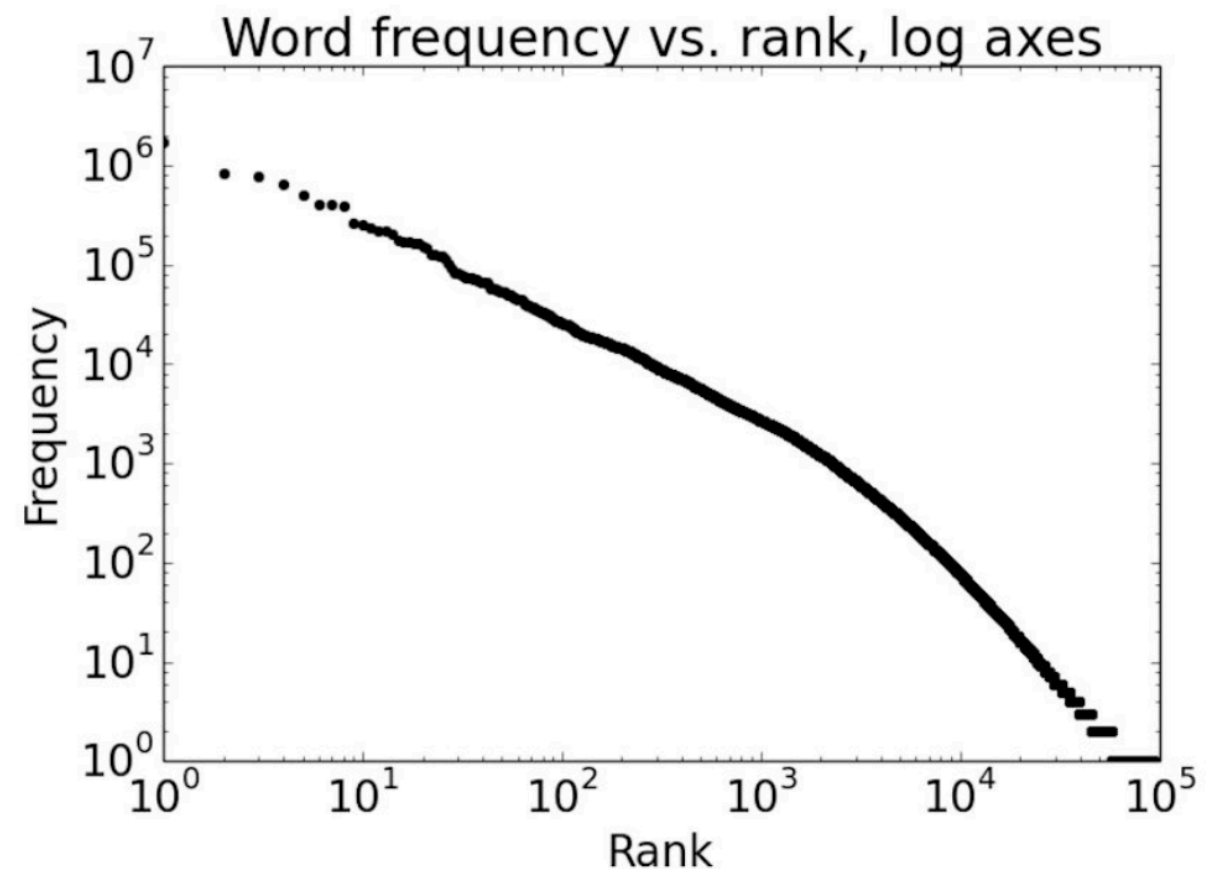
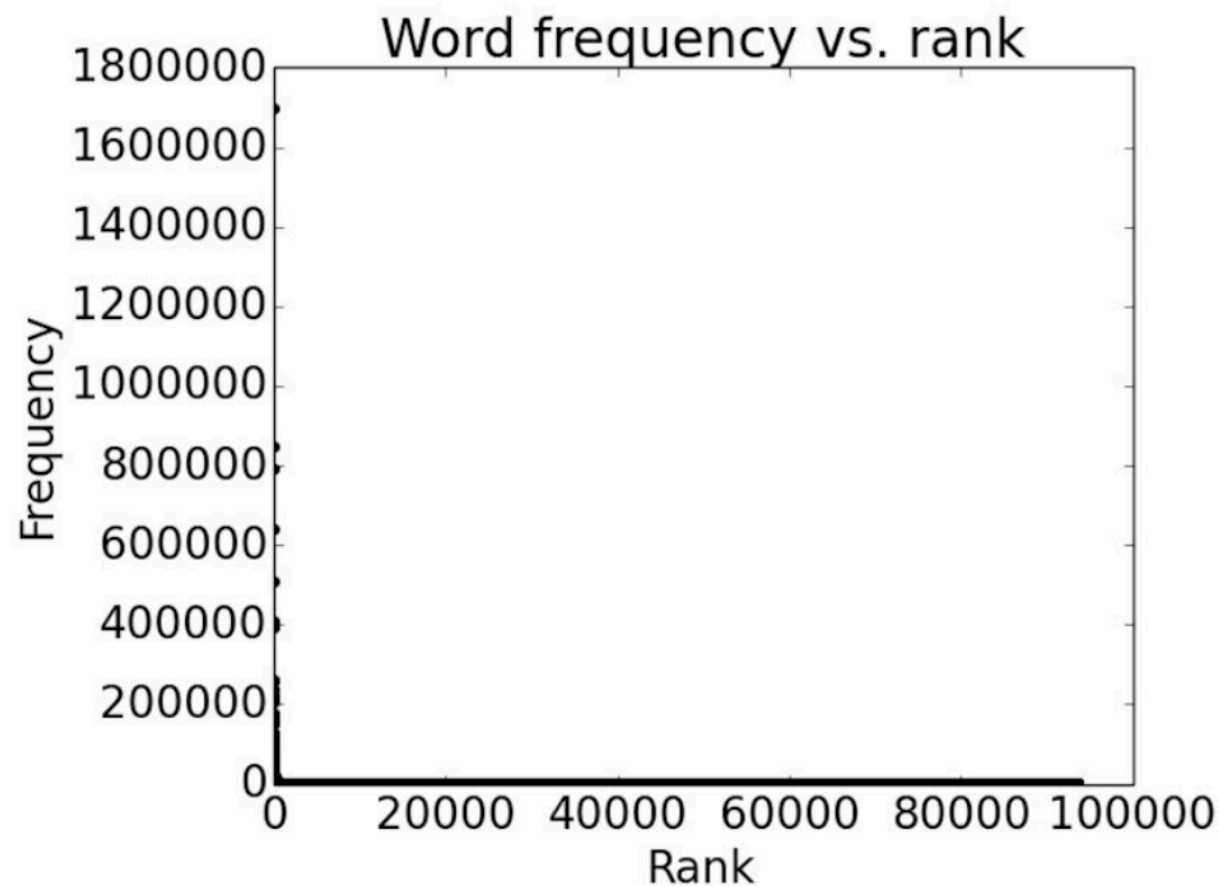
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

