Introduction

CS4248 Natural Language Processing

Week 01

Min-Yen KAN

1

Many slides borrowed with permission from Diyi Yang (Georgia Tech), Yulia Tsvetkov (CMU) and Noah Smith (UW)

Week 01 Agenda

What is NLP?

Why NLP?

Levels of Linguistic Knowledge

Why is NLP Hard?

Connections to Other Fields

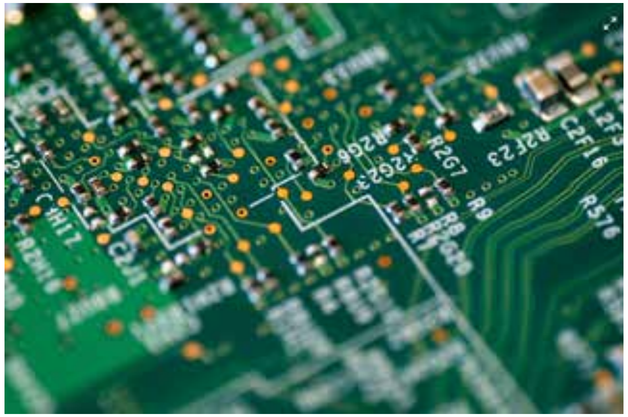
What are We Going to Learn?

Administrivia and Course Organization

What is NLP?

What does it mean to “know” a language?

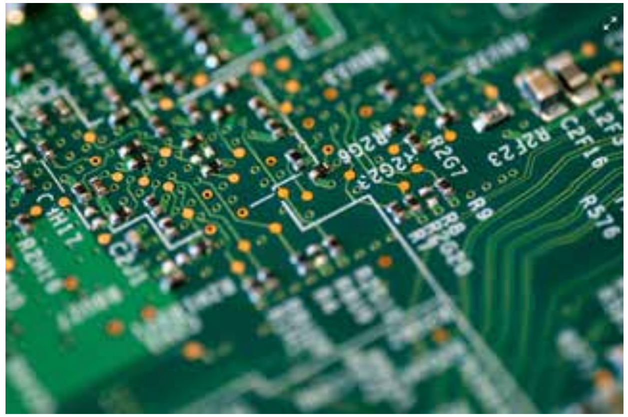
School of Computing



Natural Language

$\mathcal{R}$

Photos from Saketh Garuda and Magnus Engø @ Unsplash.
Slide Adapted from Noah Smith (UW)



Analysis

Natural Language

Generation

$\mathcal{R}$

Photos from Saketh Garuda and Magnus Engø @ Unsplash.
Slide Adapted from Noah Smith (UW)

Why NLP?

What do we use it for?

Slides adapted from Diyi Yang (GaTech)

Communication with Machines

~1960s

~1980s

Today

Slides adapted from Diyi Yang (GaTech)

Conversational Agents

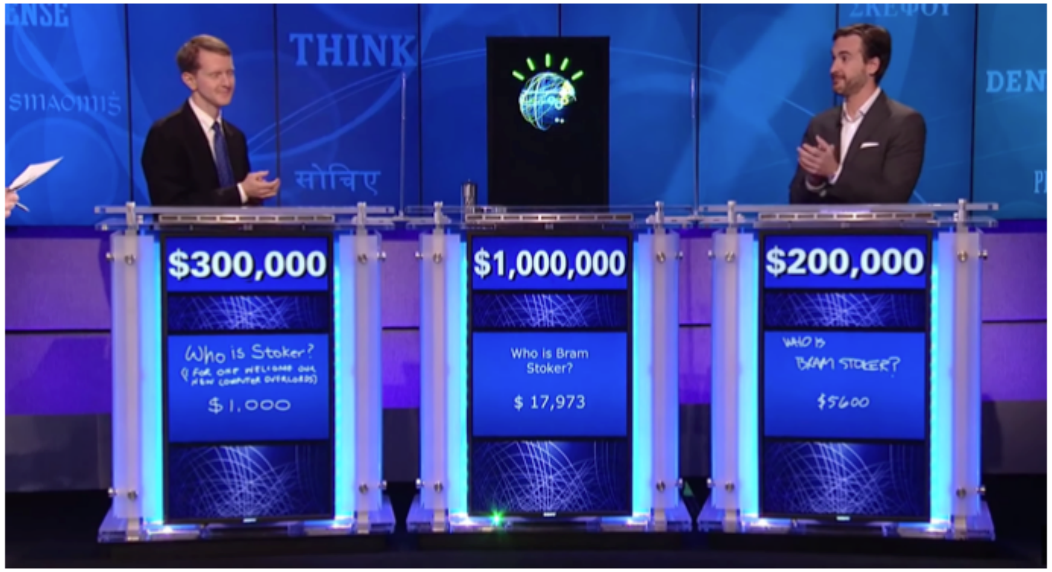
Conversational Agents contain:

