CS 4650/7650:
Natural Language Processing

# Introduction to NLP

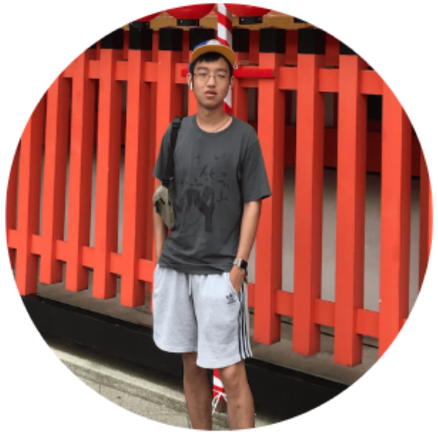
Diyi Yang

Some slides borrowed from Yulia Tsvetkov at CMU and Noah Smith at UW

# Welcome!

Diyi Yang

Ian Stewart



Jiaao Chen

Nihal Singh

# Course Website

**https://www.cc.gatech.edu/classes/AY2020/cs7650_spring**

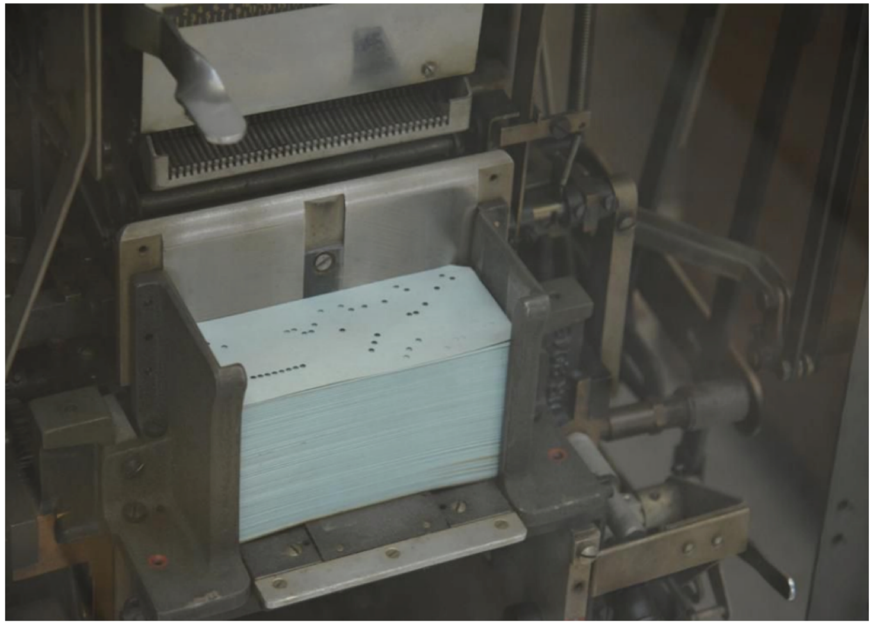
CS-4650/7650: Natural Language Processing, Spring 2020

## NLP: CS-4650/7650

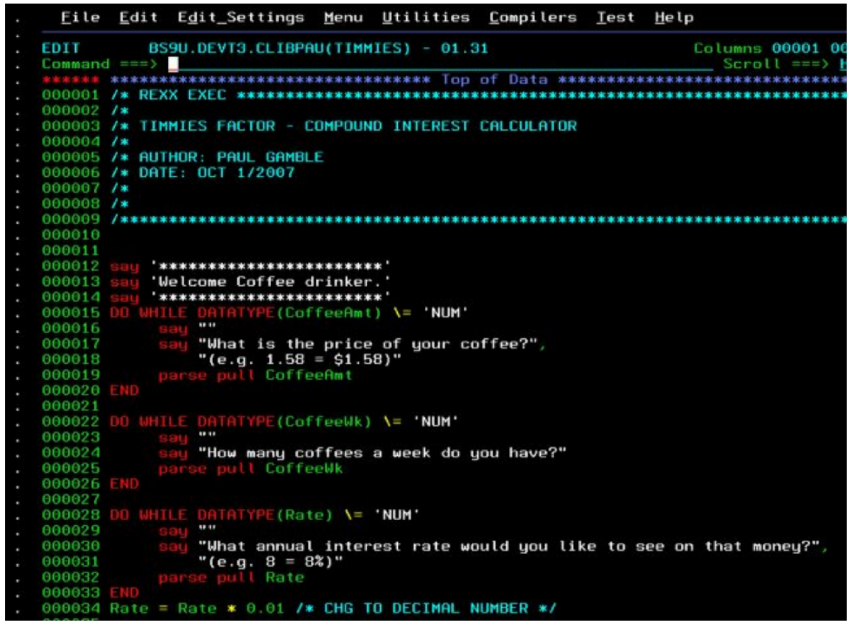
MW 3:00-4:15pm, Kendeda 215

Course Information  Schedule  Grading  Policies  Prerequisites  FAQs  Projects  External Resources

# Communication With Machines



~50-70s




~80s

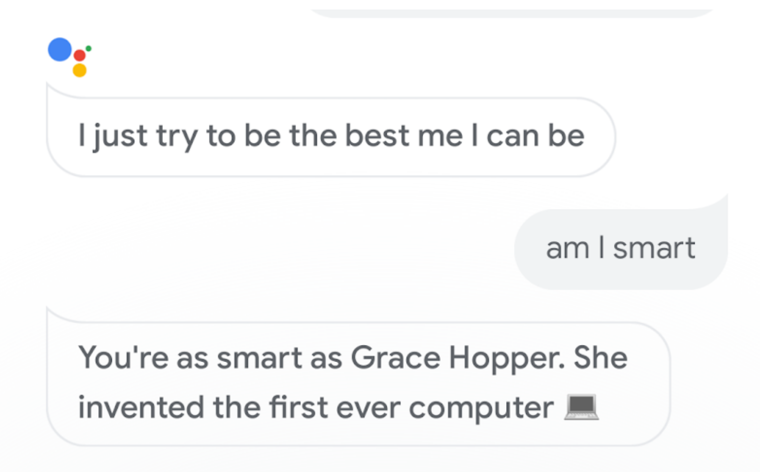

today

# Conversational Agents

Conversational agents contain:

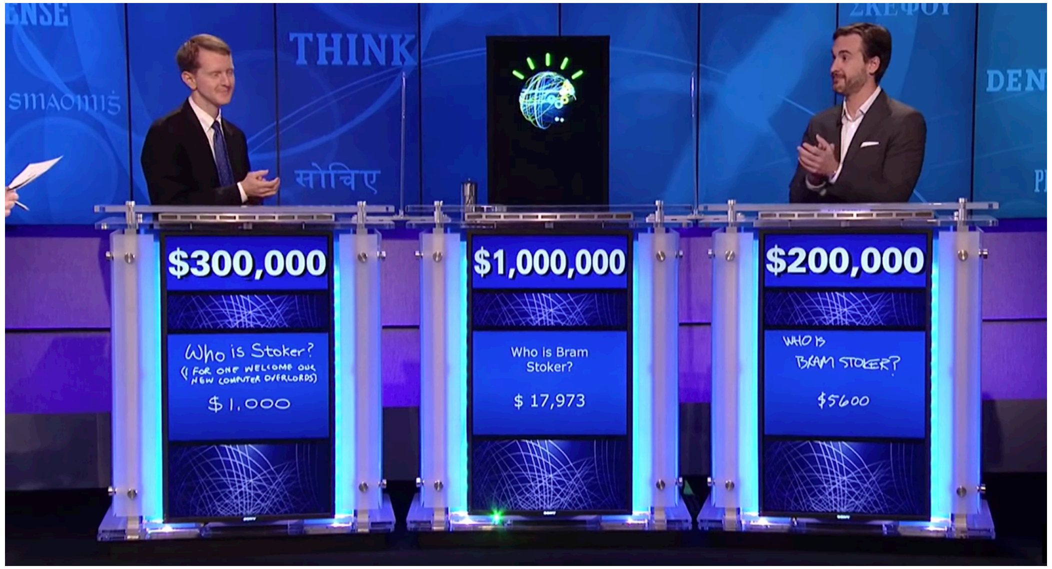
- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech

works with the Google Assistant



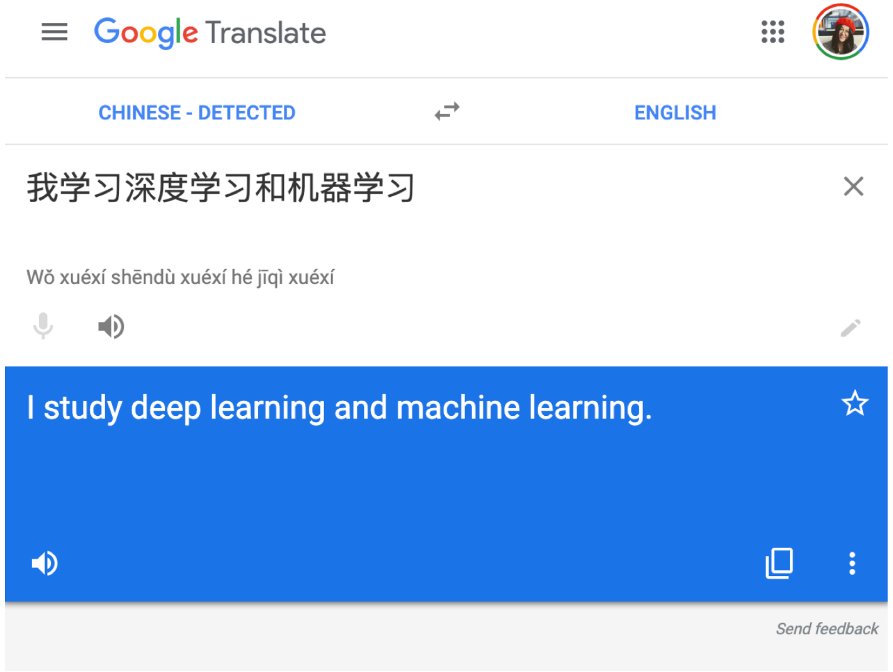

In early 2011, an IBM computing system named Watson
competed against the world's best Jeopardy! champions.

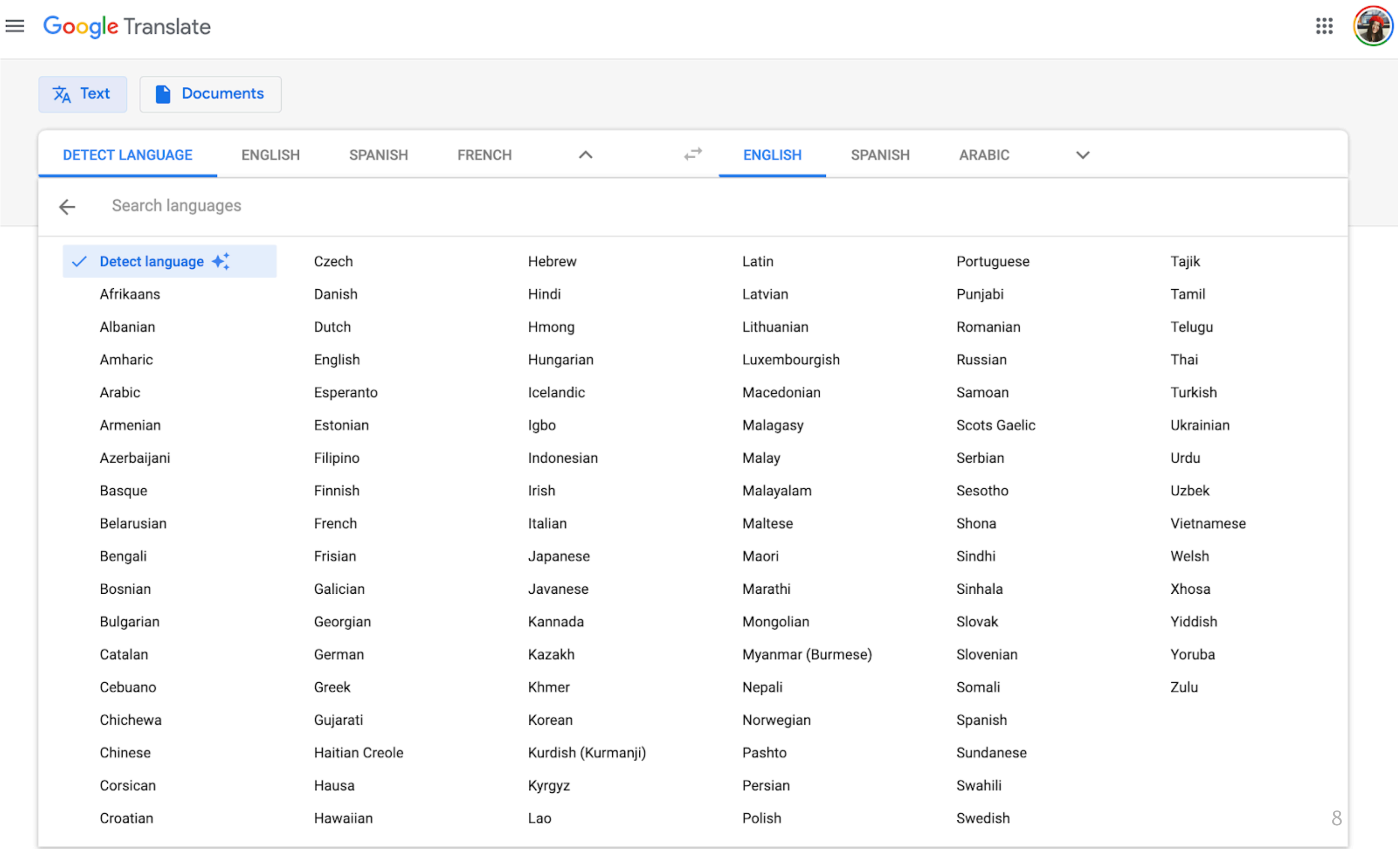
# Question Answering




- 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?

# Machine Translation



# Natural Language Processing

## Applications

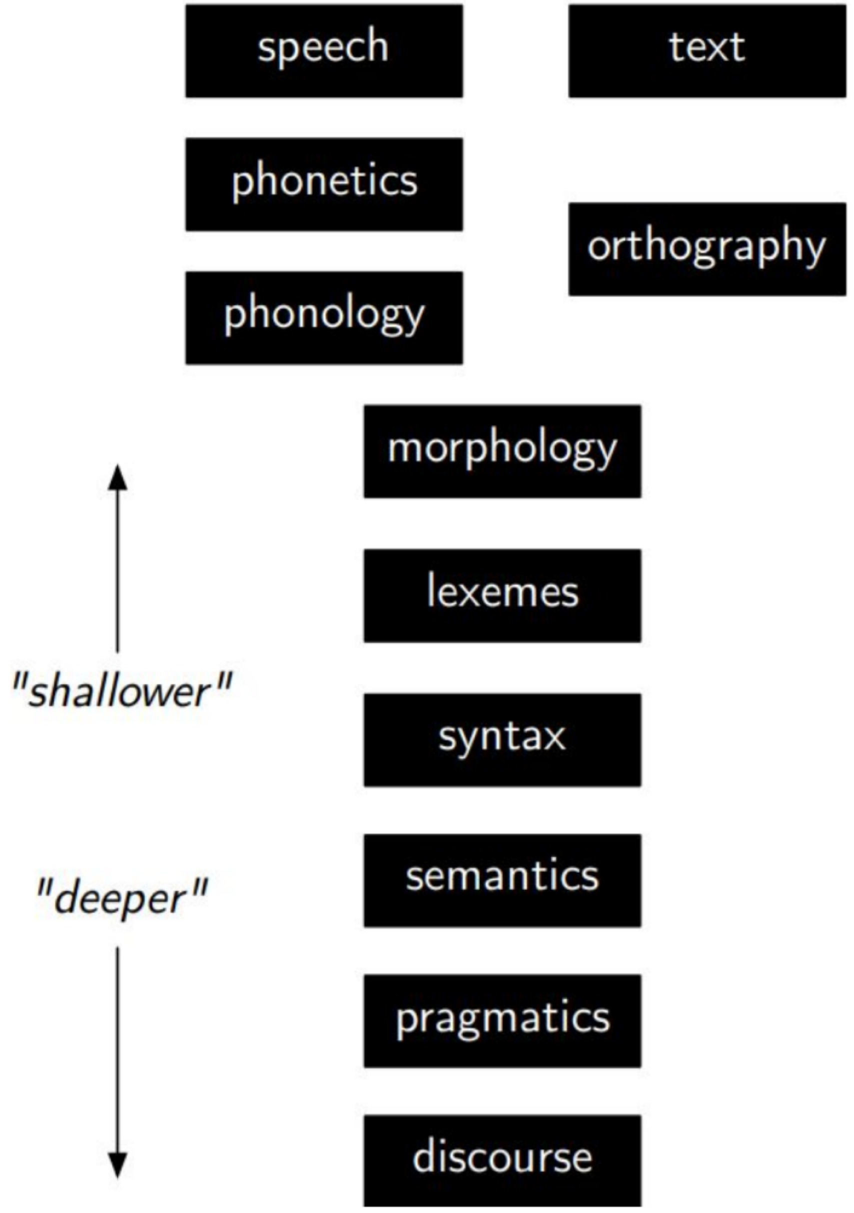
- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- …

## Core Technologies

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Word sense disambiguation
- Semantic role labeling
- …

NLP lies at the intersection of computational linguistics and machine learning.