Plotting Word Frequencies

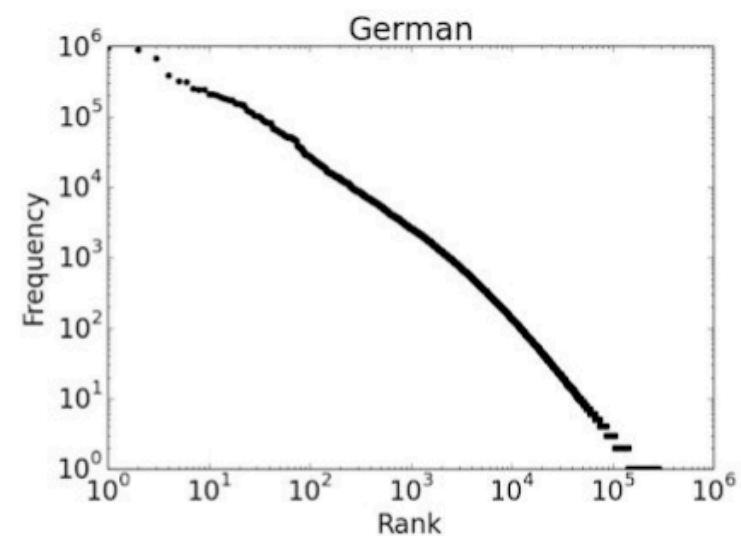
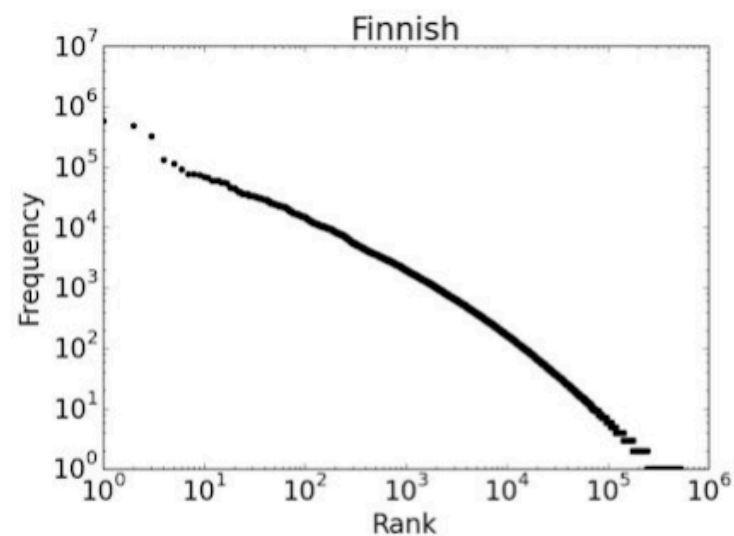
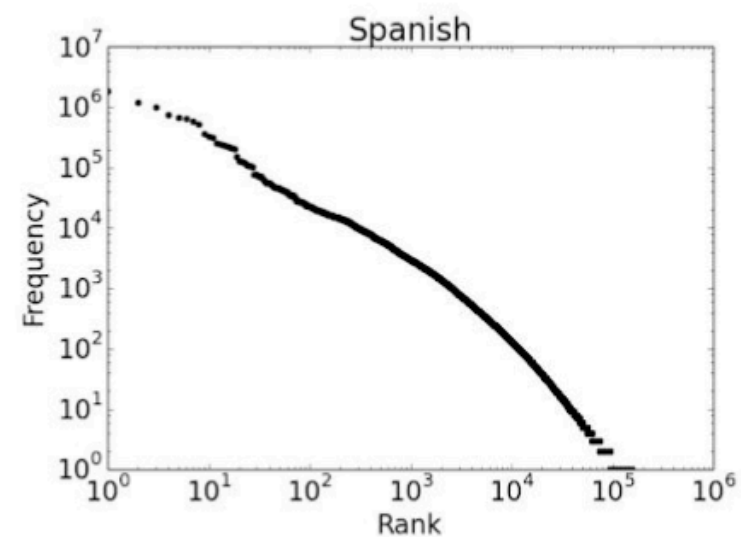
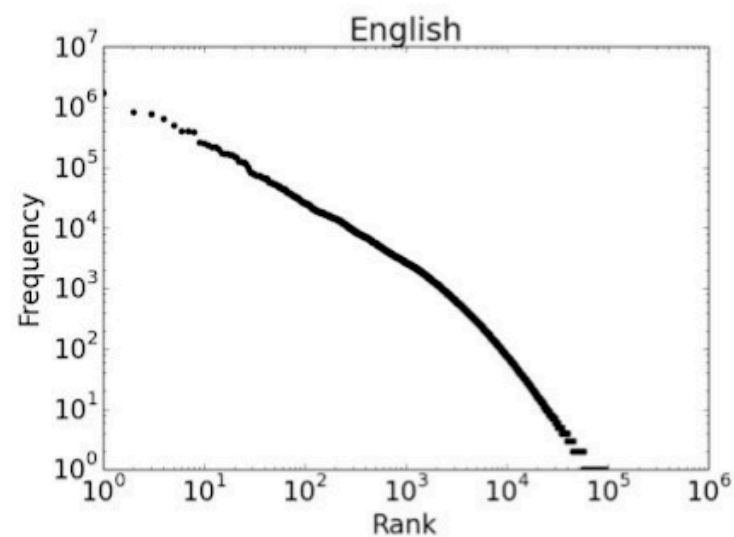
- Order words by frequency. What is the frequency of n_{th} ranked word?



Zipf's Law

Implications:

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
- We need to find clever ways to estimate probabilities for things we have rarely or never seen

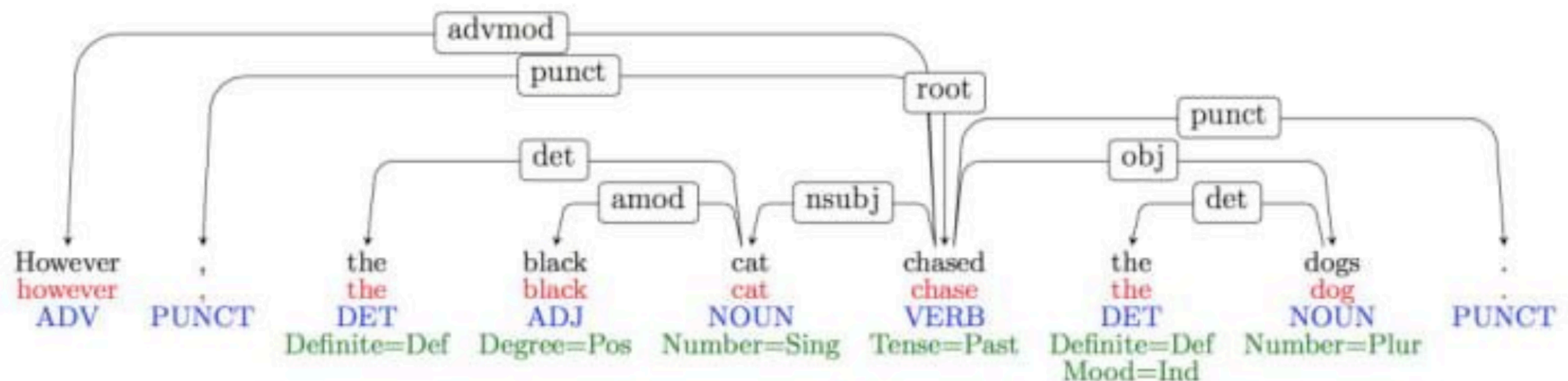


Why NLP is Hard?