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

Slides adapted from Diyi Yang (GaTech)

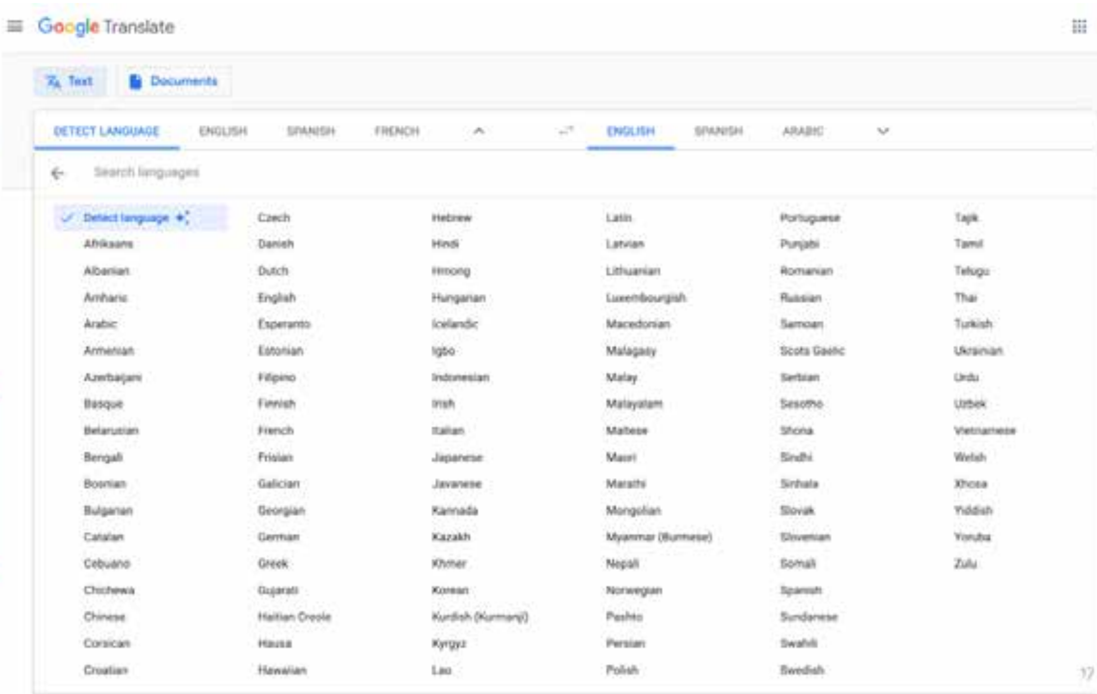
Question Answering

- 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?



Slides adapted from Diyi Yang (GaTech)

Machine Translation



Slides adapted from Diyi Yang (GaTech)

Natural Language Processing

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization

Core Technologies

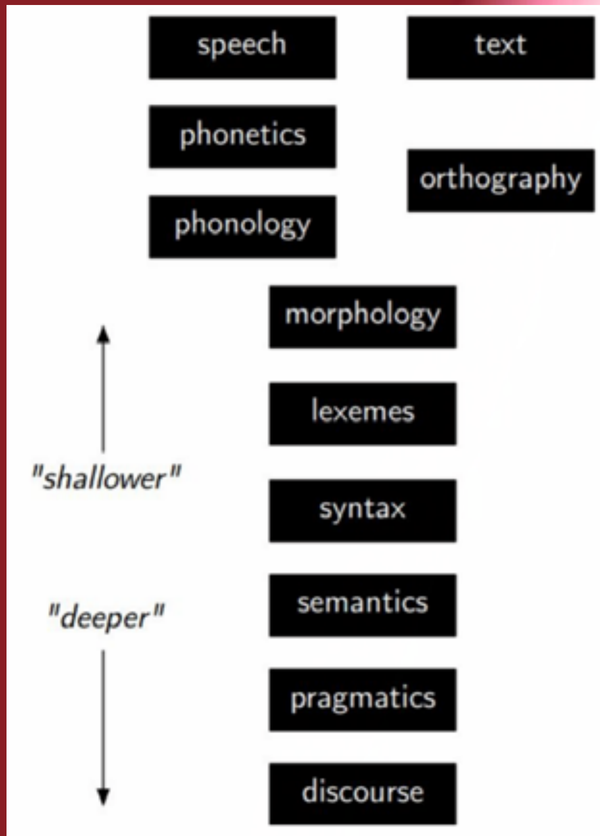
- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Word sense disambiguation
- Semantic role labeling

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

Slides adapted from Diyi Yang (GaTech)

Levels of Linguistic Knowledge

Introduction



Slide Adapted from Noah Smith (UW)

