



STC

Natural Language Processing

Introduction

Tapas Kumar Mishra

A black and white photograph of a large, modern building with a curved glass facade and a central entrance. The building has a prominent white overhang or canopy over the entrance. The foreground is a paved area with some low-lying plants. The sky is overcast.

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

Google Translate

CHINESE - DETECTED ↔ ENGLISH

我学习深度学习和机器学习 ×

Wǒ xuéxí shēndù xuéxí hé jīqì xuéxí

🔊 Send feedback

I study deep learning and machine learning. ☆

🔊 ⋮

Google Translate

Text Documents

DETECT LANGUAGE ENGLISH SPANISH FRENCH ↔ ENGLISH SPANISH ARABIC ▼

Search languages

_DETECT LANGUAGE	ENGLISH	SPANISH	FRENCH	ENGLISH	SPANISH	ARABIC
<input checked="" type="checkbox"/> Detect language ⊕	Czech	Hebrew	Latin	Portuguese	Tajik	
Afrikaans	Danish	Hindi	Latvian	Punjabi	Tamil	
Albanian	Dutch	Hmong	Lithuanian	Romanian	Telugu	
Amharic	English	Hungarian	Luxembourgish	Russian	Thai	
Arabic	Esperanto	Icelandic	Macedonian	Samoan	Turkish	
Armenian	Estonian	Igbo	Malagasy	Scots Gaelic	Ukrainian	
Azerbaijani	Filipino	Indonesian	Malay	Serbian	Urdu	
Basque	Finnish	Irish	Malayalam	Sesotho	Uzbek	
Belarusian	French	Italian	Maltese	Shona	Vietnamese	
Bengali	Frisian	Japanese	Maori	Sindhi	Welsh	
Bosnian	Galician	Javanese	Marathi	Sinhala	Xhosa	
Bulgarian	Georgian	Kannada	Mongolian	Slovak	Yiddish	
Catalan	German	Kazakh	Myanmar (Burmese)	Slovenian	Yoruba	
Cebuano	Greek	Khmer	Nepali	Somali	Zulu	
Chichewa	Gujarati	Korean	Norwegian	Spanish		
Chinese	Haitian Creole	Kurdish (Kurmanji)	Pashto	Sundanese		
Corsican	Hausa	Kyrgyz	Persian	Swahili		
Croatian	Hawaiian	Lao	Polish	Swedish		

Natural Language Processing

Applications

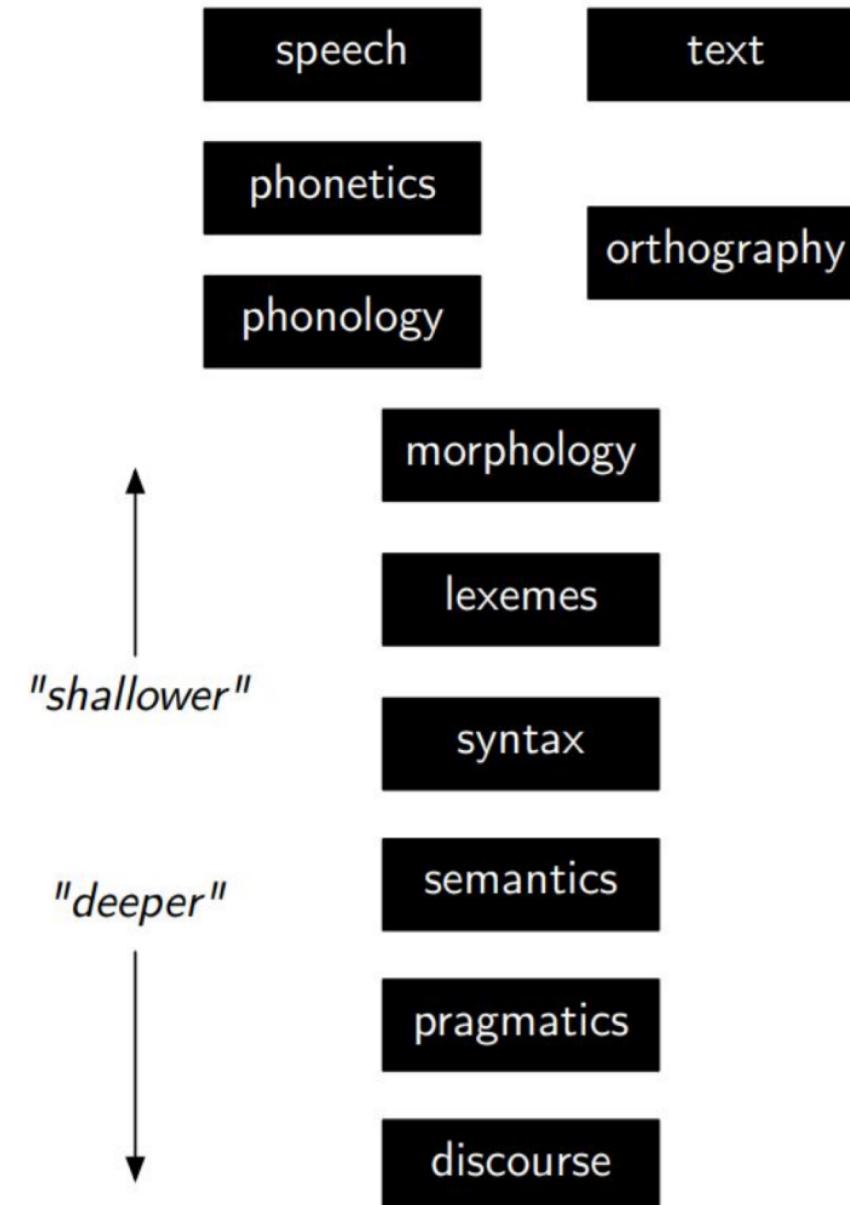
- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