- Ambiguity
- Scale
- Sparsity
- Variation
- Expressivity
- Unmodeled variables
- Unknown representations \mathbb{R}

Variation

- Suppose we train a part of speech tagger or a parser on the Wall Street Journal

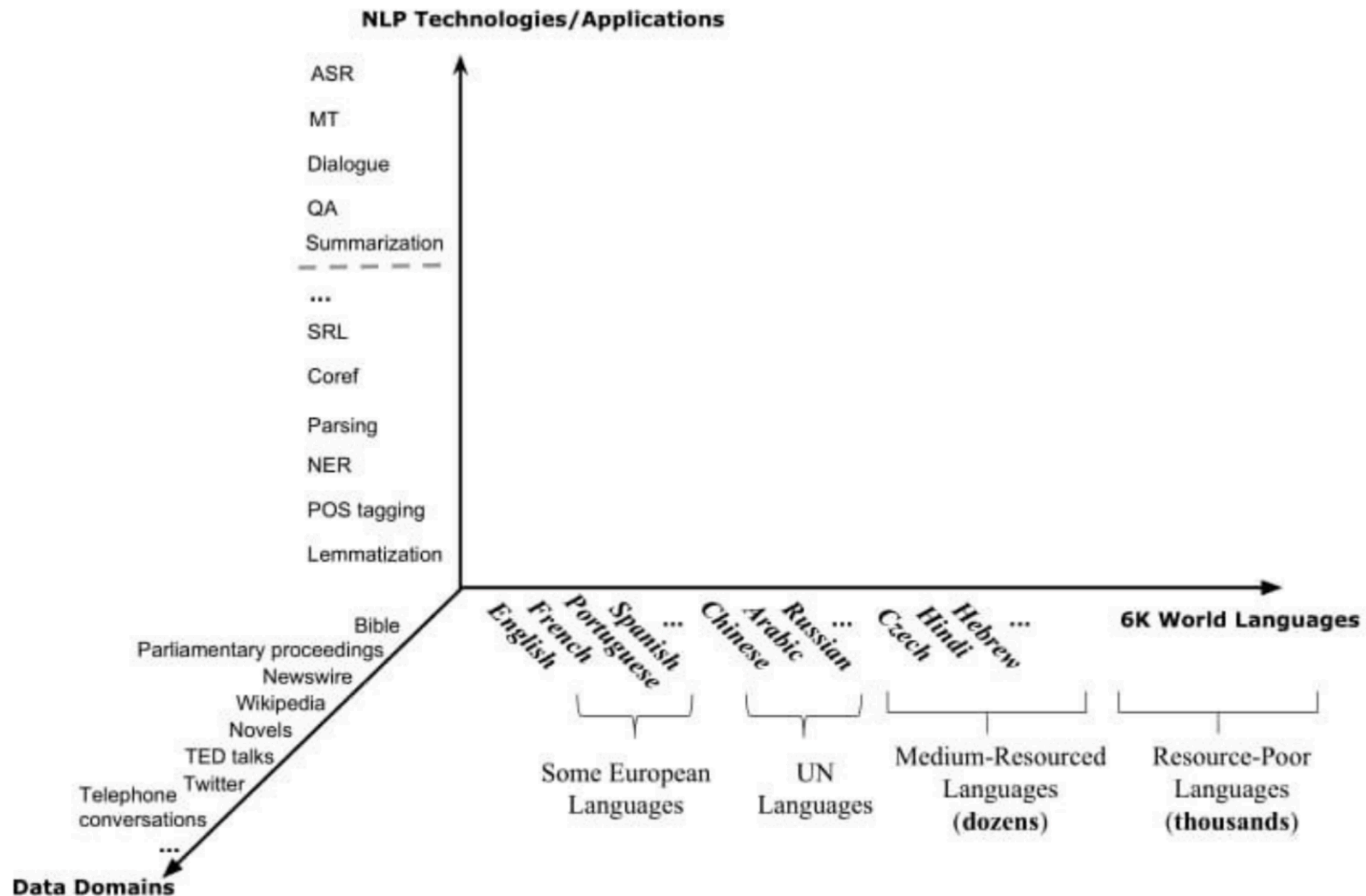


- What will happen if we try to use this tagger/parser for social media?

@_rkpnrnte hindi ko alam babe eh, absent ako
kanina I'm sick rn hahaha 🤔👏

Variation

- Training data comes from diverse domains
- Potential distributional shift between train/test data



Why NLP is Hard?

- Ambiguity
- Scale
- Sparsity
- Variation
- Expressivity
- Unmodeled variables
- Unknown representations \mathbb{R}

Expressivity

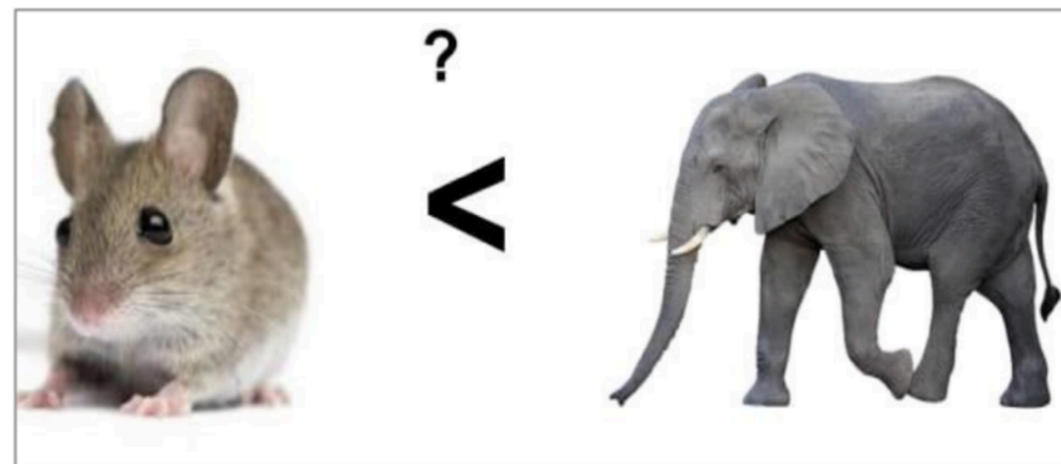
- Not only can one form have different meanings (ambiguity) but **the same meaning can be expressed with different forms:**
 - **She gave the book to Tom vs. She gave Tom the book**
 - **Some kids popped by vs. A few children visited**
 - **Is that window still open? vs. Please close the window**
 -