Phonetics, Phonology

- Pronunciation Modeling

SOUNDS Th i a si e n

Slide Adapted from Noah Smith (UW)

Words

- Language Modeling
- Tokenization
- Spelling Correction

WORDS This is a simple sentence

Slide Adapted from Noah Smith (UW)

Morphology

- Morphology Analysis
- Tokenization
- Lemmatization

WORDS This is a simple sentence

MORPHOLOGY be 3sg present

Slide Adapted from Noah Smith (UW)

Part of Speech

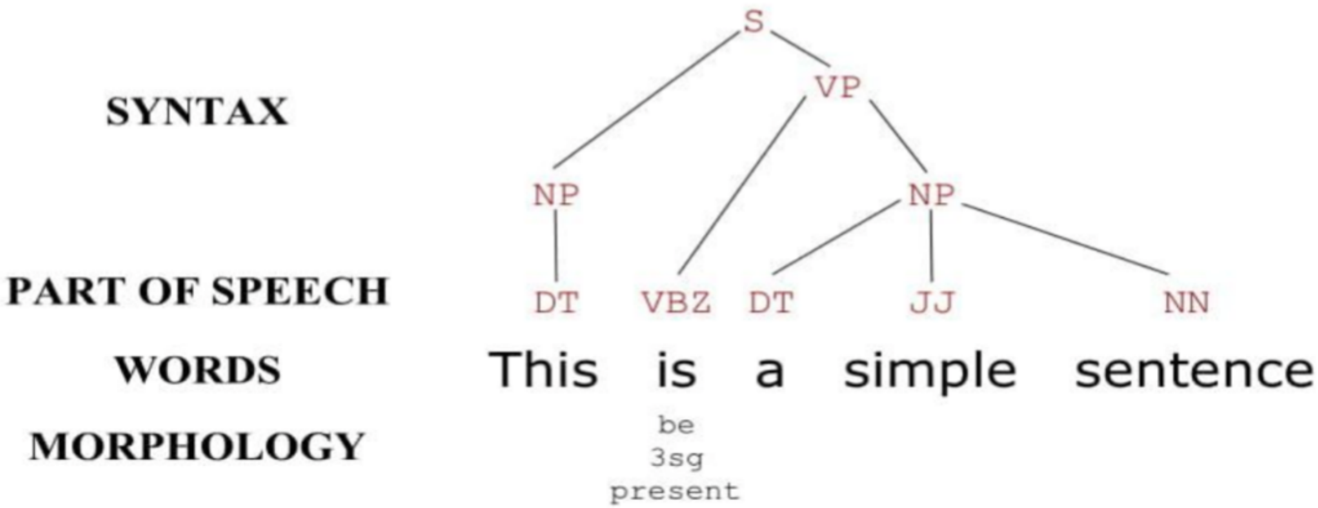
- Part of Speech Tagging

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

Slide Adapted from Noah Smith (UW)

Syntax

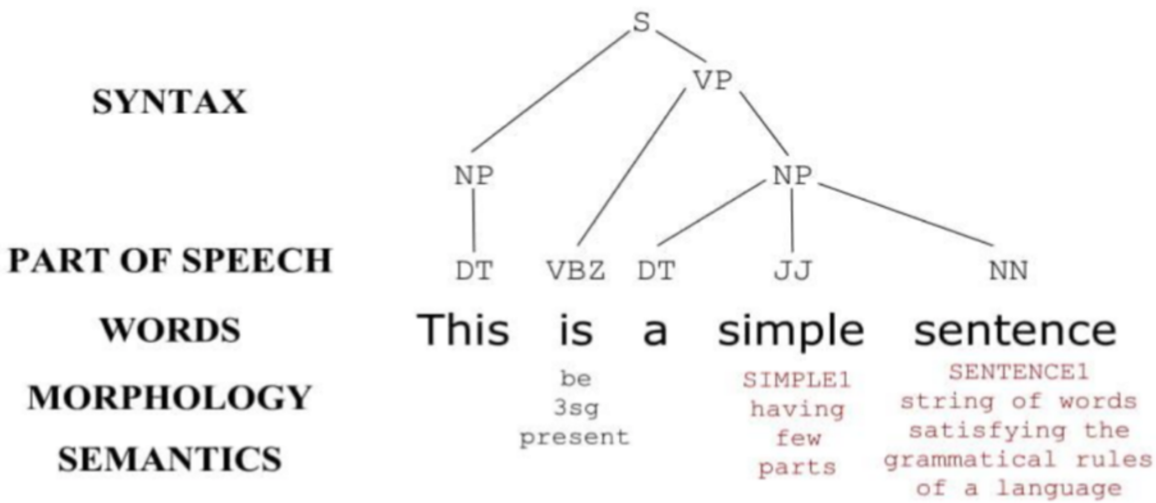
- Syntactic Parsing



Slide Adapted from Noah Smith (UW)