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

WORDS	This is a simple sentence
MORPHOLOGY	be 3sg present

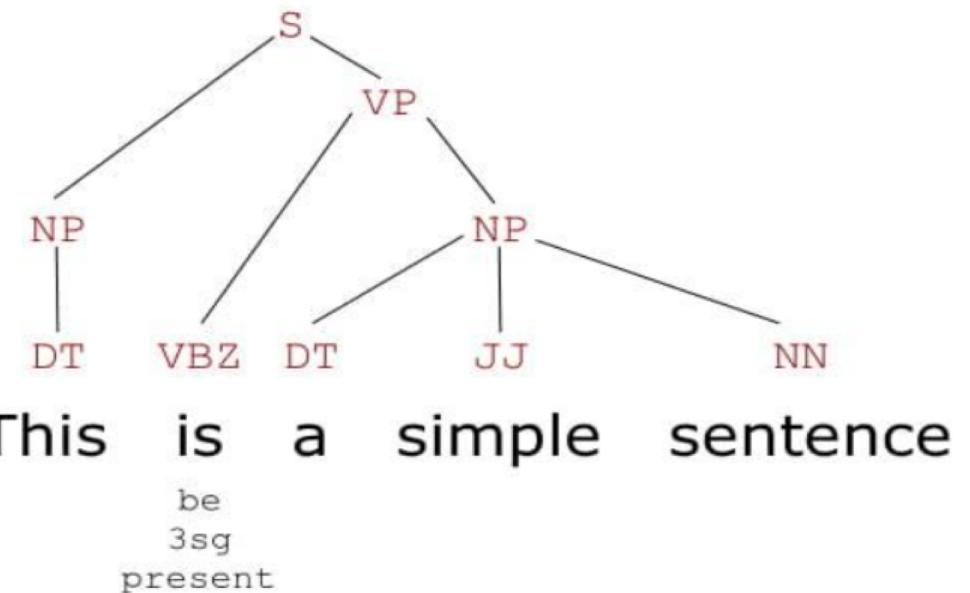
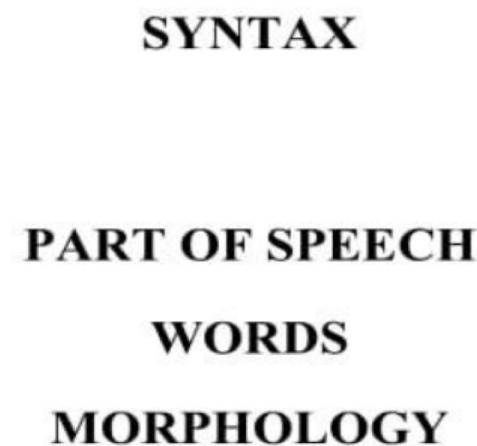
Part of Speech