Unmodeled Variables

- World knowledge
 - I dropped the glass on the floor and it broke
 - I dropped the hammer on the glass and it broke



“Drink this milk”

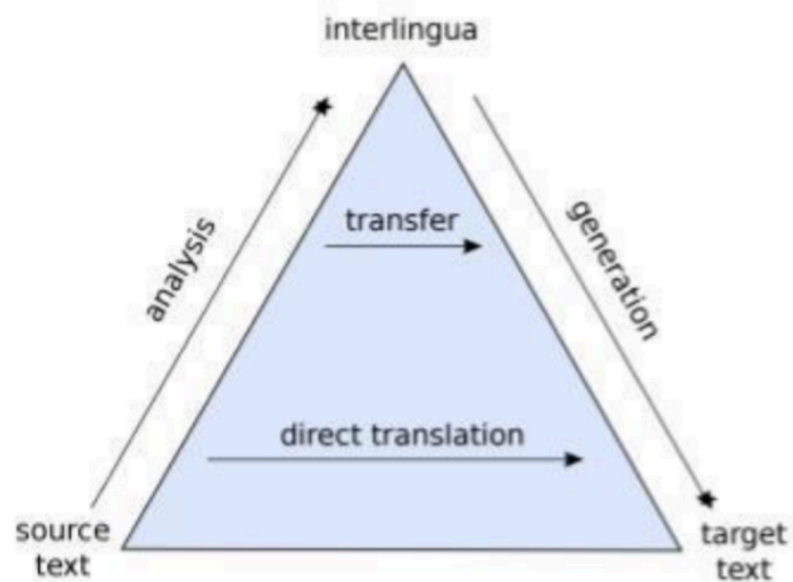


Unmodeled Meaning Representation

- Very difficult to capture what is \mathbb{R} , since we don't even know how to represent the knowledge a human has/needs:
 - What is the “meaning” of a word or sentence?
 - How to model context?
 - Other general knowledge?