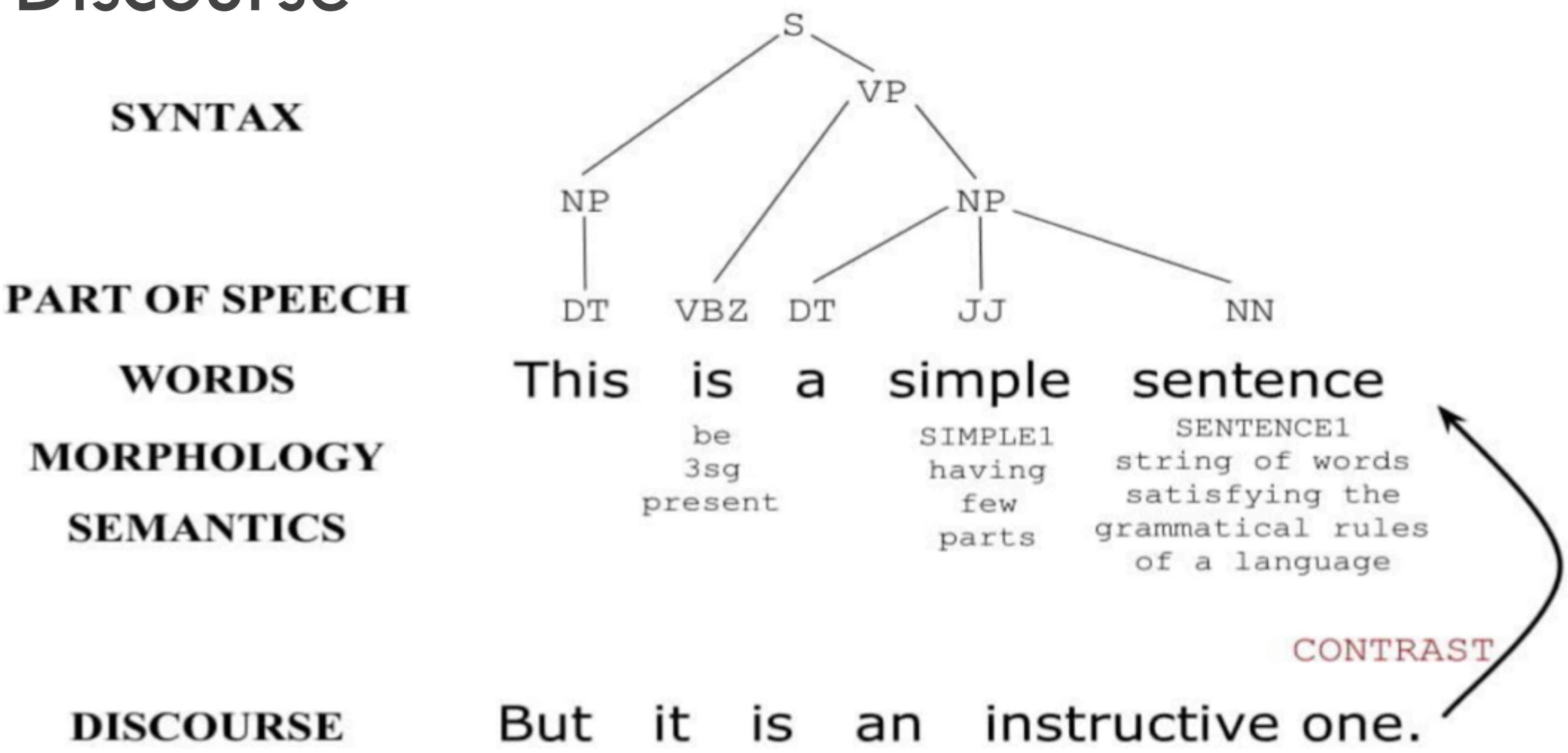
Semantics

- Named Entity Recognition
- Word Sense Disambiguation
- Semantic Role Labeling



Slide Adapted from Noah Smith (UW)

Discourse



Slide Adapted from Noah Smith (UW)

Why is NLP Hard?

Ambiguity and Scale

- Ambiguity
- Scale
- Sparsity
- Variation
- Expressivity
- Unmodeled Variables
- Unknown Representations

Slides adapted from Diyi Yang (GaTech)

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*

Slides adapted from Diyi Yang (GaTech)

Words

- Segmenting text into words: ประธานาธิบดีทรัมป์ [Praṭhānāṭhibdī thrạmp̒]
- Morphological variation: *color*, *colour*, *ka ler*, *Manfuckinghattan*, *Twitterati*, *kiasuism*
- Words with multiple meanings: *bank*, *mean*, *POS*
- Domain-specific meanings: *latex*
- Multiword expressions: *make a decision*, *make out*

Slide Adapted from Noah Smith (UW)