- Part of speech tagging

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

Syntax

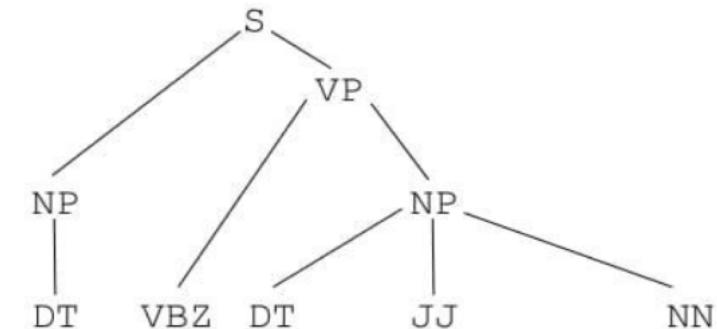
- Syntactic parsing



Semantics

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

SYNTAX



PART OF SPEECH

WORDS

MORPHOLOGY

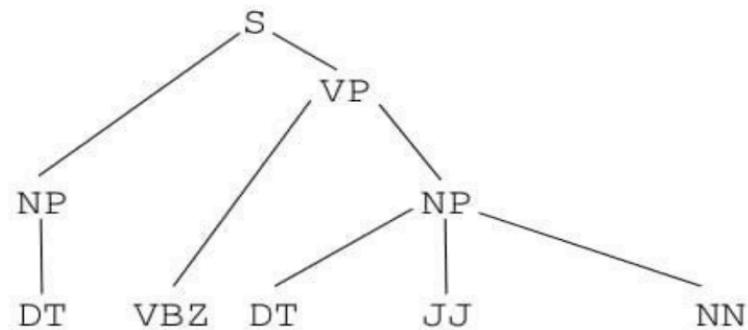
SEMANTICS

This is a simple sentence

be 3sg present SIMPLE1 having few parts SENTENCE1 string of words satisfying the grammatical rules of a language

Discourse

SYNTAX



PART OF SPEECH

WORDS

This is a simple sentence

be
3sg
present

SIMPLE1
having
few
parts

SENTENCE1
string of words
satisfying the
grammatical rules
of a language

MORPHOLOGY

SEMANTICS

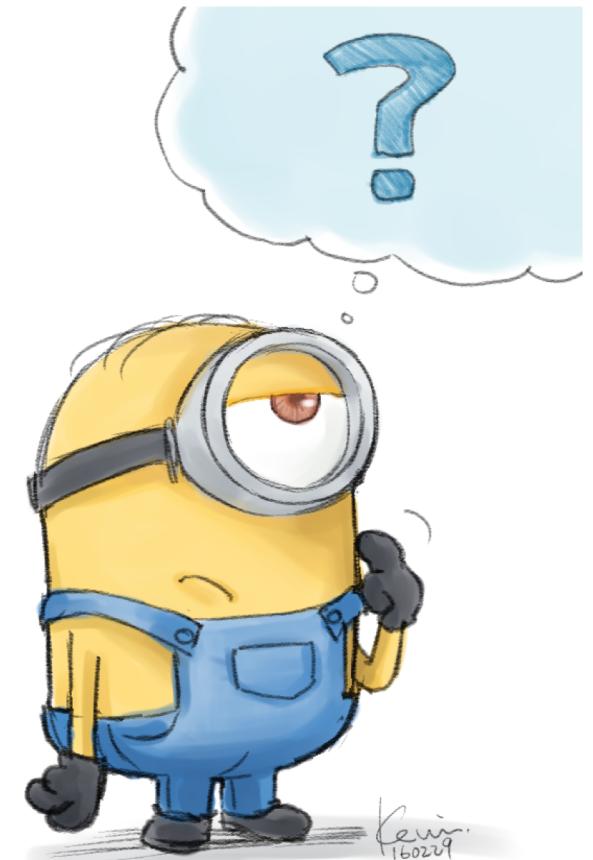
CONTRAST

DISCOURSE

But it is an instructive one.

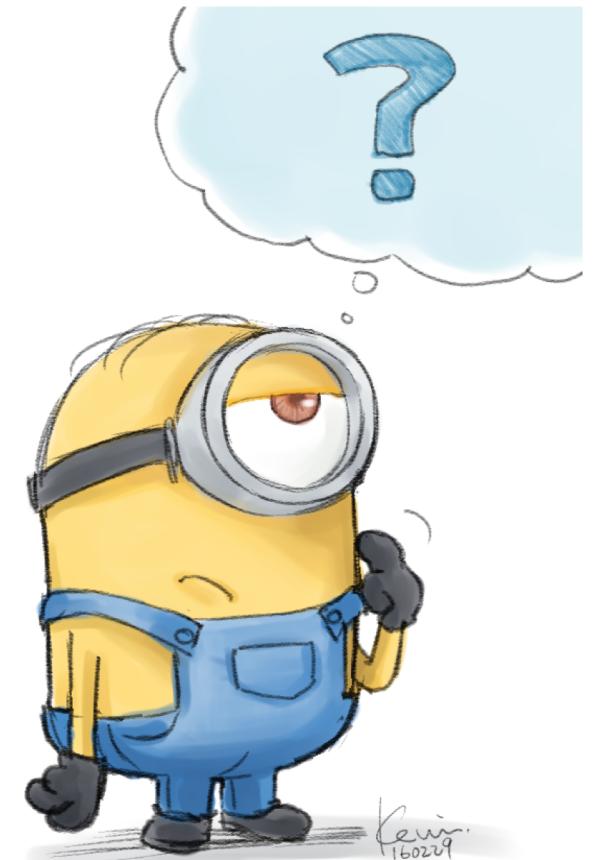
Why NLP is Hard?