Symbolic and Probabilistic NLP

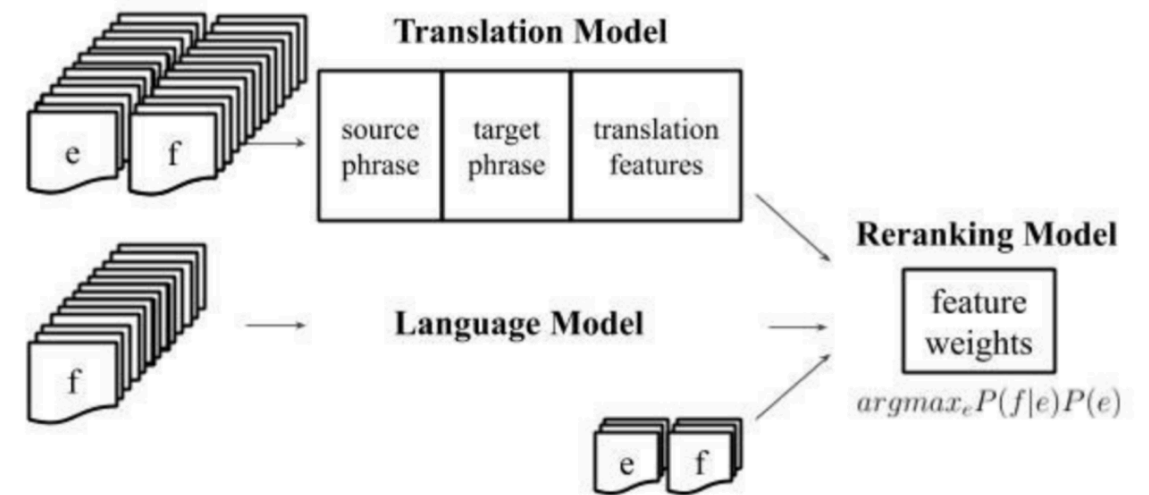
Logic-based/Rule-based NLP



~90s

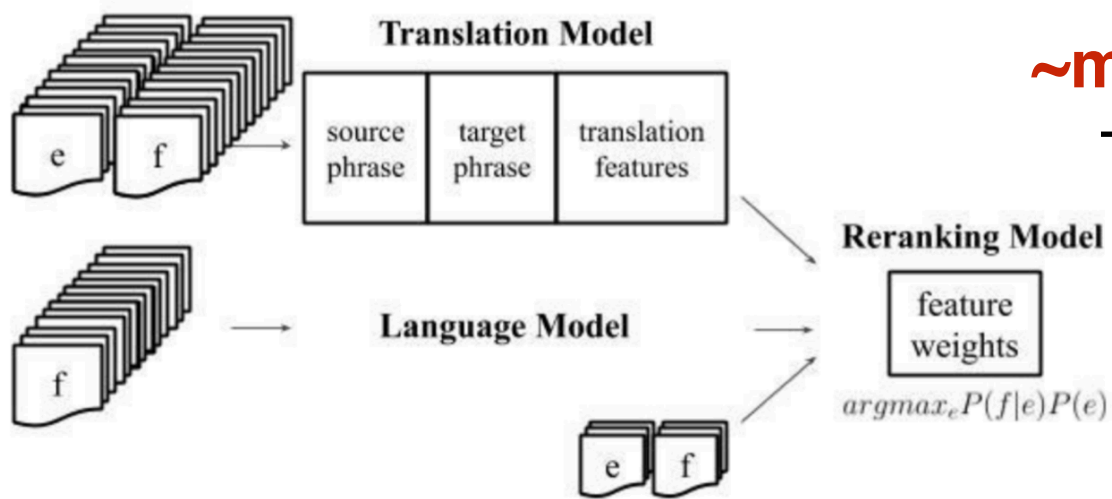


Statistical NLP

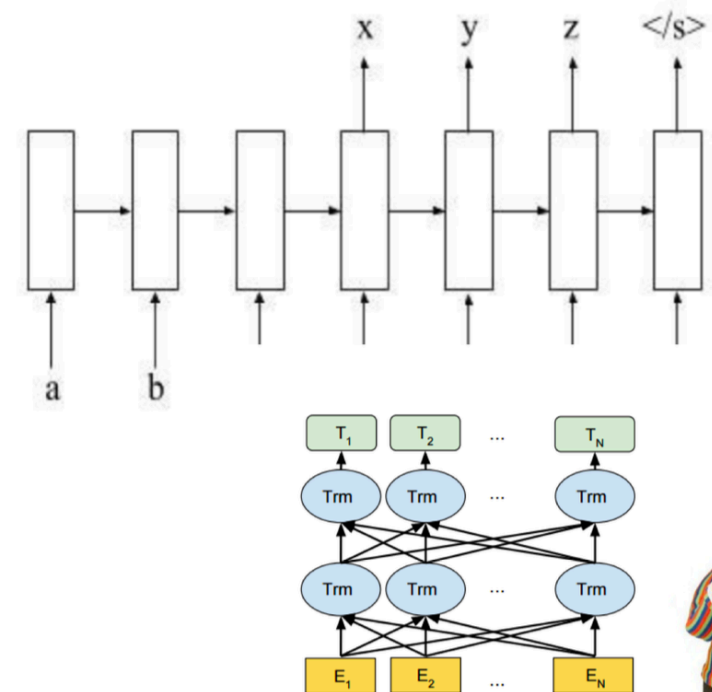


Probabilistic and Connectionist NLP

Engineered Features/Representations



Learned Features/Representations



NLP vs Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- R is not directly observable.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

NLP vs Linguistics

- NLP must contend with NL data as found in the world
- NLP \approx computational linguistics
- Linguistics has begun to use tools originating in NLP!

Fields with Connections to NLP

- Machine learning
- Deep Learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Data science
- Political science
- Psychology
- Economics
- Education

NLP System Building Overview

A General Framework for NLP Systems

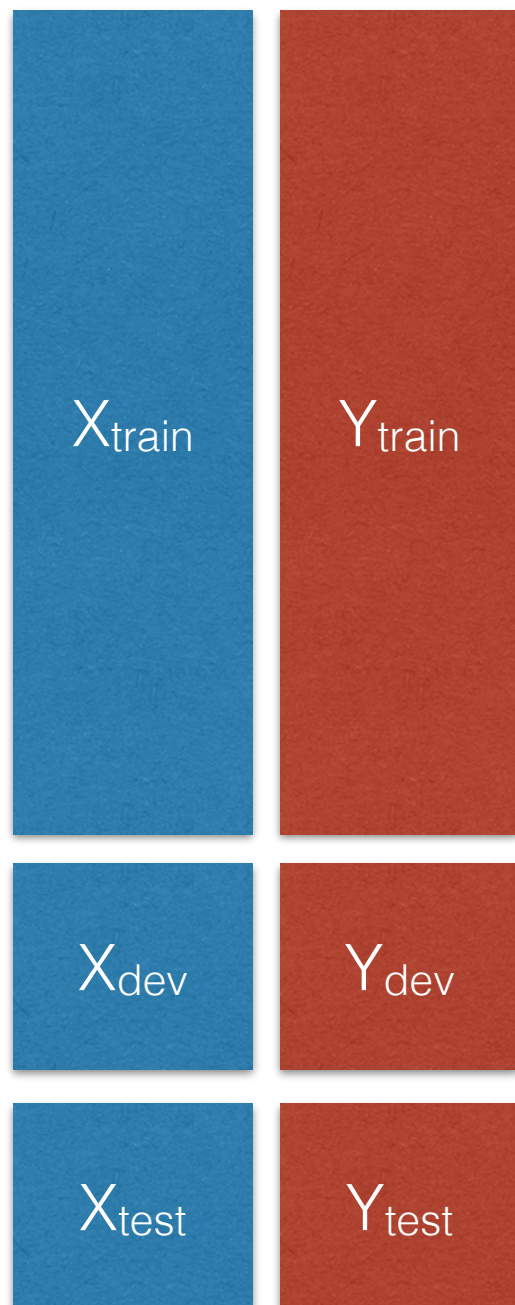
- Formally, create a function to map an **input X (language)** into an **output Y** . Examples:

<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

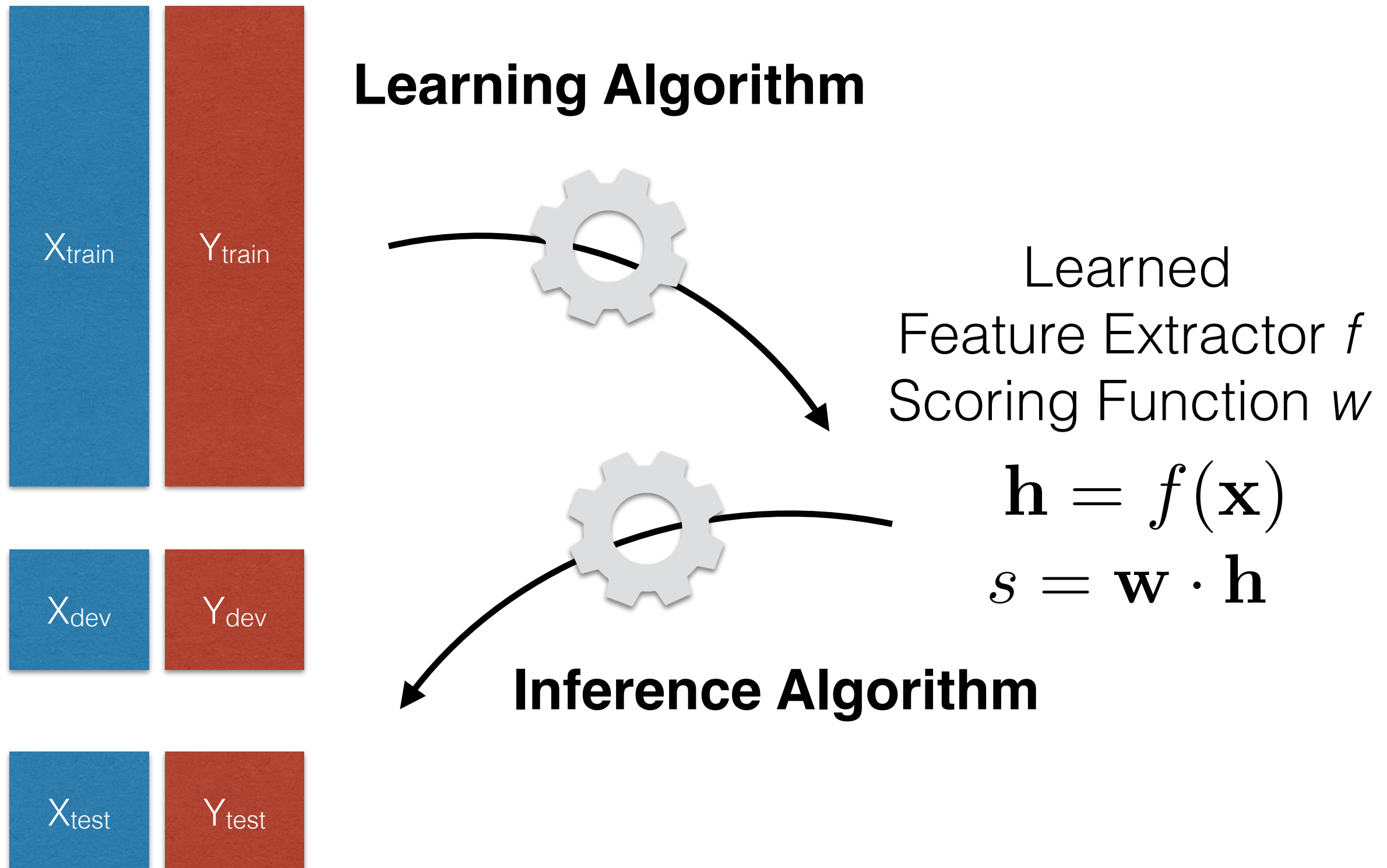


Training Set: Generally larger dataset, used during system design, creation, and learning of parameters.