Part of Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol

Slide Adapted from Noah Smith (UW)

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

Slide Adapted from Noah Smith (UW)

NUS School of Computing

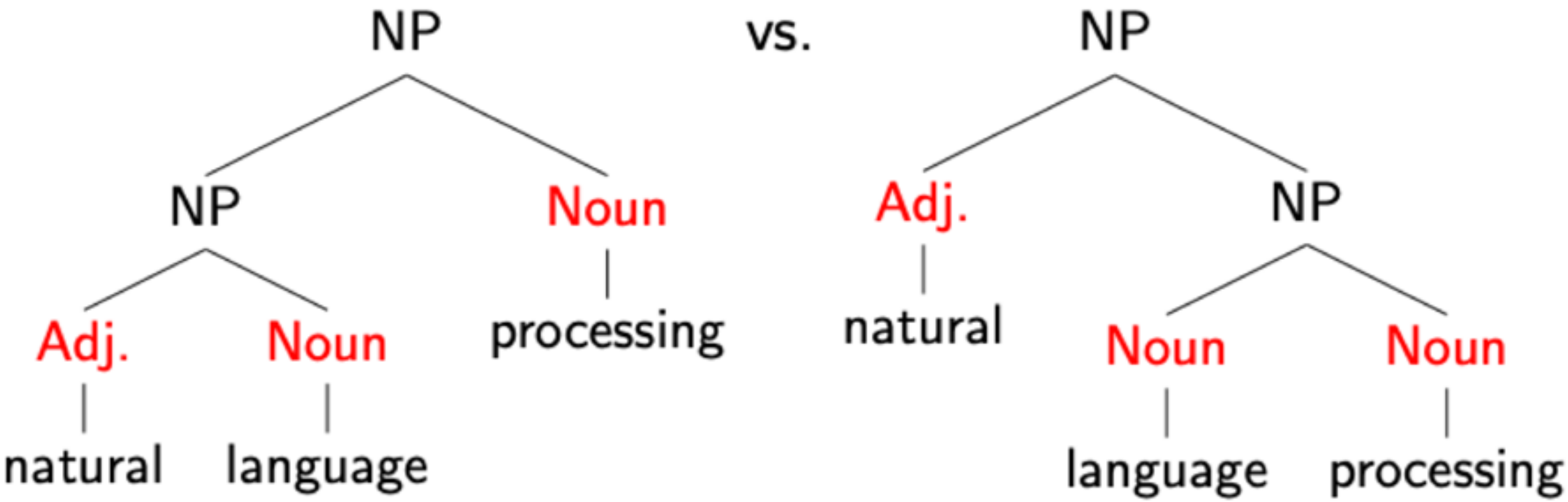
Part of Speech Tagging

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	

Slide Adapted from Noah Smith (UW)

Syntax



Slide Adapted from Noah Smith (UW)

Syntax + Semantics

We saw the woman with the telescope wrapped in paper.

Slide Adapted from Noah Smith (UW)

Syntax + Semantics

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?

Slide Adapted from Noah Smith (UW)

Semantics

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

Slide Adapted from Noah Smith (UW)

Semantics

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

– Groucho Marx

Which "woman" is that? Quantifier Scope

Slide Adapted from Noah Smith (UW)

Pragmatics

Do you know what time it is?

Do you want to come with me to the Esplanade?

What are the contexts of

- the speaker
- the hearer

Why is NLP Hard?

Sparsity

- Ambiguity
- Scale
- Sparsity
- Variation
- Expressivity
- Unmodeled Variables
- Unknown Representations

Slides adapted from Diyi Yang (GaTech)

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 Fr/En sentences
- Facebook Business reviews
- The Web!

Photo courtesy StickPNG.
Slides adapted from Diyi Yang (GaTech).

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

Slides adapted from Diyi Yang (GaTech)

Sparsity

Sparse data due to
Zipf's Law

Example: the frequency of different words in a large text corpus

any word

Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns

Frequency	Token
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

Slides adapted from Diyi Yang (GaTech)