# Level Of Linguistic Knowledge



# Phonetics, Phonology

- Pronunciation Modeling

**SOUNDS** Th i a si e n

# Words

- Language Modeling
- Tokenization
- Spelling correction

**WORDS** This is a simple sentence

# Morphology

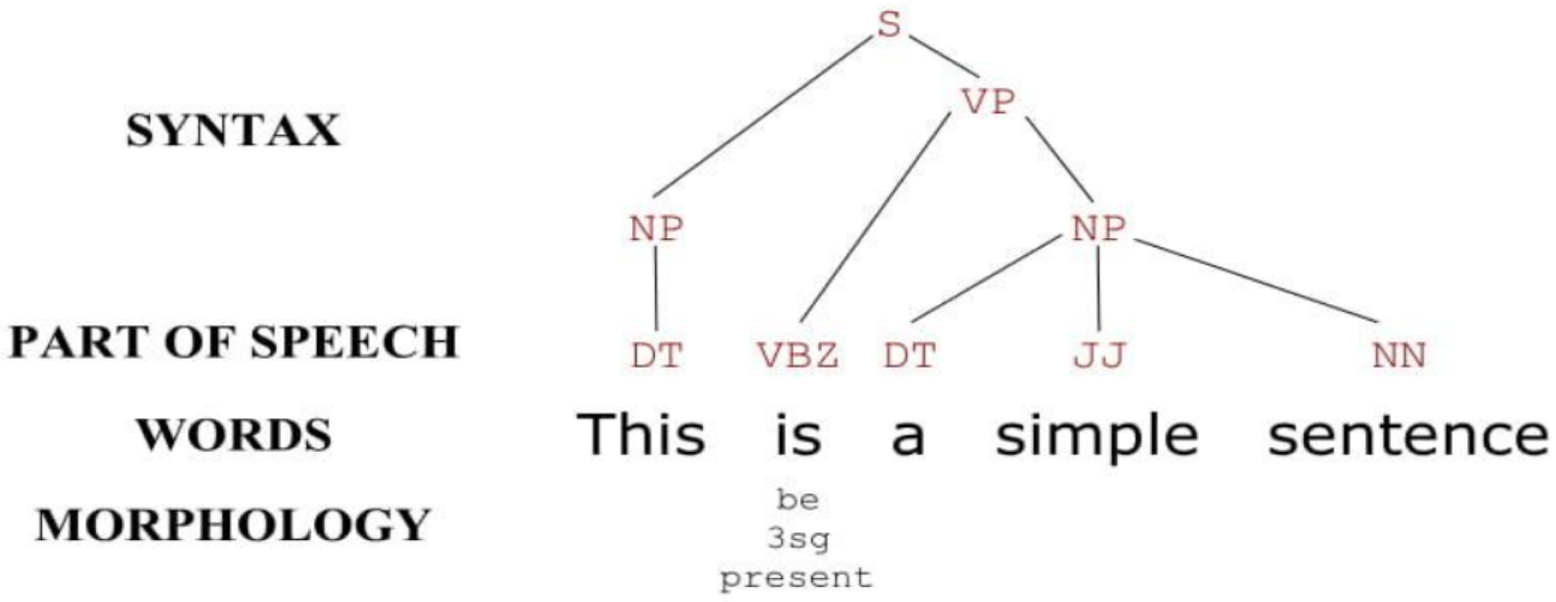
- Morphology analysis
- Tokenization
- Lemmatization

