Natural Language Processing (CSE 490U): Introduction

Noah Smith

University of Washington
nasmith@cs.washington.edu

January 4, 2017

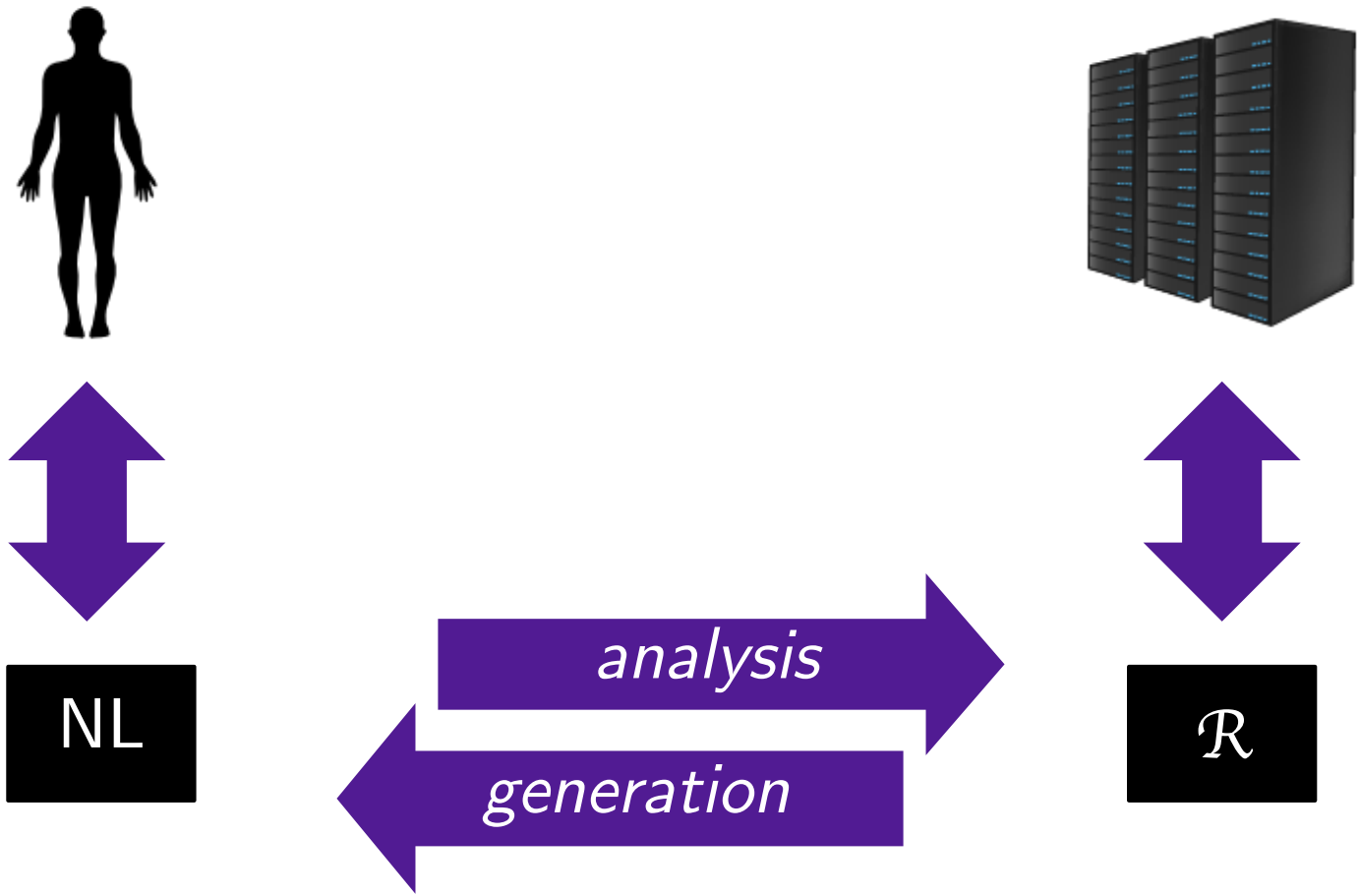
What is NLP?