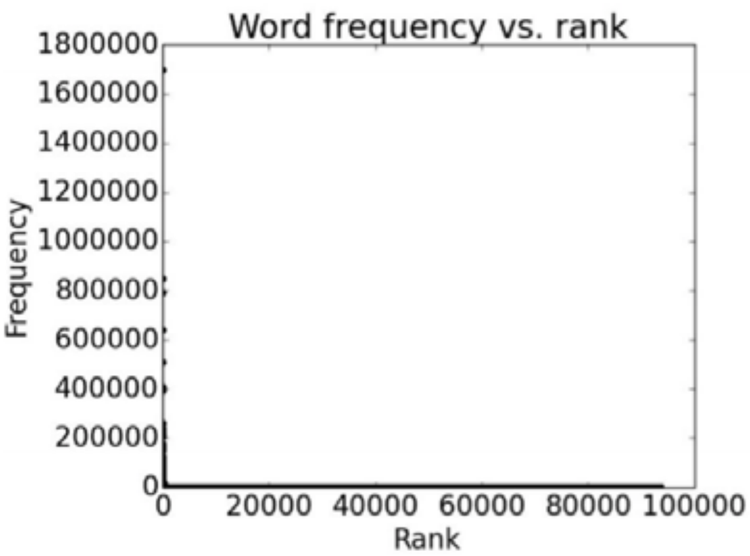
Sparsity

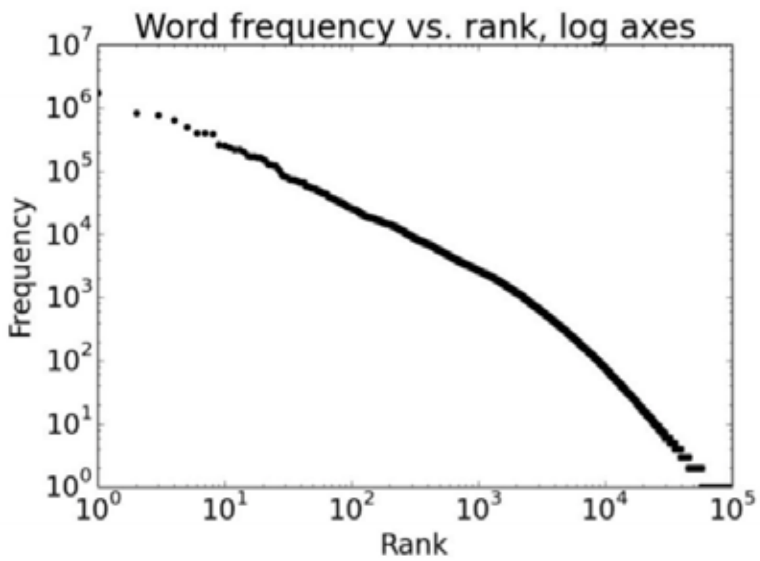
Order words by frequency. What is the frequency of nth ranked word?

Slides adapted from Diyi Yang (GaTech)

Sparsity

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



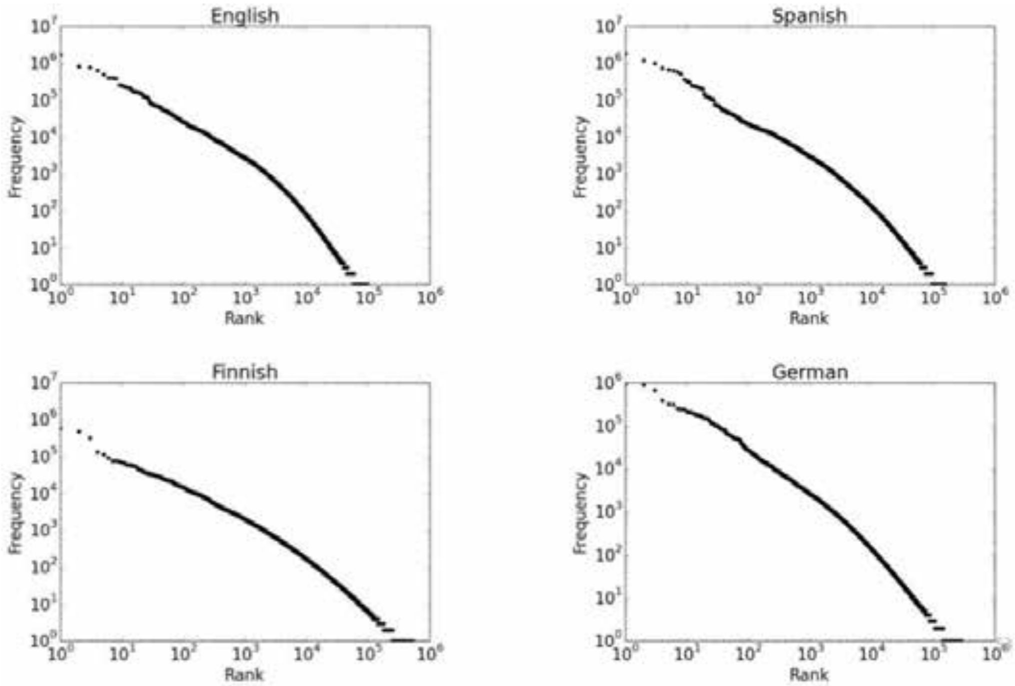


Slides adapted from Diyi Yang (GaTech)

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



Slides adapted from Diyi Yang (GaTech)

Why is NLP Hard?