# Part of Speech

- Part of speech tagging

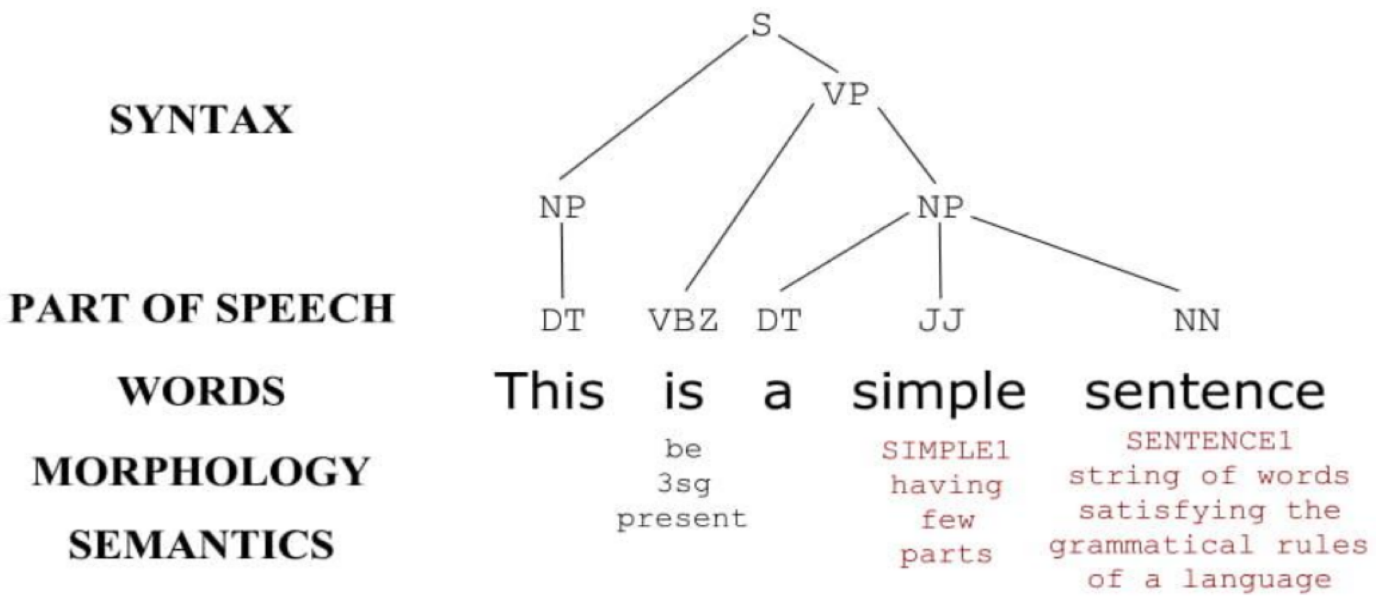
# Syntax

- Syntactic parsing

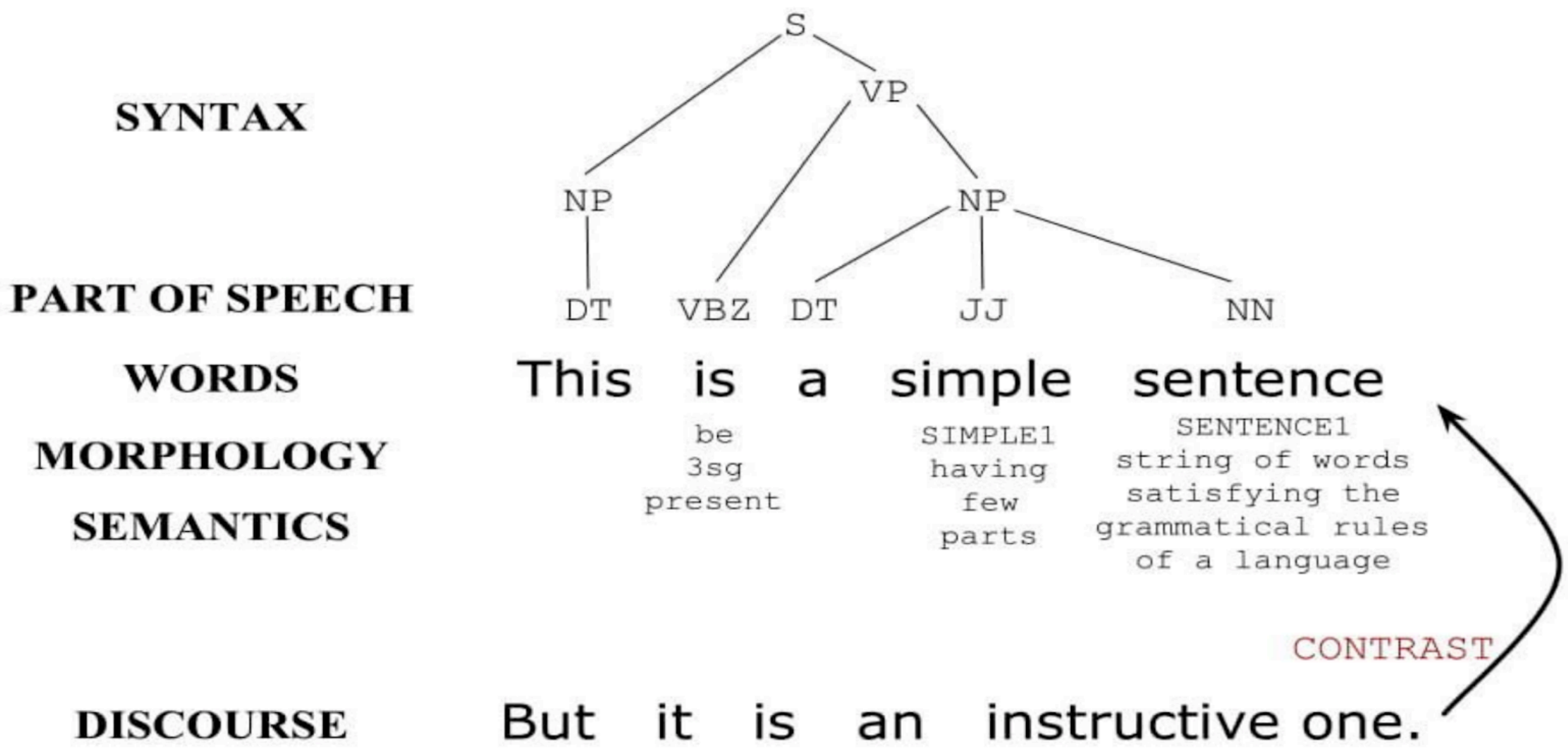


# Semantics

- Named entity recognition
- Word sense disambiguation
- Semantic role labeling



# Discourse


SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
SEMANTICS
DISCOURSE
S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This is a simple sentence
be
3sg
present
SIMPLE1
having
few
parts
SENTENCE1
string of words
satisfying the
grammatical rules
of a language
CONTRAST
But it is an instructive one.

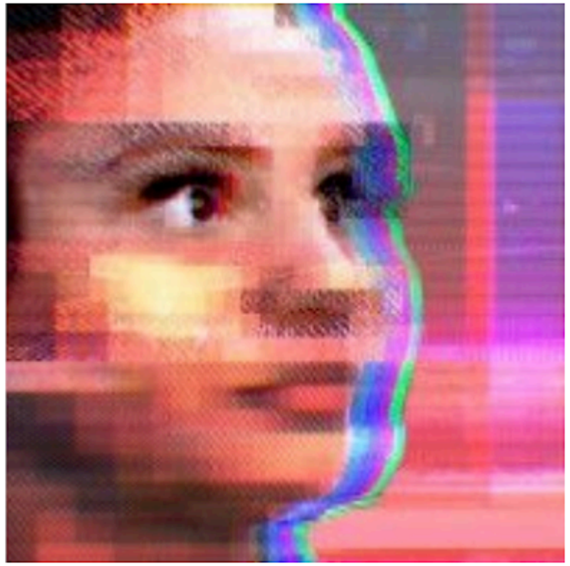
# Where Are We Now?

# Where Are We Now?

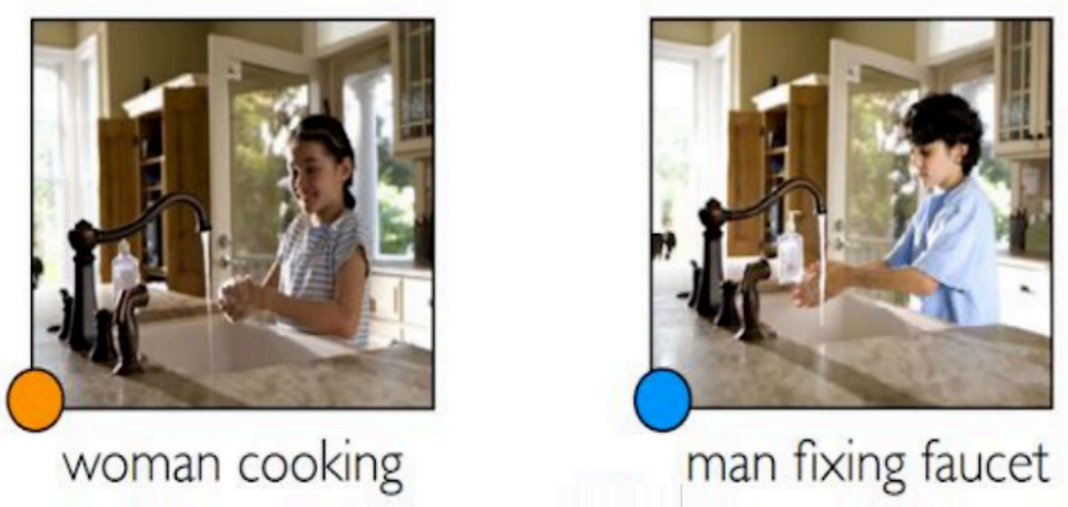
| Baseline mutual information model (Li et al. 2015) |
| --- |
| A: Where are you going? (1) |
| B: I'm going to the restroom. (2) |
| A: See you later. (3) |
| B: See you later. (4) |
| A: See you later. (5) |
| B: See you later. (6) |
| ... |
| ... |
| A: how old are you? (1) |
| B: I'm 16. (2) |
| A: 16? (3) |
| B: I don't know what you are talking about. (4) |
| A: You don't know what you are saying. (5) |
| B: I don't know what you are talking about . (6) |
| A: You don't know what you are saying. (7) |
| ... |

VS

# Where Are We Now?



https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist




Zhao, J., Wang, T., Yatskar, M., Ordonez, V and Chang, M.-W. (2017) Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraint. EMNLP

# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations

# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations

# Ambiguity

- Ambiguity at multiple levels
  - Word senses: **bank** (finance or river ?)
  - Part of speech: **chair** (noun or verb ?)
  - Syntactic structure: **I can see a man with a telescope**
  - Multiple: **I made her duck**



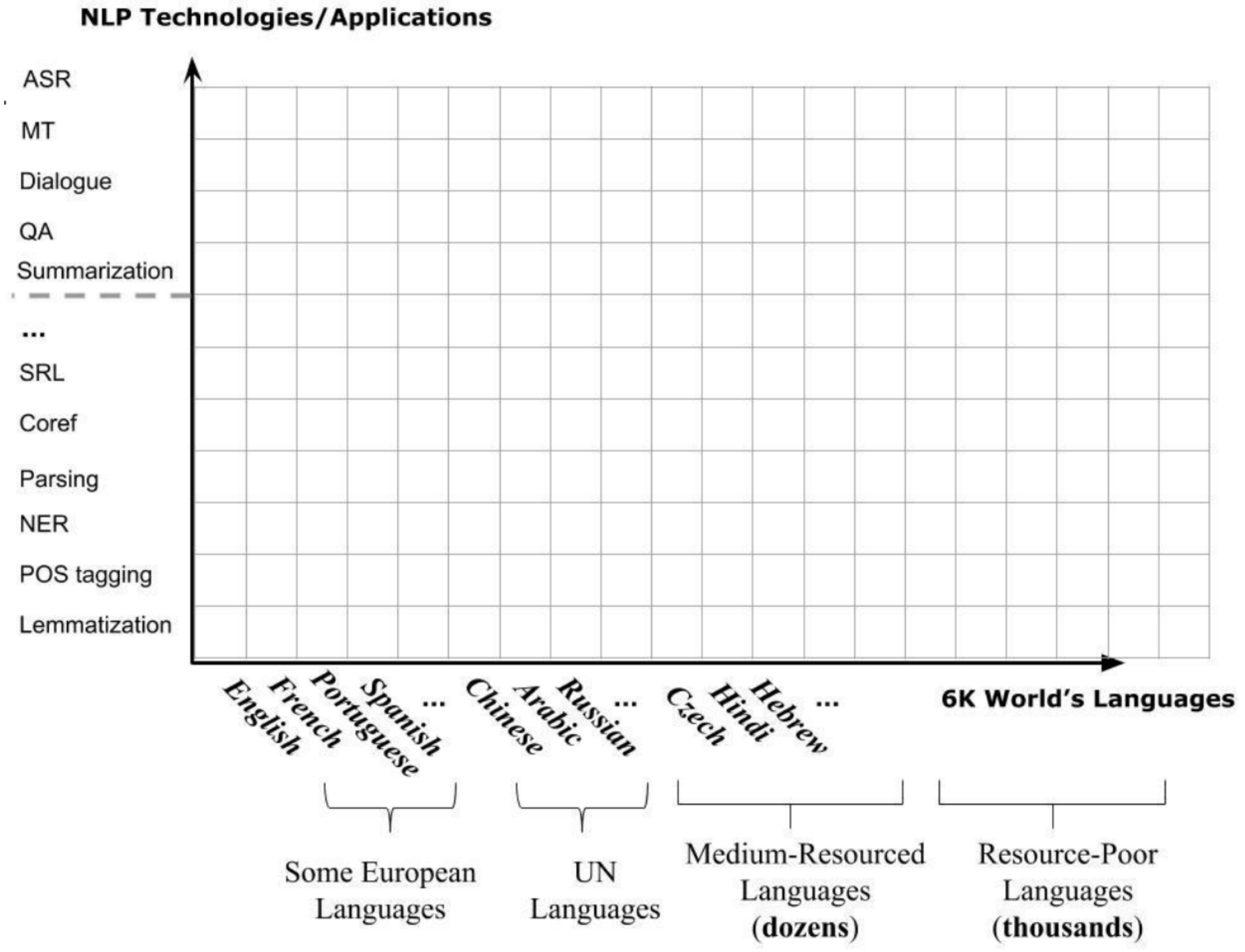
"One morning I shot
an elephant in my pajamas"