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**



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

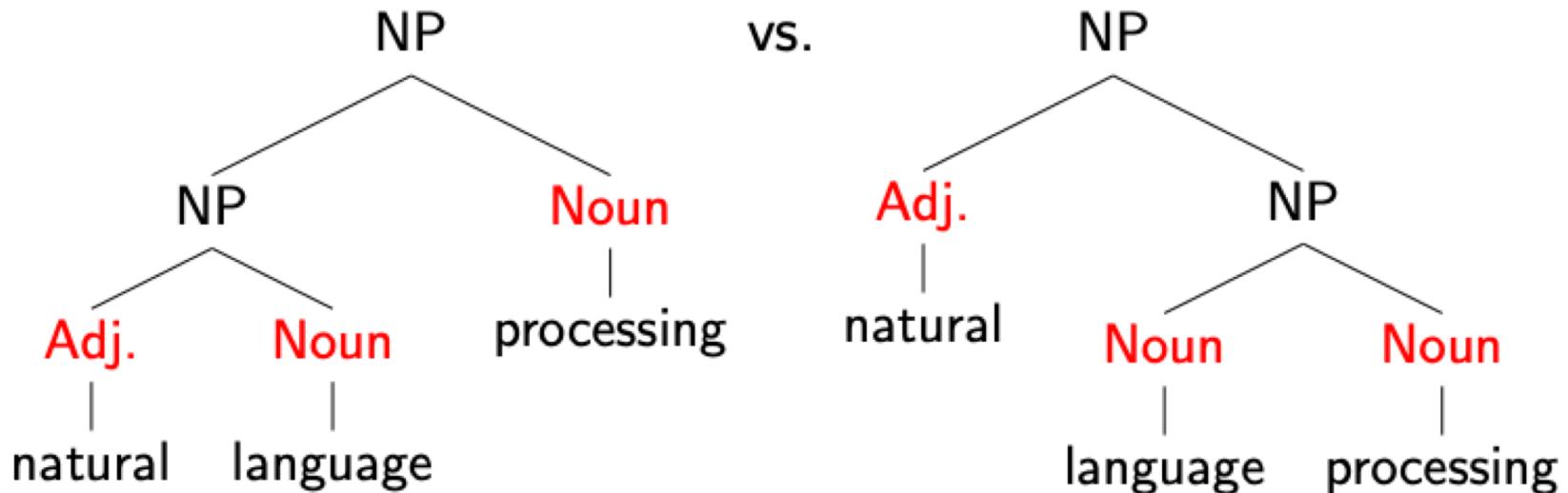
I know, right shake my head for your
ikr smh he asked fir yo last name

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



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?