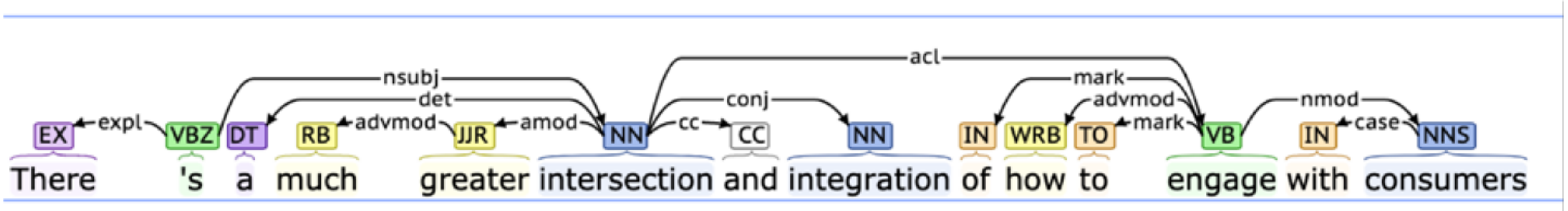
Variation

- Ambiguity
- Scale
- Sparsity
- Variation
- Expressivity
- Unmodeled Variables
- Unknown Representations

Slides adapted from Diyi Yang (GaTech)

Variation

Suppose we train a POS tagger or a parser on formal news

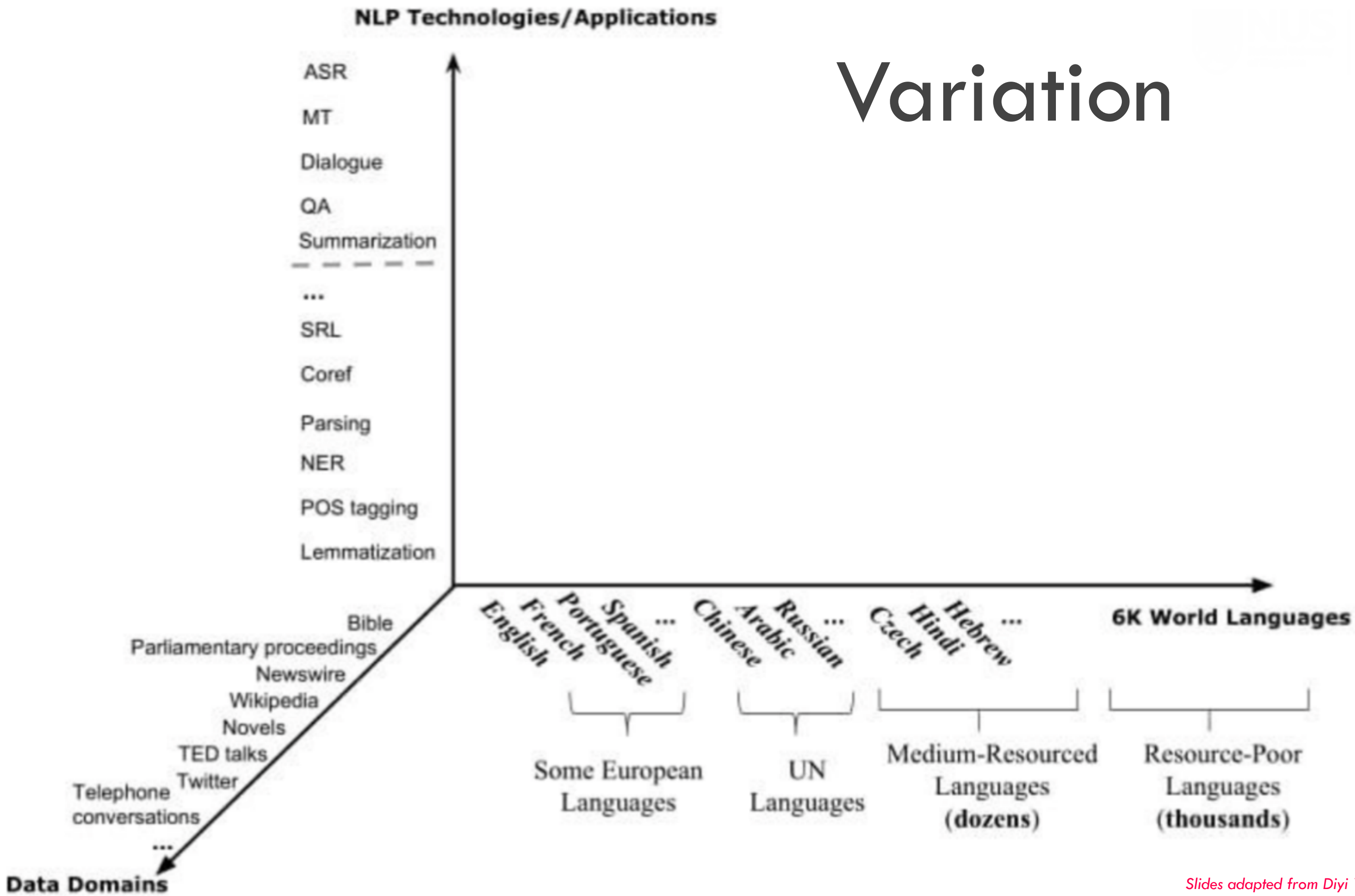


What will happen if we try to use this tagger/parser for social media?

ikr smh he asked fir yo last name so he can add u on fb lololol

Slides adapted from Diyi Yang (GaTech)

Variation



Slides adapted from Diyi Yang (GaTech)

Why is NLP Hard?