*I made her duck*

[SLP2 ch. 1]

- 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
- …

# Ambiguity and Scale



# The Challenges of “Words”

- Segmenting text into words
- Morphological variation
- Words with multiple meanings: bank, mean
- Domain-specific meanings: latex
- Multiword expressions: make a decision, take out, make up

# Part of Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol

# Part of Speech Tagging

| I know, right | shake my head | | | for | your | | |
|---|---|---|---|---|---|---|---|
| ikr | smh | he | asked | fir | yo | last | name |

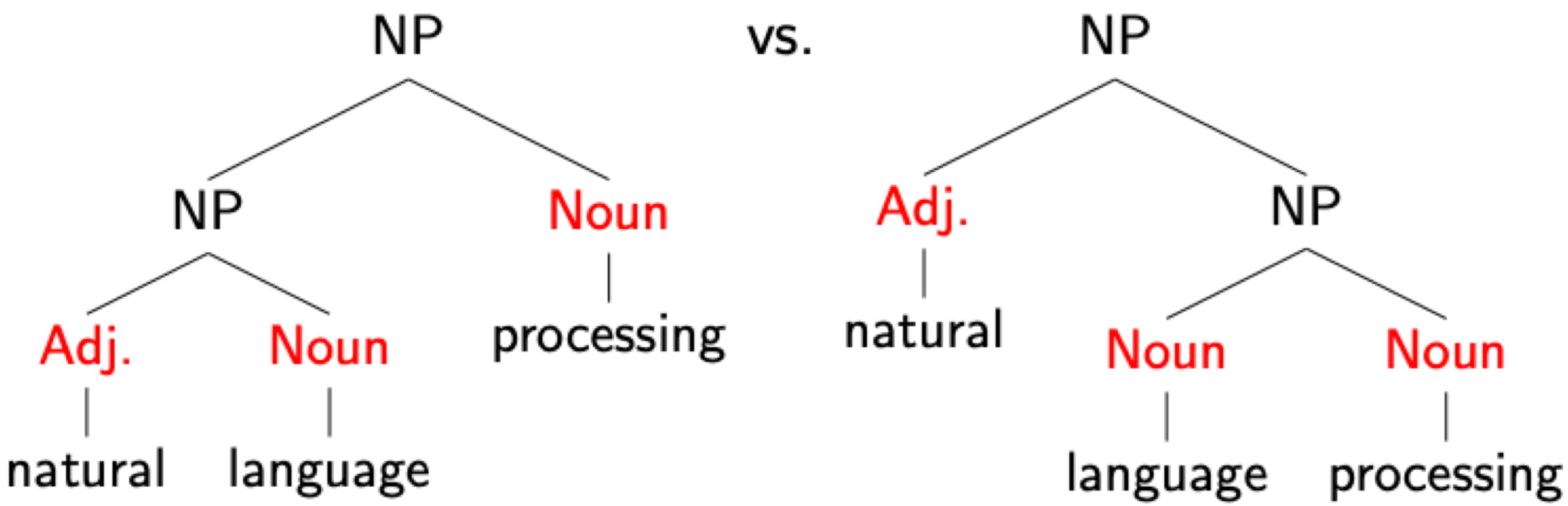
| | | | | you | | Facebook | laugh out loud |
|---|---|---|---|---|---|---|---|
| so | he | can | add | u | on | fb | lololol |

# Part of Speech Tagging

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I know, right | shake my head | | | for | your | | |
| ikr | smh | he | asked | fir | yo | last | name |
| ! | G | O | V | P | D | A | N |
| interjection | acronym | pronoun | verb | prep. | det. | adj. | noun |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | you | | Facebook | laugh out loud |
| so | he | can | add | u | on | fb | lololol |
| P | O | V | V | O | P | ^ | ! |
| preposition | | | | | | proper noun | |

# Syntax


NP
NP
Noun
Adj.
Noun
natural
language
processing
vs.
NP
Adj.
NP
natural
Noun
Noun
language
processing

# Morphology + Syntax


A ship-shipping ship, shipping shipping-ships

# Semantics

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

# Semantics

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

– Groucho Marx

# Syntax + Semantics

- We saw the woman with the telescope wrapped in paper.

# Syntax + Semantics

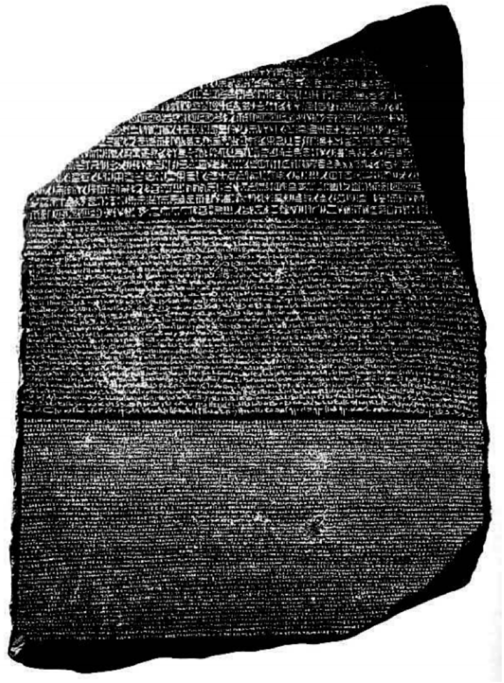
- We saw the woman with the telescope wrapped in paper.
  - Who has the telescope?
  - Who or what is wrapped in paper?
  - An even of perception, or an assault?

# Dealing with Ambiguity

- How can we model ambiguity?
  - Non-probabilistic methods (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 analyses, i.e., the most probable one
- 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 French/English sentences
  - Yelp reviews
  - The Web!