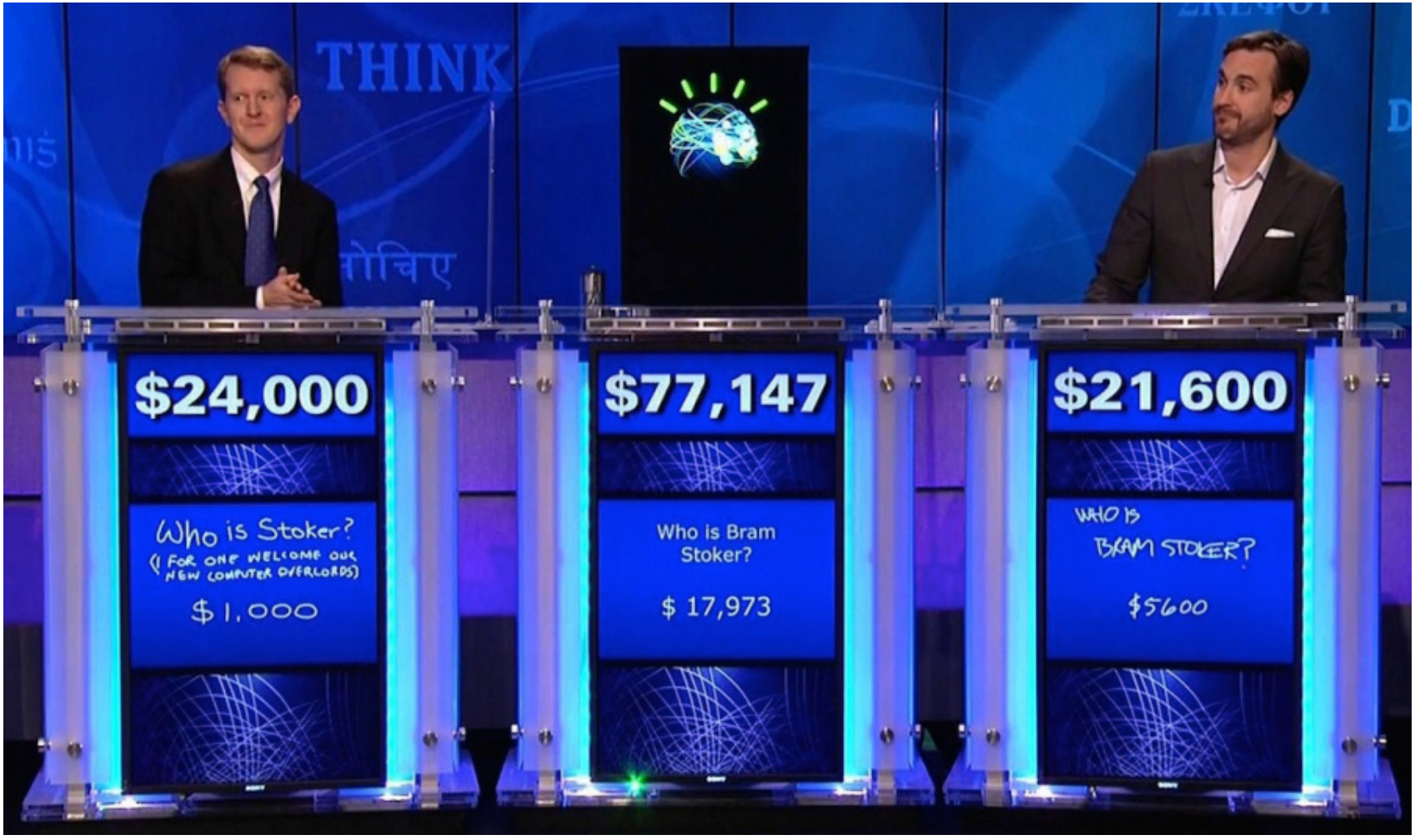
$\text{NL} \in \{\text{Mandarin Chinese}, \text{English}, \text{Spanish}, \text{Hindi}, \ldots, \text{Lushootseed}\}$

Automation of:

- analysis ($\text{NL} \rightarrow \mathcal{R}$)
- generation ($\mathcal{R} \rightarrow \text{NL}$)
- acquisition of $\mathcal{R}$ from knowledge and data

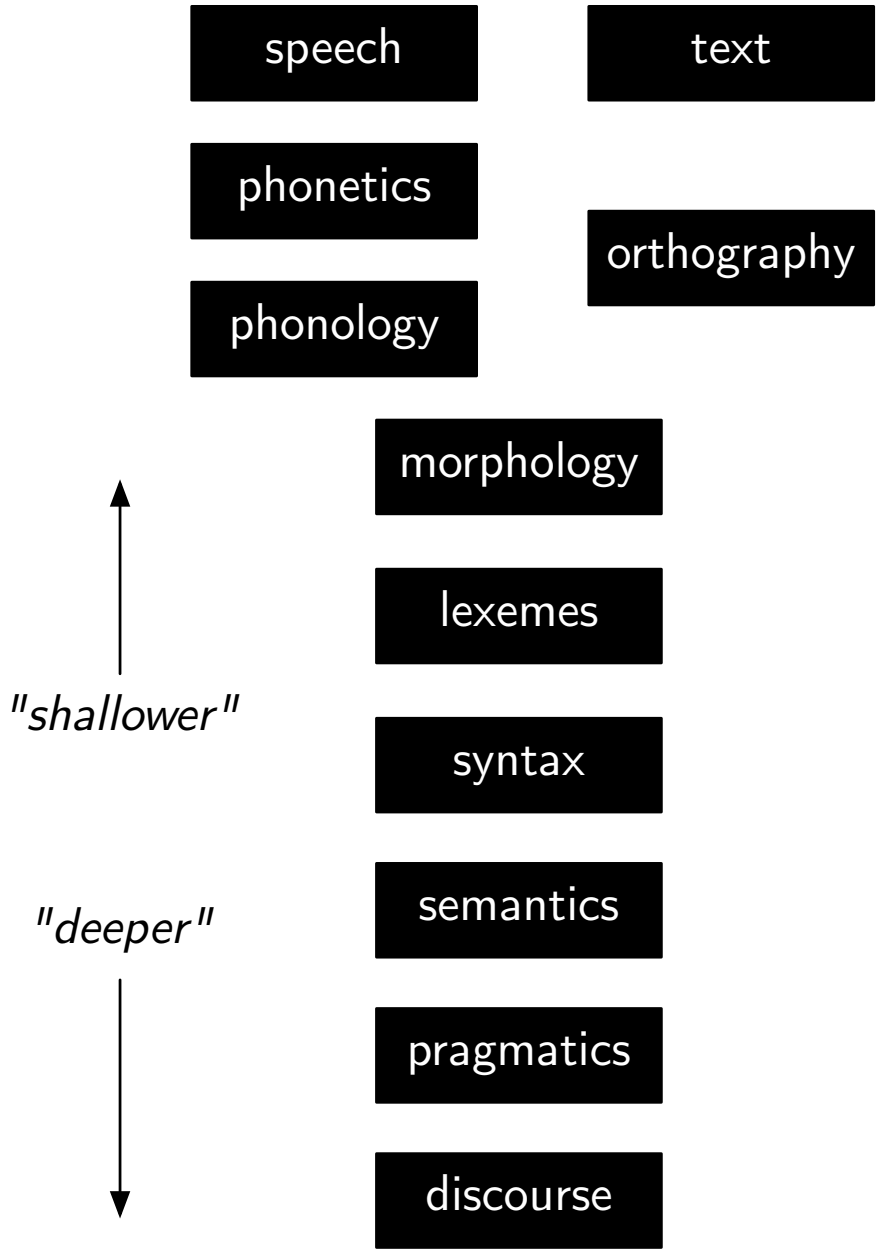
What is $\mathcal{R}$?


NL
analysis
generation
$\mathcal{R}$


THINK
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

What does it mean to “know” a language?