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



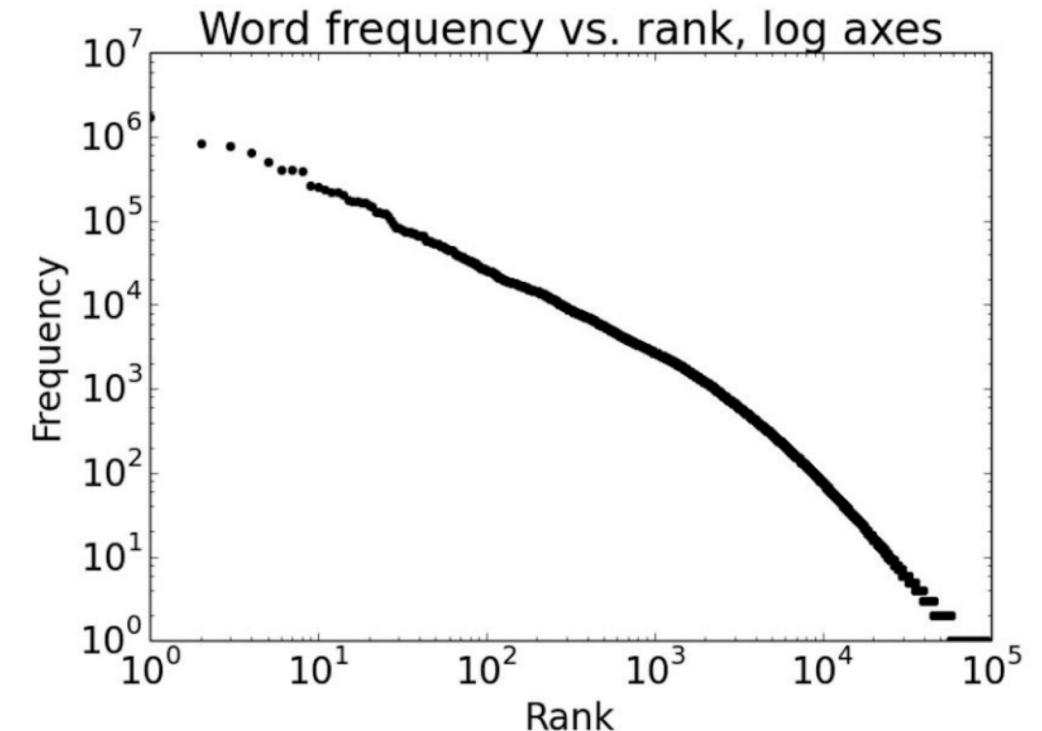
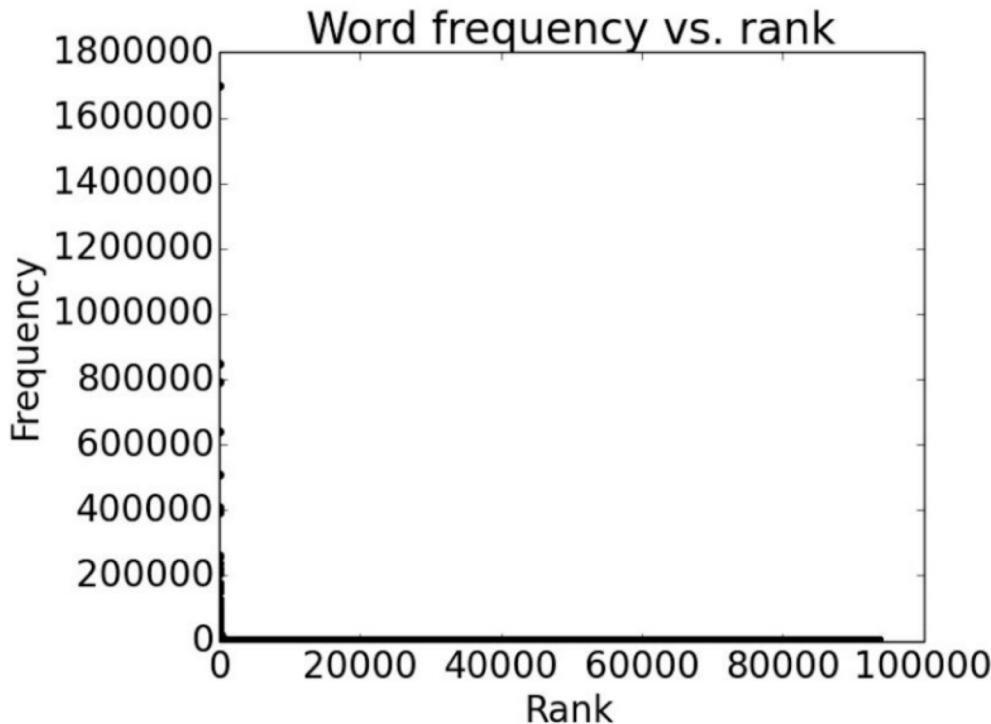
Sparsity