Rosetta Stone

# Statistical NLP

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

# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations

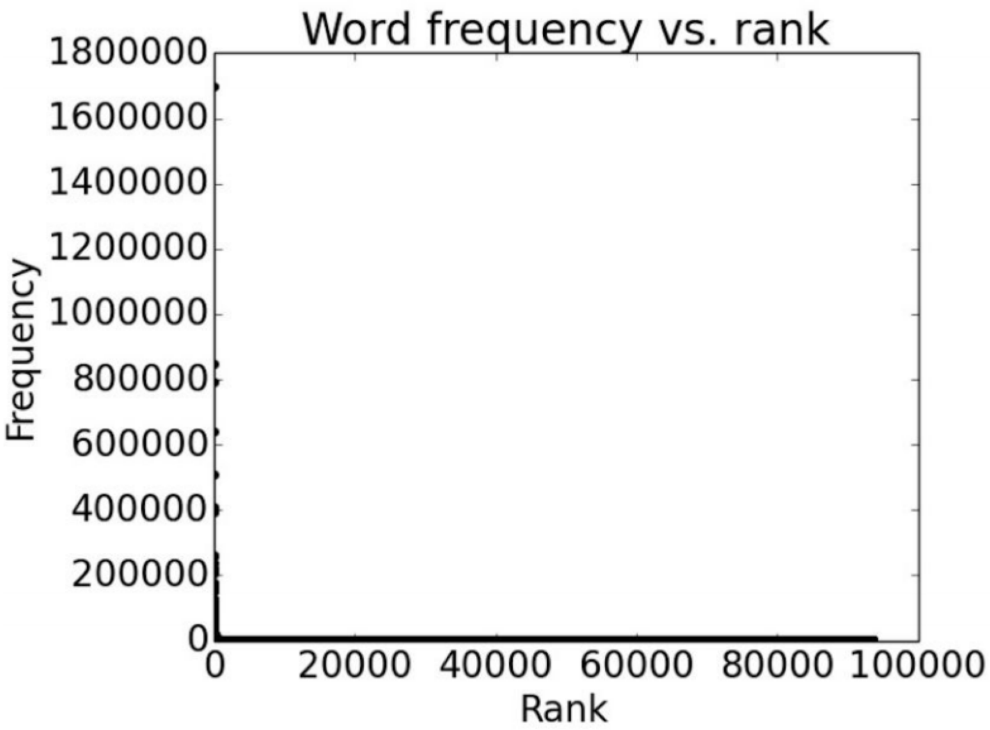
# Sparsity

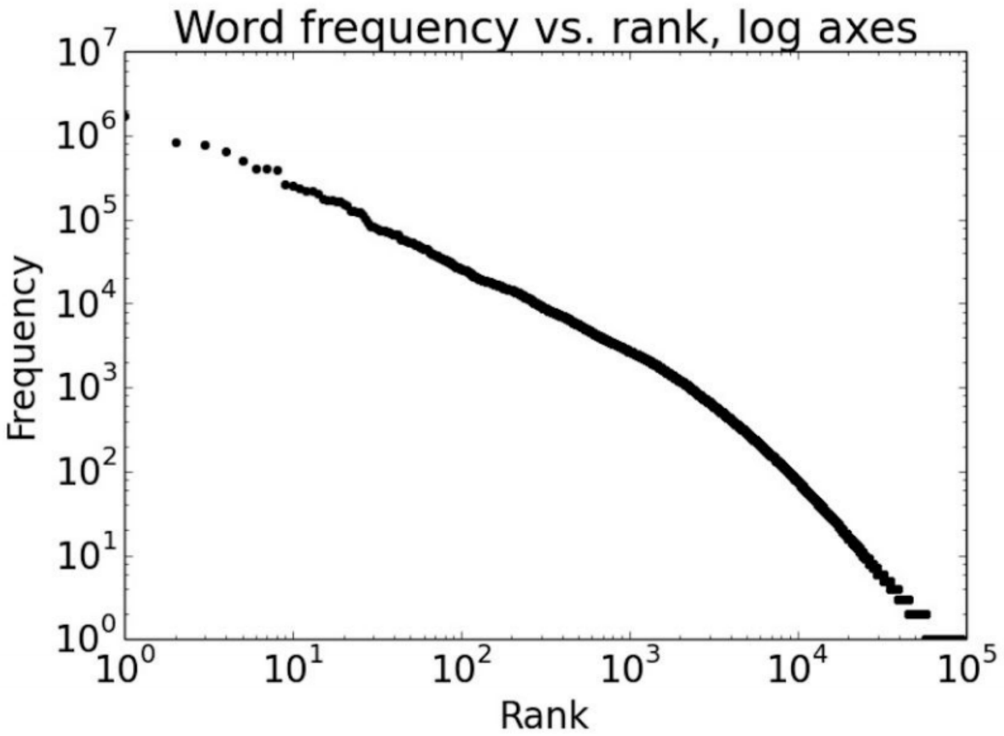
- Sparse data due to Zipf's Law
- Example: the frequency of different words in a large text corpus

| 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 |

# Sparsity

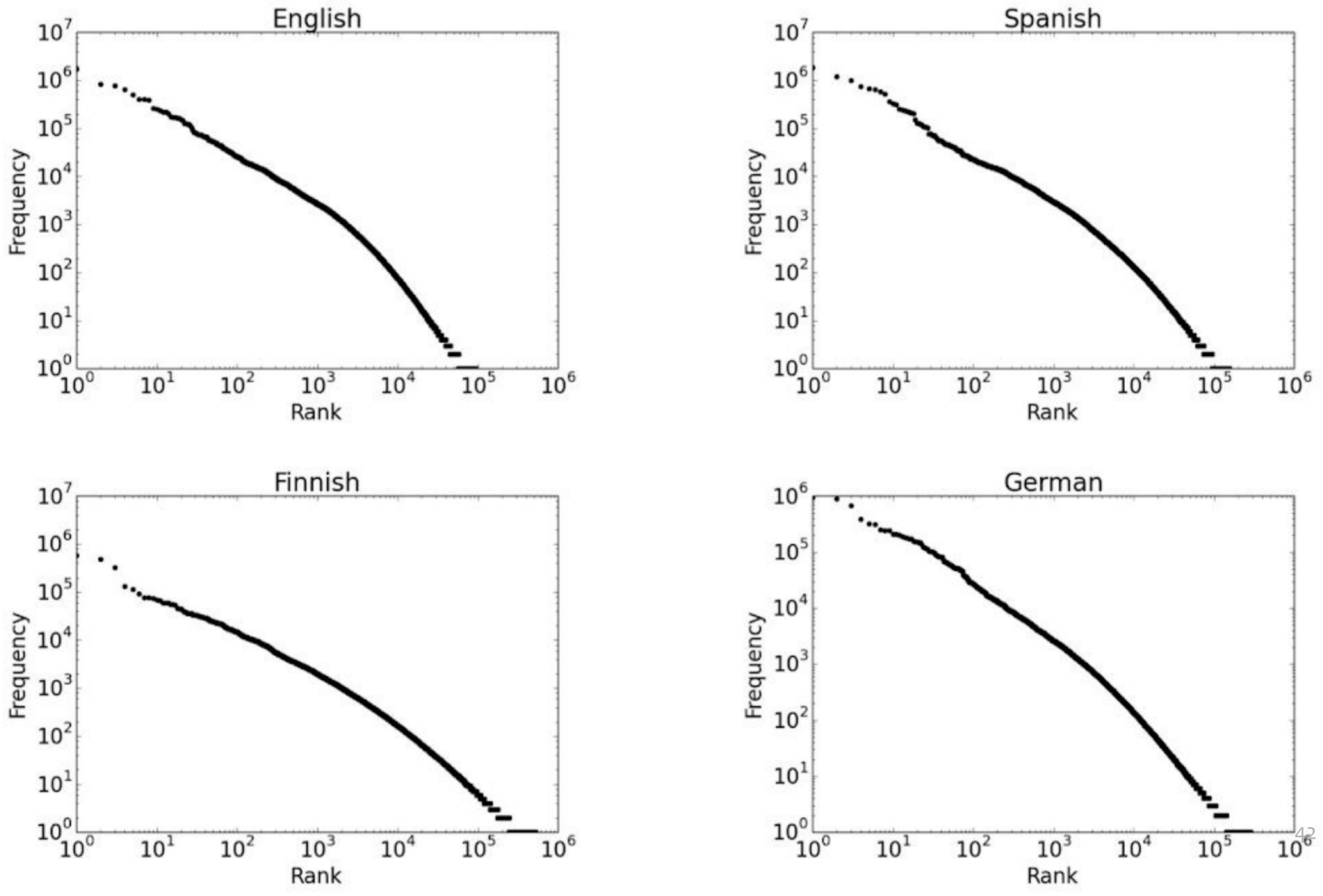
- Order words by frequency. What is the frequency of nth ranked word?