Expressivity, Unmodeled Variables and Unknown Representations

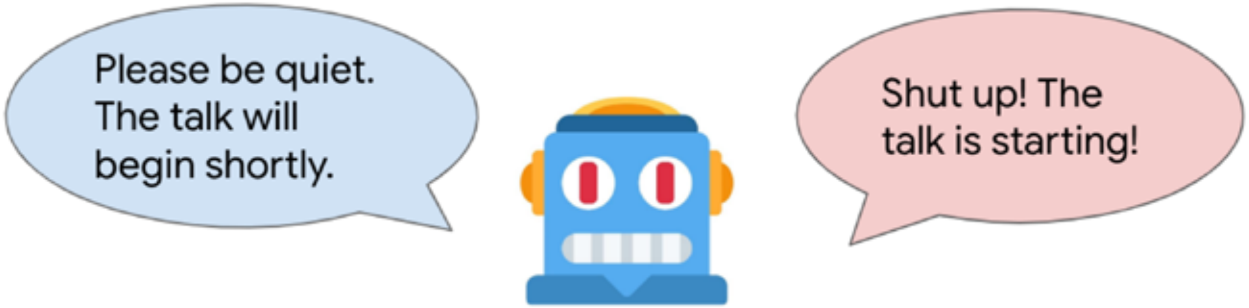
- Ambiguity
- Scale
- Sparsity
- Variation
- Expressivity
- Unmodeled Variables
- Unknown Representations

Slides adapted from Diyi Yang (GaTech)

Expressivity

Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:

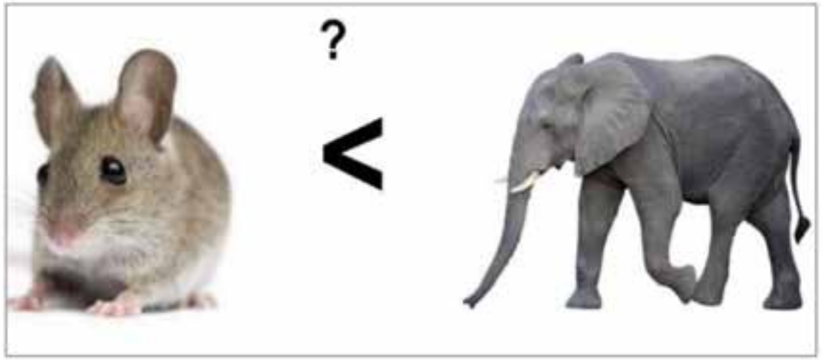
- *She gave the book to Tom* vs. *She gave Tom the book*
- *Some kids popped by* vs. *A few children visited*
- *Is that window still open?* vs. *Please close the window*



Slides adapted from Diyi Yang (GaTech)

Unmodeled Variables

“Drink this milk”



World Knowledge: [Winograd Schemas]

The trophy wouldn’t fit in the suitcase. It was too big.

The trophy wouldn’t fit in the suitcase. It was too small.

Slides adapted from Diyi Yang (GaTech)

Unmodeled Representation

Difficult to capture what is $\mathcal{R}$, as we don't even know how to represent the knowledge a human has or needs:

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

Slides adapted from Diyi Yang (GaTech)

Connections to Other Fields

Is NLP Machine Learning?

To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.

$\mathcal{R}$ is a theorized construct, not directly observable.

Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

Slide Adapted from Noah Smith (UW)

Is NLP Linguistics?

NLP must contend with NL data as found in the world.

NLP ≈ computational linguistics.

Linguistics now use tools originating in NLP!

Slide Adapted from Noah Smith (UW)

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

Slide Adapted from Noah Smith (UW)

What are We Going to Learn?

Overview of our course

Desiderata 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

Computational efficiency during training and testing (construction and production)

Explainable to human users; transparent

Ethical

Slide Adapted from Noah Smith (UW)

NLP is changing

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

Slide Adapted from Noah Smith (UW)

Course Meta Topics

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

Slide Adapted from Diyi Yang (GaTech)

Administrivia and Course Organization

Let's go over the website!

http://www.comp.nus.edu.sg/~cs4248