- 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

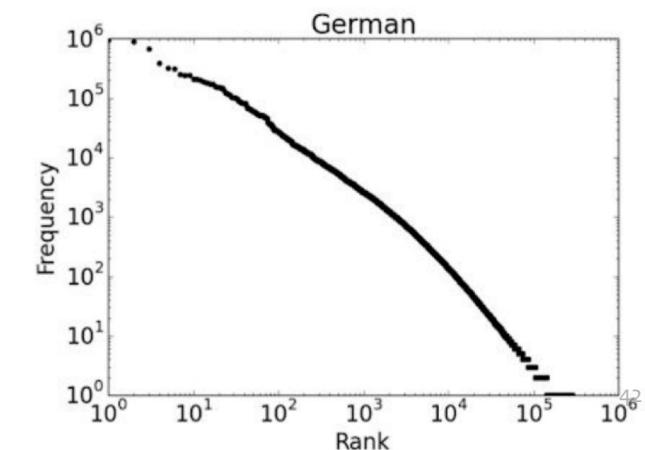
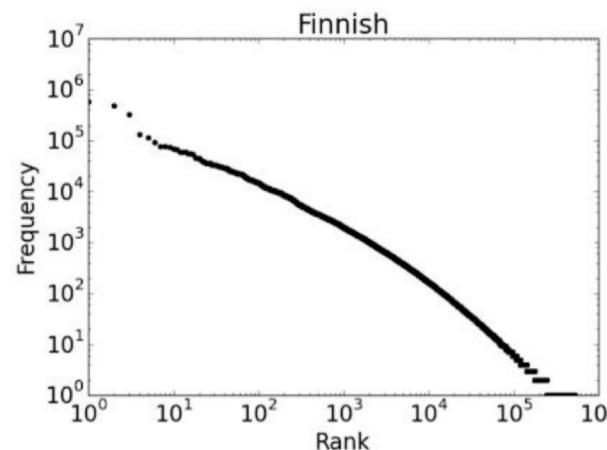
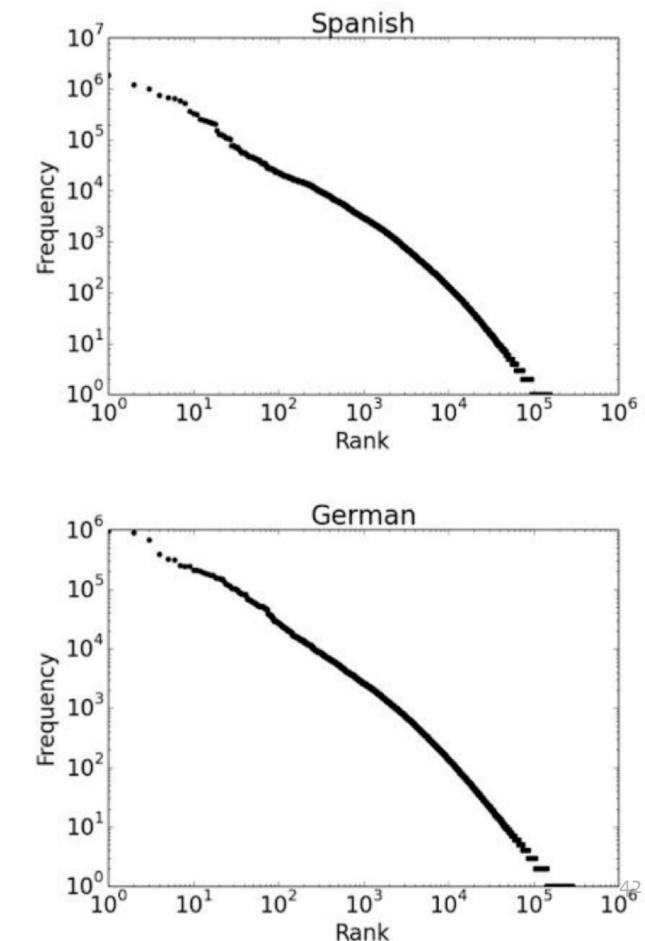
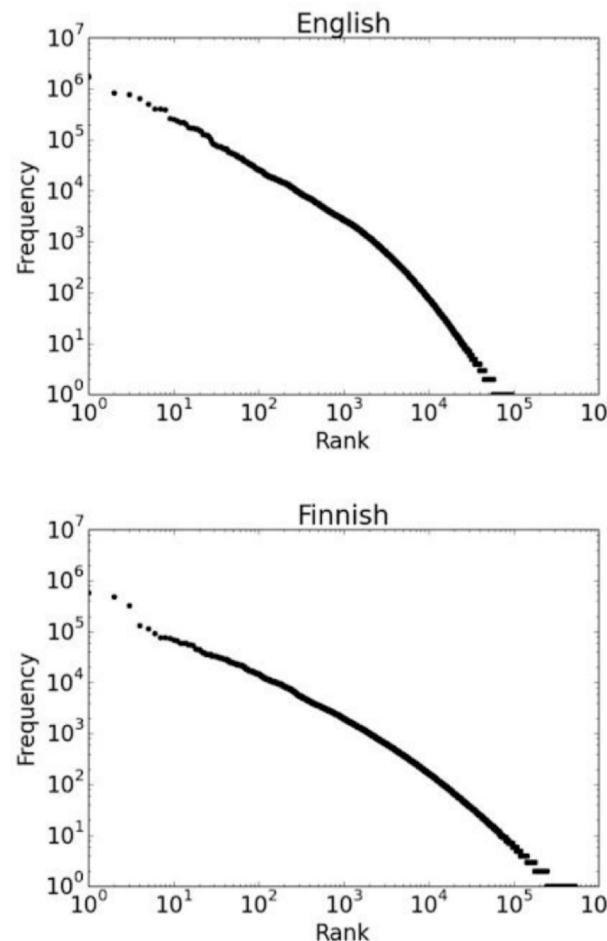
Sparsity

- 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



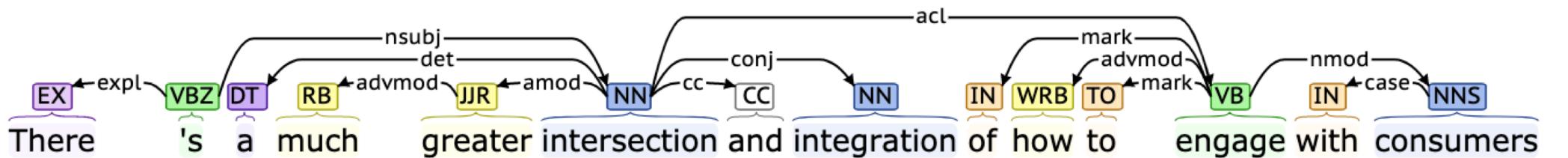
Why NLP is Hard?