# Sparsity

- Regardless of how large our corpus is, there will be a lot of infrequent words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen

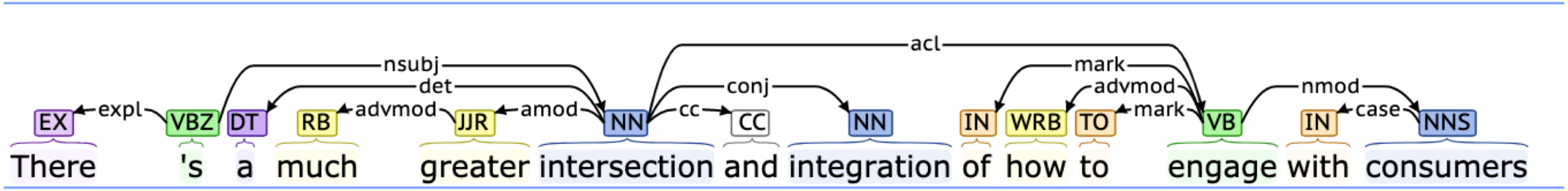


# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations

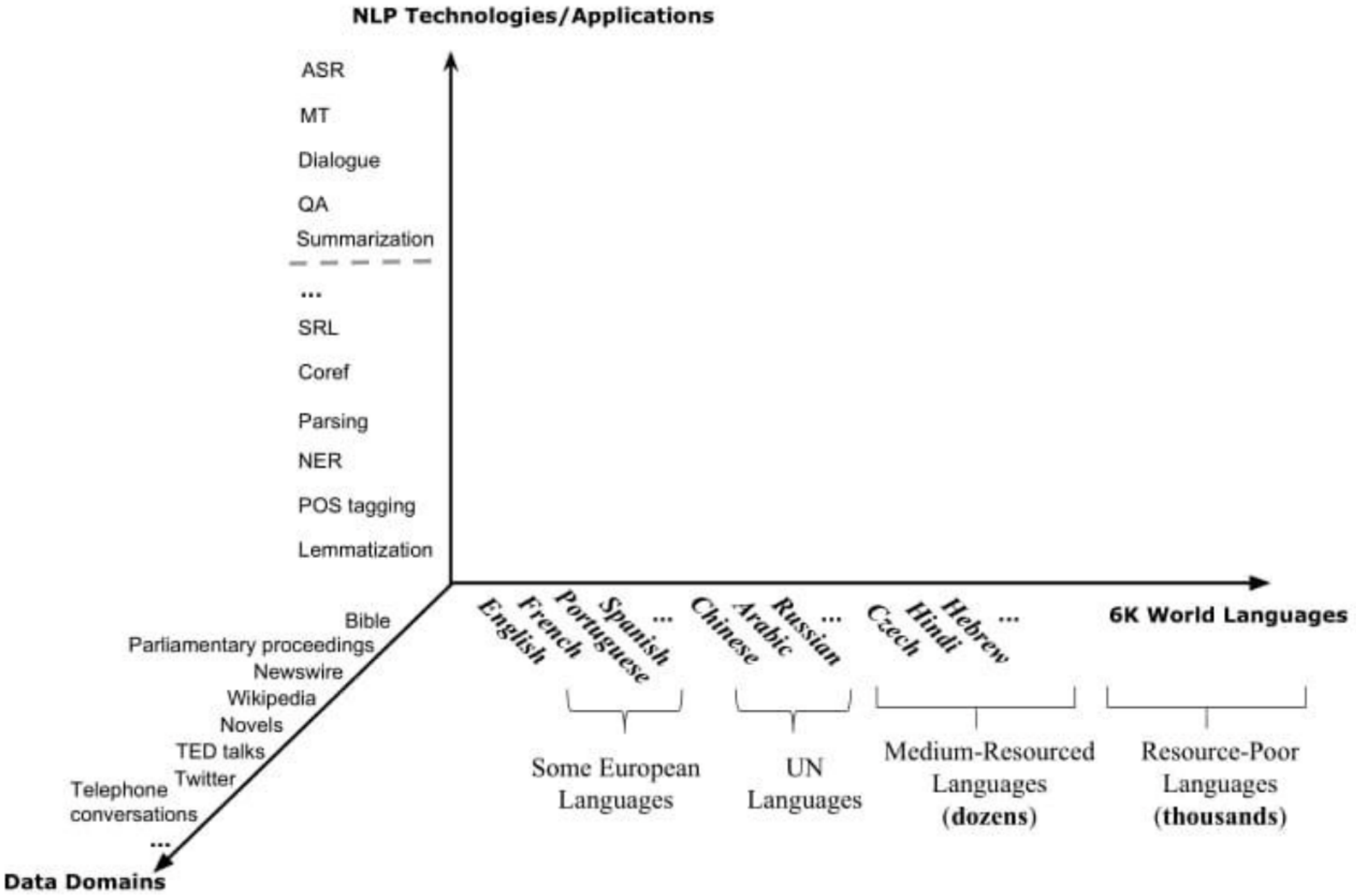
# Variation

- 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?
  - *"ikr smh he asked fir yo last name so he can add u on fb lololol"*

# Variation

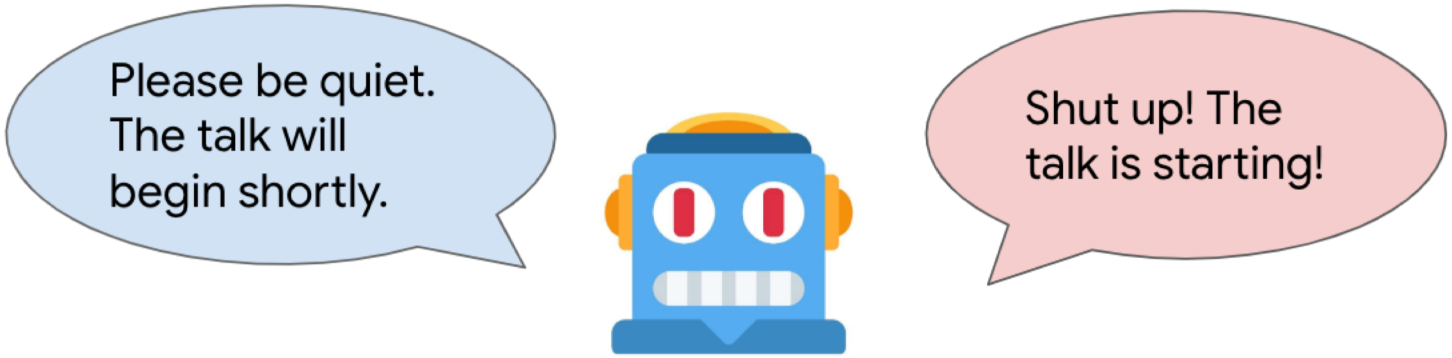

NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
6K World Languages
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)
Data Domains
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...

# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations

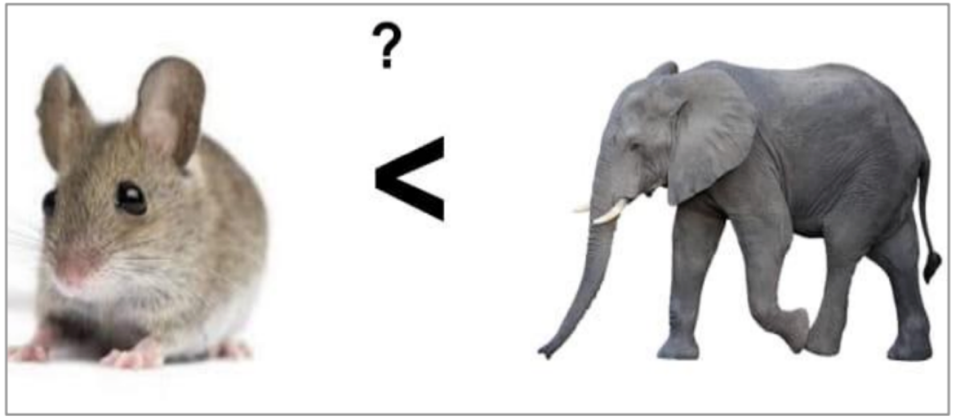
# Expressivity

- 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

"Drink this milk"




## World knowledge

I dropped the glass on the floor and it broke

I dropped the hammer on the glass and it broke

# Unmodeled Representation

Very difficult to capture what is $\mathcal{R}$, since we don't even know how to represent the knowledge a human has/needs:

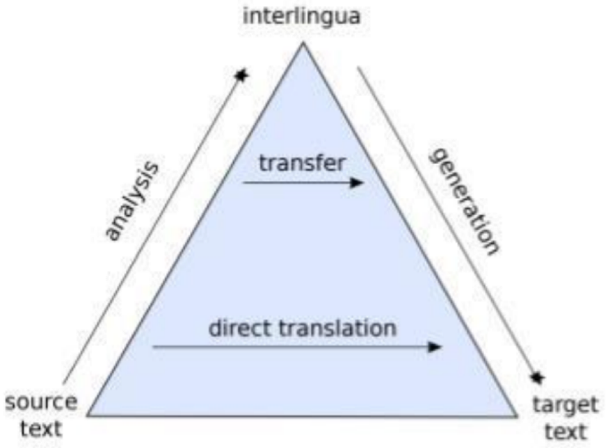
- What is the "meaning" of a word or sentence?
- How to model context?
- Other general knowledge?