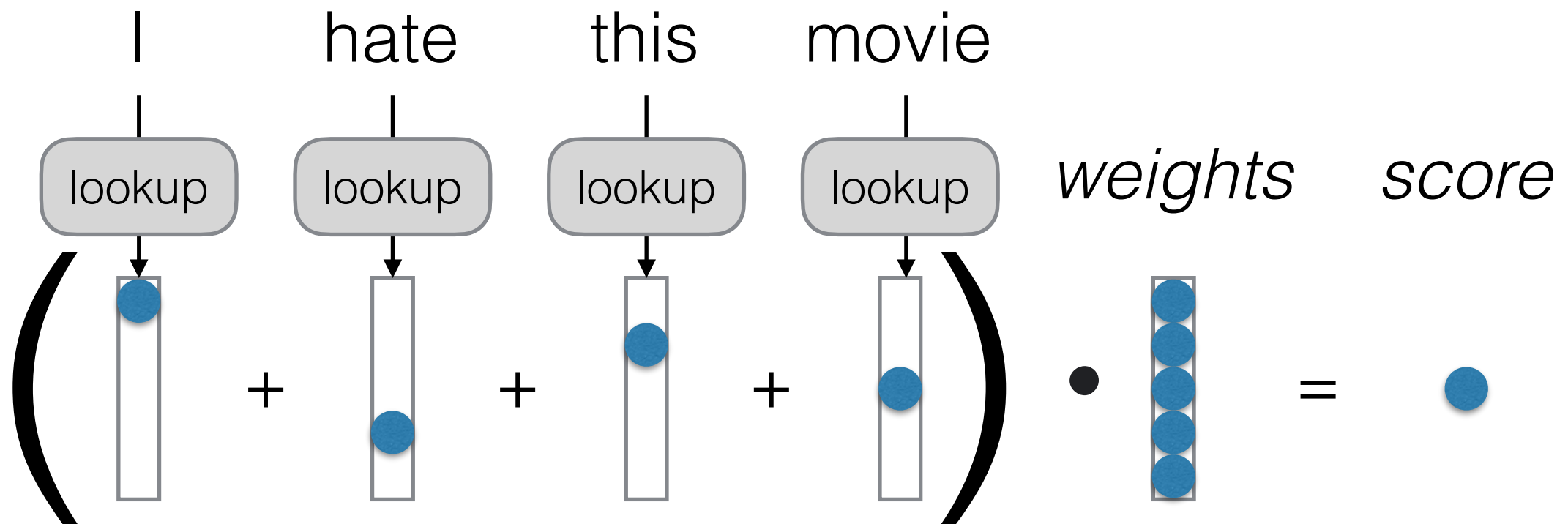
Development ("dev", "validation") Set: Smaller dataset for testing different design decisions ("hyper-parameters").

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

Machine Learning



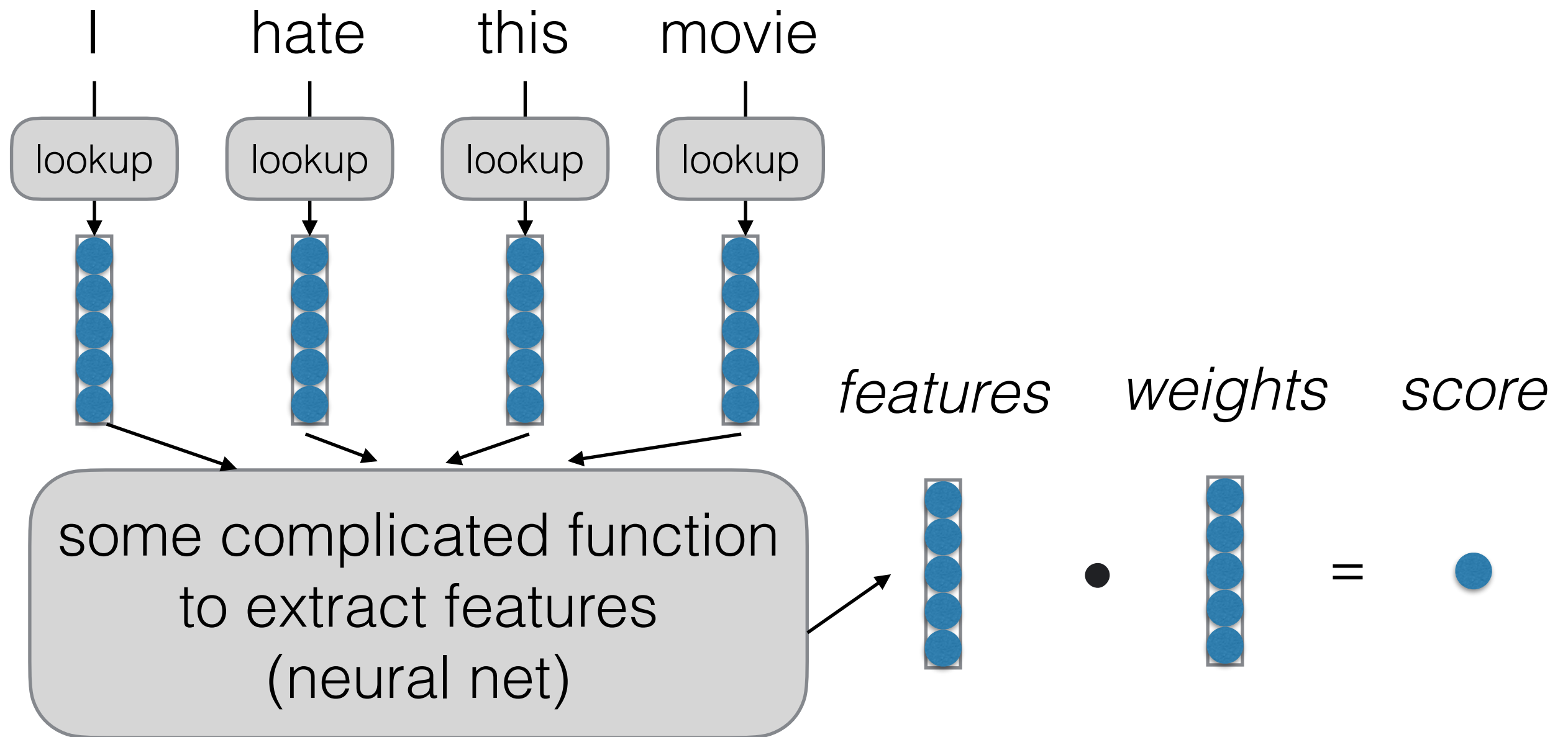
Bag of Words (BOW)



Features f are based on word identity, weights w learned

Which problems mentioned before would this solve?