Levels of Linguistic Knowledge



Orthography

ลูกศิษย์วัดกระทิงยังยื้อปิดถนนทางขึ้นไปนมัสการพระบาทเขาคิชฌกูฏ หวิดปะทะกับเจ้าถิ่นที่ออกมาเผชิญหน้าเพราะเดือดร้อนสัญจรไม่ได้ ผวจ.เร่งทุกฝ่ายเจรจาก่อนที่ชื่อเสียงของจังหวัดจะเสียหายไปมากกว่านี้ พร้อมเสนอหยุดจัดงาน 15 วัน....

Morphology

uygarlaştıramadıklarımızdanmışsınızcasına
"(behaving) as if you are among those whom we could not civilize"

TIFGOSH ET HA-LELED BA-GAN
"you will meet the boy in the park"

unfriend, Obamacare, Manfuckinghattan

The Challenges of "Words"

- Segmenting text into words (e.g., Thai example)
- Morphological variation (e.g., Turkish and Hebrew examples)
- Words with multiple meanings: *bank*, *mean*
- Domain-specific meanings: *latex*
- Multiword expressions: *make a decision*, *take out*, *make up*, *bad hombres*

Example: Part-of-Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lolololol

Example: Part-of-Speech Tagging

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

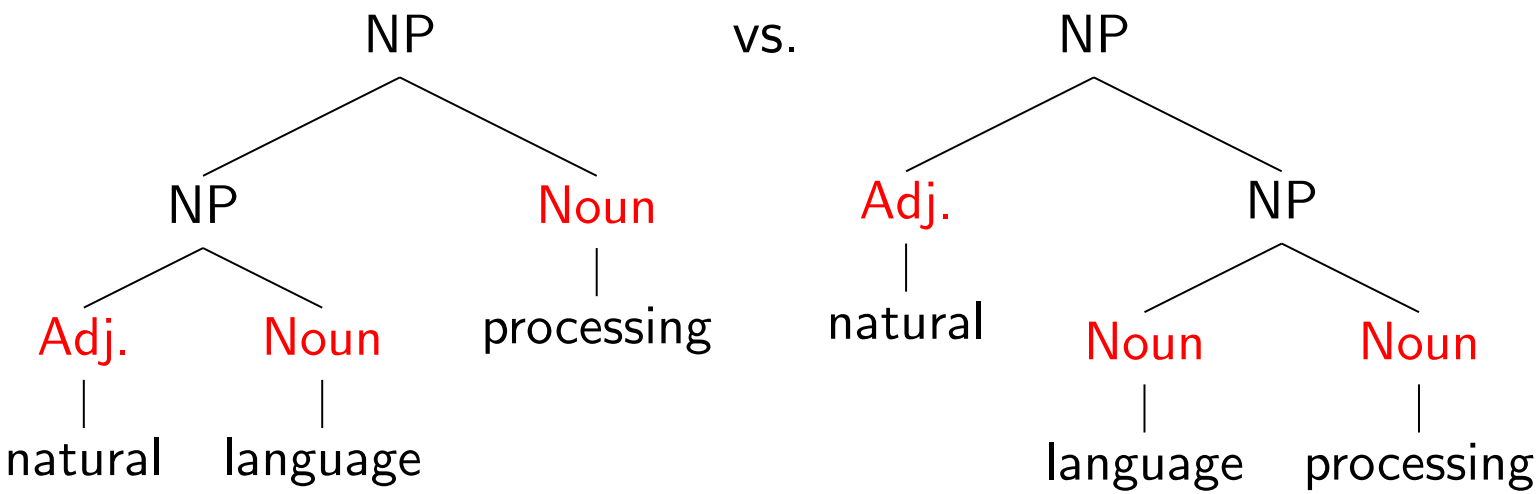
				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lolololol

Example: 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	lolololol
P	O	V	V	O	P	∧	!
preposition						proper noun	

Syntax


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

Morphology + Syntax

A ship-shipping ship, shipping shipping-ships.

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?

Syntax + Semantics

We saw the woman with the telescope wrapped in paper.

- Who has the telescope?
- Who or what is 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 event of perception, or an assault?

Semantics

Every fifteen minutes a woman in this country gives birth.