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



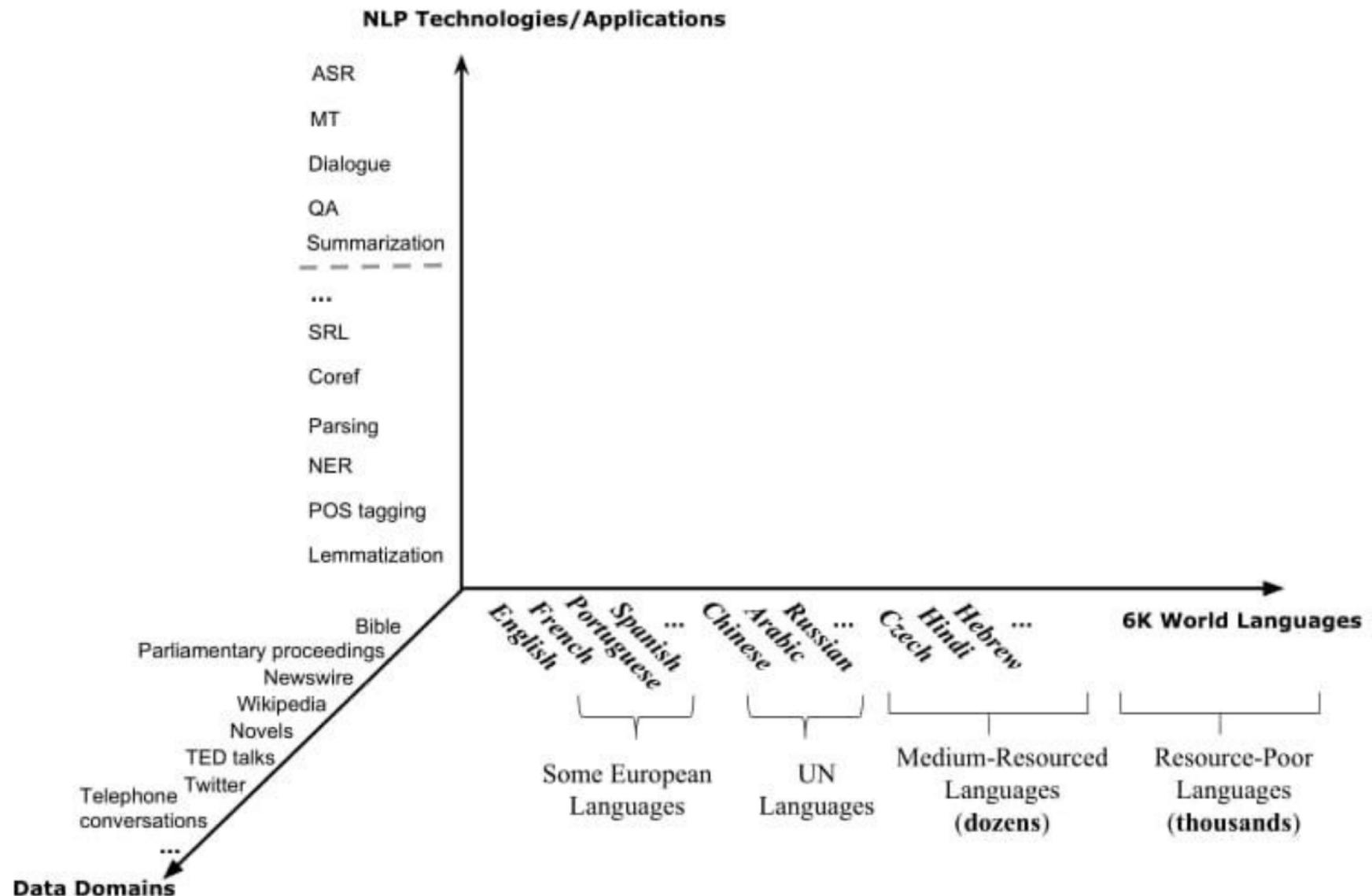
Variation

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



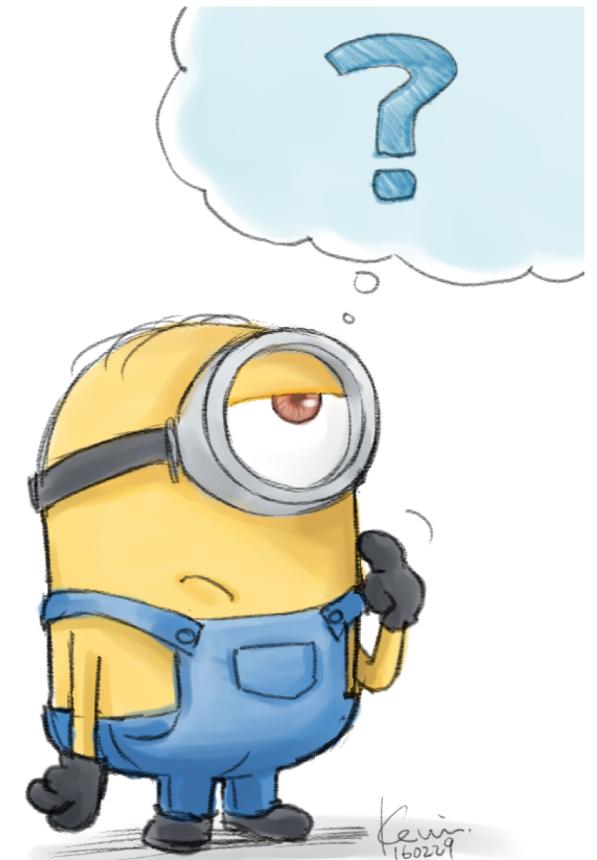
- What will happen if we try to use this tagger/parser for **social media**?
 - "ikr smh he asked fir yo last name so he can add u on fb lololol"

Variation



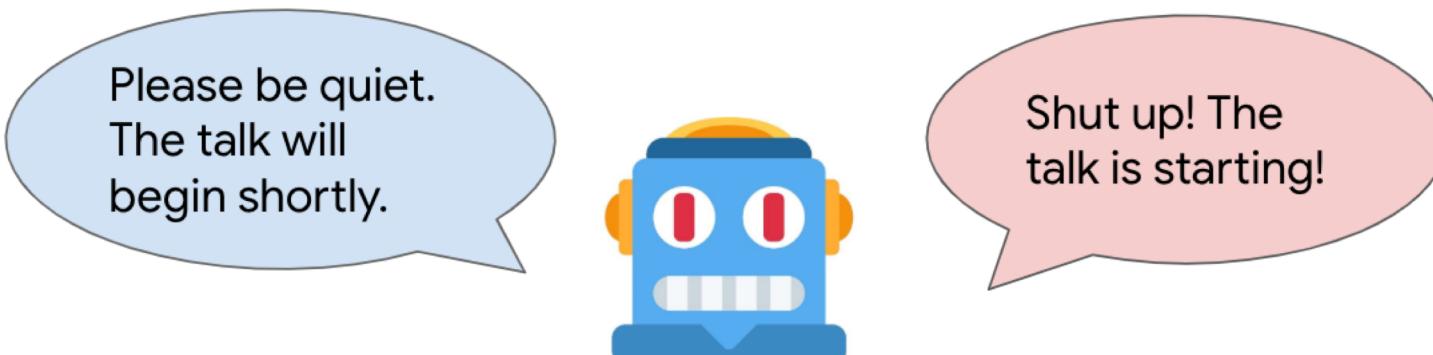
Why NLP is Hard?