# Desiderate for NLP Models

- Sensitivity to a wide range of phenomena and constraints in human language
- Generality across languages, modalities, genres, styles
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)
- High accuracy when judged against expert annotations or test data
- Ethical

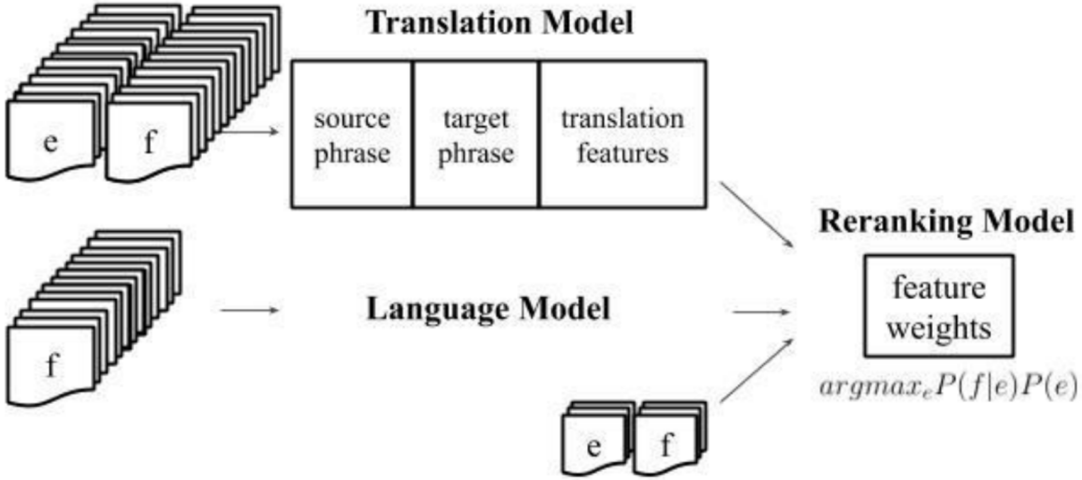
# Symbolic and Probabilistic NLP

**Logic-based/Rule-based NLP**



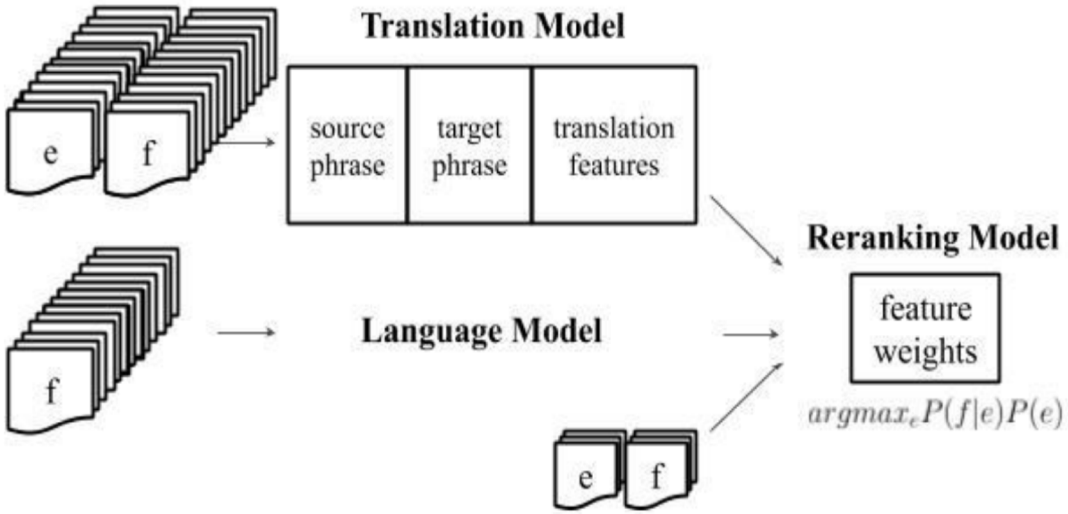

**~ 90s**

**Statistical NLP**



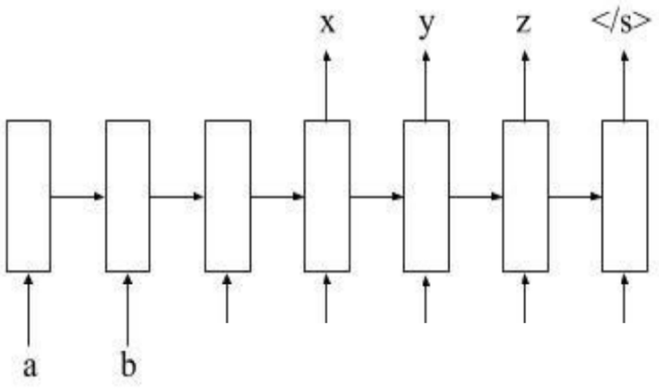
# Probabilistic and Connectionist NLP

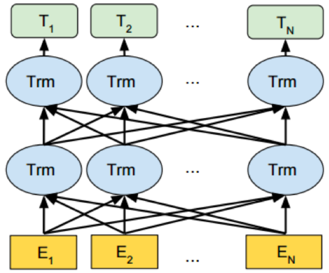
**Engineered Features/Representations**




**~mid 2010s**

**Learned Features/Representations**




# NLP vs. Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- $\mathcal{R}$ 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
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Data science
- Political science
- Psychology
- Economics
- Education

# Today's Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature

# Factors Changing NLP Landscape

1. Increases in computing power
2. The rise of the web, then the social web
3. Advances in machine learning
4. Advances in understanding of language in social context

# Logistics

# What is this Class?

- **Linguistic Issues**
  - What are the range of language phenomena?
  - What are the knowledge sources that let us disambiguate?
  - What representations are appropriate?
  - How do you know what to model and what not to model?
- **Statistical Modeling Methods**
  - Increasingly complex model structures
  - Learning and parameter estimation
  - Efficient inference: dynamic programming, search
  - Deep neural networks for NLP: LSTM, CNN, Seq2seq

# Outline of Topics

- Words and Sequences
  - Text classifications
  - Probabilistic language models
  - Vector semantics and word embeddings
  - Sequence labeling: POS tagging, NER
  - HMMs, Speech recognition
- Parsers
- Semantics
- Applications
  - Machine translation, Question Answering, Dialog Systems

# Readings

- Books:
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd or 3 rd Edition
    - https://web.stanford.edu/~jurafsky/slp3/
  - Also: Eisenstein, Natural Language Processing
    - https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

# Course Website & Piazza

www.cc.gatech.edu/classes/AY2020/cs7650_spring/

piazza.com/gatech/spring2020/cs7650cs4650

# Your Instructors

- Instructor:
  - Diyi Yang
    - Assistant professor
    - Research interests: NLP, Computational Social Science
- TAs:
  - Ian Stewart: PhD, Computational Sociolinguistics
  - Jiaao Chen: PhD, NLP/ML
  - Nihal Singh: MSCS, NLP

# TA Office Hours

- Ian Stewart: Tuesdays, 2-4pm, CODA C1106
- Jiaao Chen: Thursdays, 2-4pm
- Nihal Singh: Fridays, 9-11am

# Grading

- 4 Homework Assignments (45%)
- 1 Midterm (15%)
- 1 Course Project (40%)

- 45% Homework Assignments
  - Homework 1: 6%
  - Homework 2: 13%
  - Homework 3: 13%
  - Homework 4: 13%
- 15% Midterm Exam
  - No make-up exam unless under emergency situation
- 40% Course Project
  - Project proposal (2 pages): 5%
  - Midway report (4 pages): 10%
  - Final report (8 pages): 20%
  - Presentation (in class presentation): 5%

# Late Polices

- Late Policy
  - 4 late days to use over the duration of the semester for homework assignments only. There are no restrictions on how the late days can be used (e.g., all 4 can be used on one homework). Using late days will not affect your grade. But homework submitted late after all 4 late days have been used will receive no credit.

- No make-up exam
  - Unless under emergency situation

# Course Project

- Semester-long project (2-3 students) involving natural language processing – either focusing on core NLP methods or using NLP in support of an empirical research question
  - 2-page Project proposal (5%)
  - 4-page Midway report (10%)
  - 8-page Final report (20%)
  - Project presentation (5%)
    - 10-min in-class presentation (tentative)

# Other Announcements

- Course Contacts:
  - Webpage: materials and announcements
  - Piazza: discussion forum
  - Homework questions: Piazza, TAs' office hours
- Computing Resources:
  - Experiments can take up to hours, even with efficient computation
  - Recommendation: **start assignments early**

# What's Next?

- Text Classification