– Groucho Marx

Semantics

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

– Groucho Marx

Can $\mathcal{R}$ be "Meaning"?

Depends on the application!

- Giving commands to a robot
- Querying a database
- Reasoning about relatively closed, grounded worlds

Harder to formalize:

- Analyzing opinions
- Talking about politics or policy
- Ideas in science

Why NLP is Hard

1. Mappings across levels are complex.
   - A string may have many possible interpretations in different contexts, and resolving **ambiguity** correctly may rely on knowing a lot about the world.
   - **Richness**: any meaning may be expressed many ways, and there are immeasurably many meanings.
   - Linguistic **diversity** across languages, dialects, genres, styles, ...
2. Appropriateness of a representation depends on the application.
3. Any $\mathcal{R}$ is a theorized construct, not directly observable.
4. There are many sources of variation and noise in linguistic input.

Desiderata for NLP Methods

(ordered arbitrarily)

1. Sensitivity to a wide range of the phenomena and constraints in human language
2. Generality across different languages, genres, styles, and modalities
3. Computational efficiency at construction time and runtime
4. Strong formal guarantees (e.g., convergence, statistical efficiency, consistency, etc.)
5. High accuracy when judged against expert annotations and/or task-specific performance

NLP $\stackrel{?}{=}$ 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.
- Early connections to information theory (1940s)
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

NLP $\stackrel{?}{=}$ 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
- Theory of computation
- Data science
- Political science
- Psychology
- Economics
- Education

The Engineering Side

- Application tasks are difficult to define formally; they are always evolving.
- Objective evaluations of performance are always up for debate.
- Different applications require different $\mathcal{R}$.
- People who succeed in NLP for long periods of time are foxes, not hedgehogs.

Today's Applications

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

Factors Changing the NLP Landscape

(Hirschberg and Manning, 2015)

- Increases in computing power
- The rise of the web, then the social web
- Advances in machine learning
- Advances in understanding of language in social context

Administrivia

Course Website

http://courses.cs.washington.edu/courses/cse490u/17wi/

Your Instructors

Noah (instructor):

- UW CSE professor since 2015, teaching NLP since 2006, studying NLP since 1998, first NLP program in 1991
- Research interests: machine learning for structured problems in NLP, NLP for social science

Joshua (TA):

- Linguistics Ph.D. student
- Research interests: computational resources for Lushootseed

Sam (TA):

- Computer Science Ph.D. student
- Research interests: machine learning for natural language semantics

Outline of CSE 490U

1. **Probabilistic language models**, which define probability distributions over text passages. (about 1 week)
2. **Text classifiers**, which infer attributes of a piece of text by "reading" it. (about 1 week)
3. **Sequence models** (about 1.5 weeks)
4. **Parsers** (about 2 weeks)
5. **Semantics** (about 2 weeks)
6. **Machine translation** (about 1 week)
7. Another advanced topic (about 1 week, time permitting)

Readings

- Main reference text: Jurafsky and Martin, 2008, some chapters from new edition (Jurafsky and Martin, forthcoming) when available
- Course notes from others
- Research articles

Lecture slides will include references for deeper reading on some topics.

Evaluation

- Approximately five assignments (A1–5), completed individually (50%).
- Quizzes (15%), given without warning in class, in quiz sections, or online
- An exam (30%), to take place at the end of the quarter
- Participation (5%)

Evaluation

- Approximately five assignments (A1–5), completed individually (50%).
  - Some pencil and paper, mostly programming
  - Graded mostly on attempt, not correctness
- Quizzes (15%), given without warning in class, in quiz sections, or online
- An exam (30%), to take place at the end of the quarter
- Participation (5%)

To-Do List

- Section meetings start next week (January 12), not tomorrow.
- Read: Jurafsky and Martin (2008, ch. 1), Hirschberg and Manning (2015).
- Entrance survey (on Canvas).
- Print, sign, and return the academic integrity statement.

References I

Julia Hirschberg and Christopher D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015. URL `https://www.sciencemag.org/content/349/6245/261.full`.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, second edition, 2008.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, third edition, forthcoming. URL `https://web.stanford.edu/~jurafsky/slp3/`.