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



Expressivity

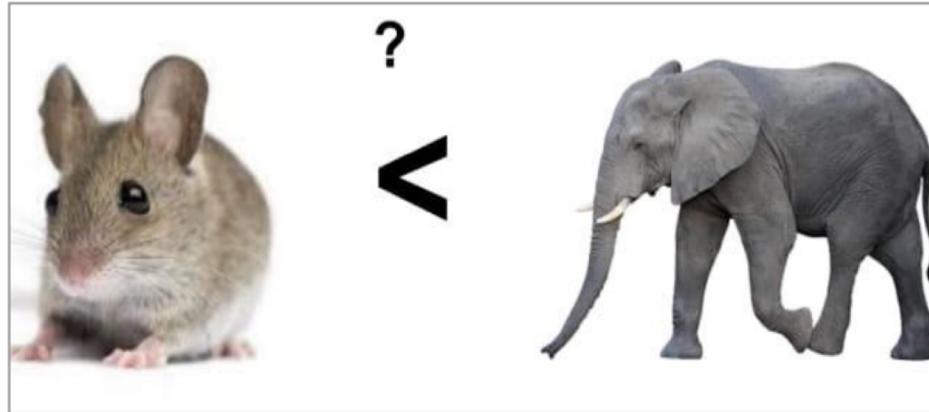
- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
 - *She gave the book to Tom* vs. *She gave Tom the book*
 - *Some kids popped by* vs. *A few children visited*
 - *Is that window still open?* vs. *Please close the window*



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

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