Neural Network Models

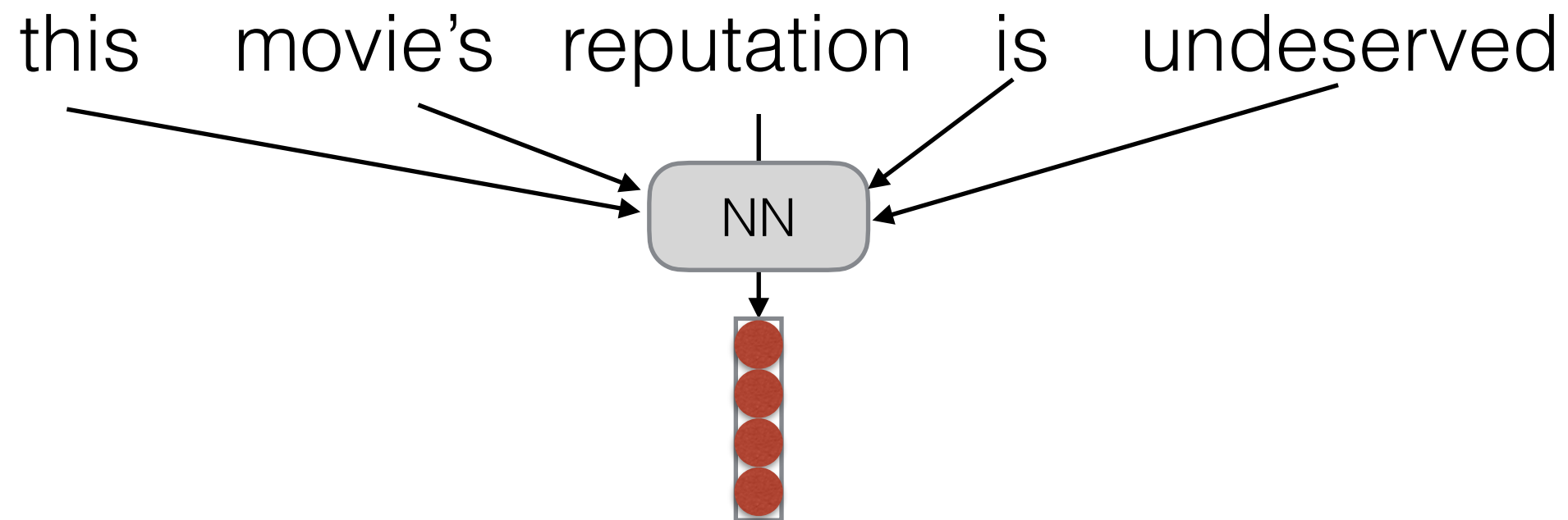


Class Goals

- 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

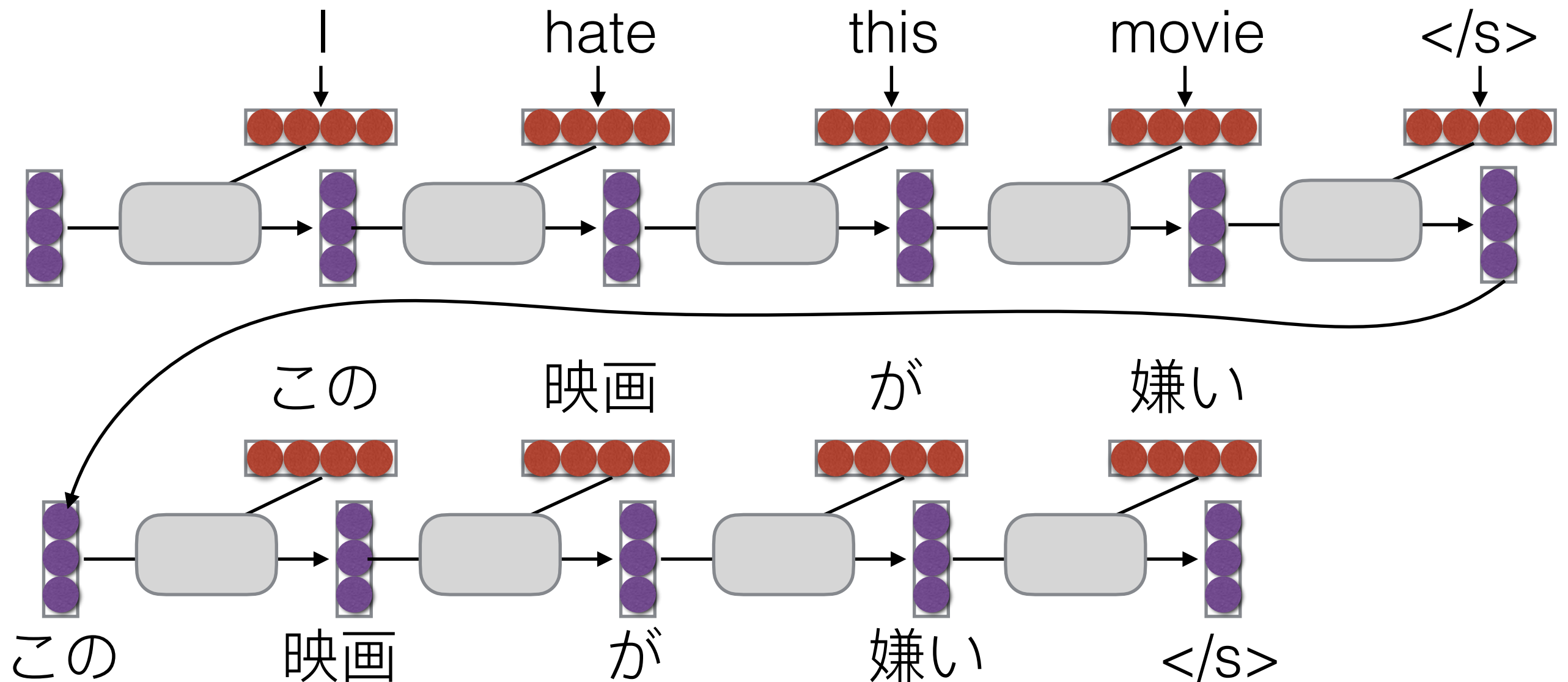
Roadmap Going Forward

Topic 1: NLP Fundamentals



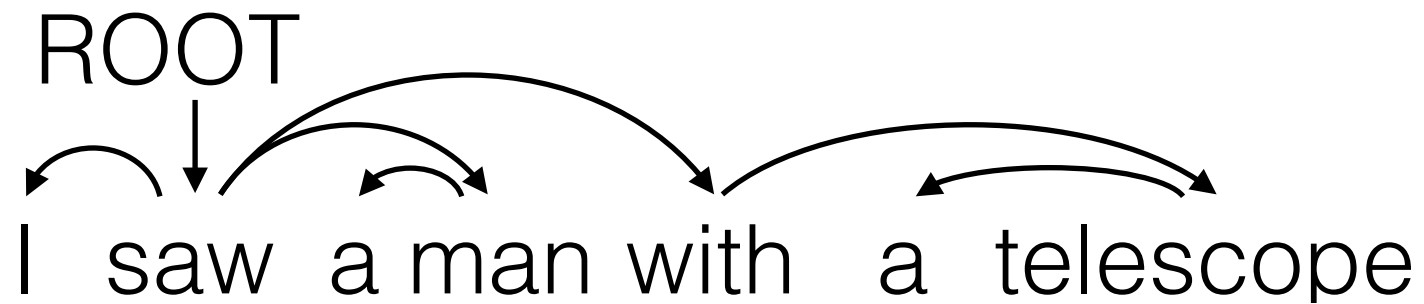
- Text Classification and ML Fundamentals
- Language Modeling and NN Training Tricks
- Word Vectors
- Neural Network Basics and Toolkit Construction

Topic 2: Modeling and Neural Net Basics



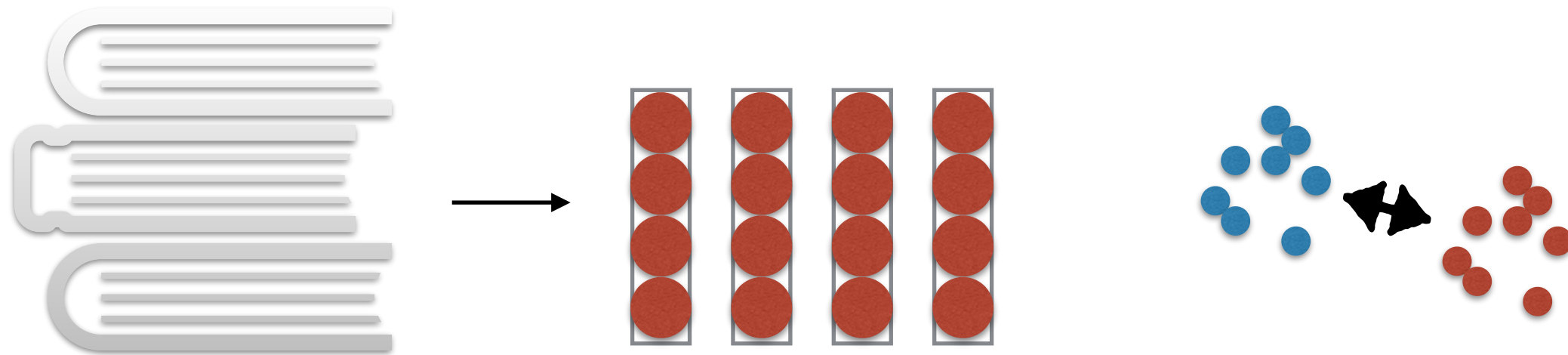
- Recurrent Networks
- Conditioned Generation
- Attention

Topic 3: Natural Language Analysis



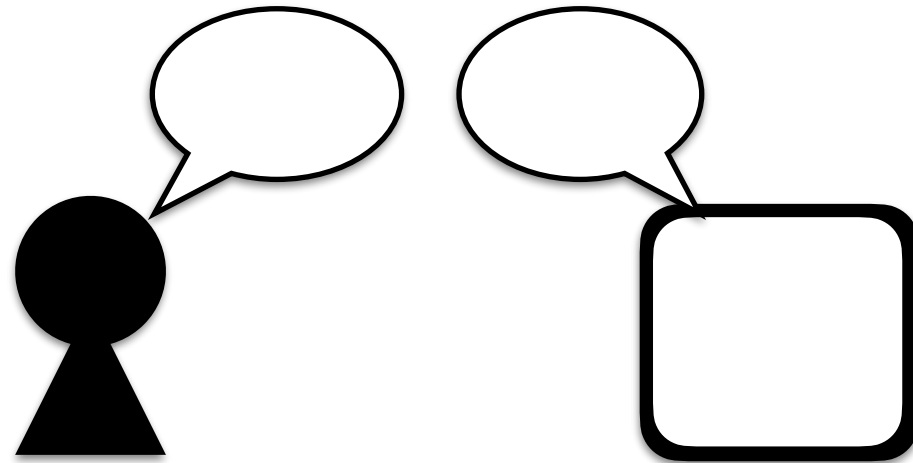
- Word Segmentation and Morphology
- Syntactic Parsing
- Semantic Parsing
- Discourse Structure and Analysis

Topic 4: Representation Learning and Algorithms



- Pre-training and Self-supervised Learning
- Multi-task and Multi-lingual Learning
- Prompting and Few-shot Learning
- Long Sequence Models
- Structured Learning Algorithms
- Latent Variable Models

Topic 5: NLP Applications



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

Class Format/Structure

Class Delivery Format: In Person

- **Keep wearing masks!**
- Maintain social distance as much as possible.
- Office hour section remains **online on Zoom.**
- Contact instructors if there's any concerns

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 materials on Canvas, due on the same day.

Assignments

- **Assignment 1 - Build-your-own Neural Network Toolkit:** *Individually* implement some parts of a neural network (15%)
- **Assignment 2 - Text Classifier / Questionnaire:** *Individually* implement a text classifier and fill in questionnaire on topics of interest (15%)
- **Assignment 3 - SOTA Survey / Re-implementation:** Re-implement and reproduce results from a recently published NLP paper (20%)
- **Final Project:** Perform a unique project that either (1) improves on state-of-the-art, or (2) applies NLP models to a unique task. Have an oral presentation and write a report. (30%)
- **Quiz:** 20%

Instructors

- **Instructor:**
 - Junjie Hu
- **TAs:**
 - Yibing Wei
- **Piazza:** <https://piazza.com/class/kykntsenzmc30o>
- **Canvas:**
 - **HW:** <https://canvas.wisc.edu/courses/292771/assignments>
 - **Quiz:** <https://canvas.wisc.edu/courses/292771/quizzes>

Thanks, Any Questions?