Natural Language Processing (CSE 517): Introduction

Noah Smith

University of Washington
nasmith@cs.washington.edu

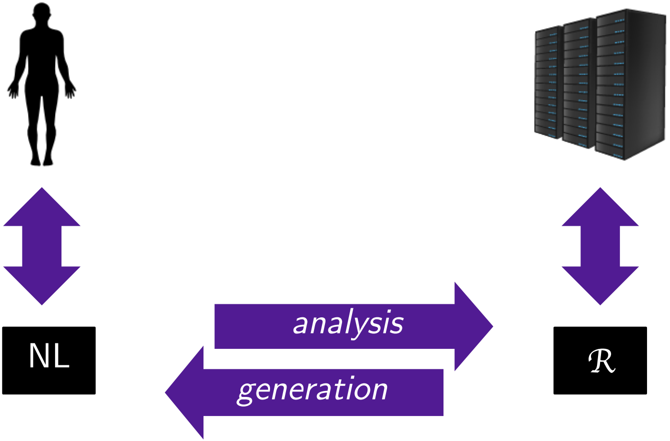
January 4, 2016

What is NLP?

Automation of:

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

What is $\mathcal{R}$?

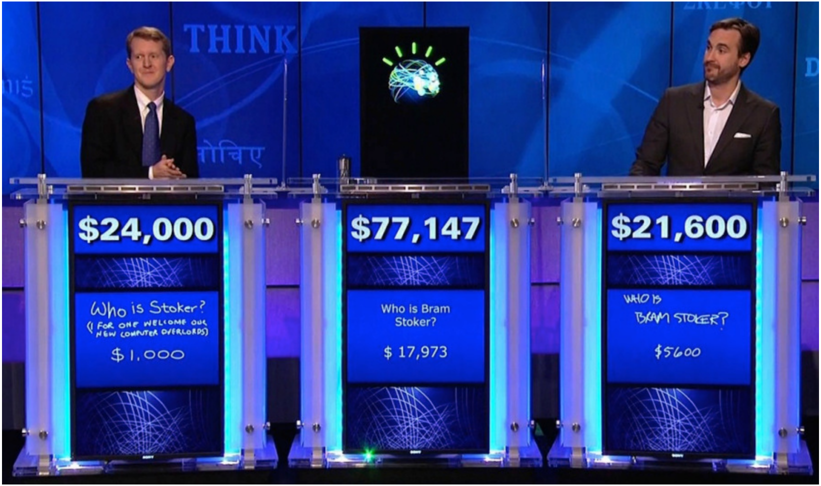


NL

analysis

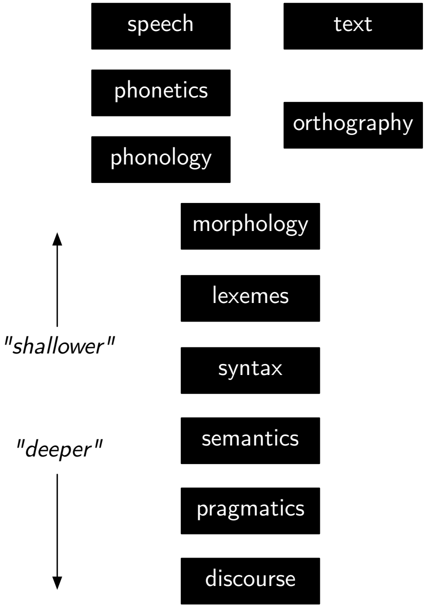
generation

$\mathcal{R}$


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



speech

text

phonetics

orthography

phonology

morphology

lexemes

syntax

semantics

pragmatics

discourse

"shallower"

"deeper"

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*

Example: Part-of-Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol

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	lololol

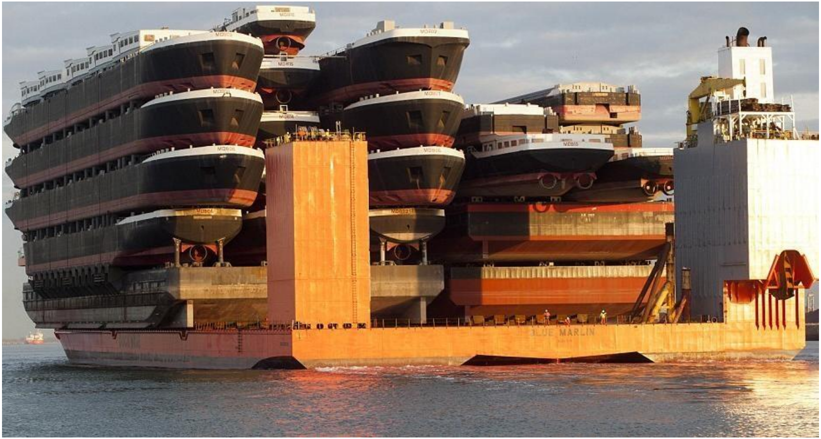
Example: Part-of-Speech Tagging

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	lololol
P	O	V	V	O	P	∧	!
preposition						proper noun	

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 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 linguistic representations
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 $\overset{?}{=}$ Machine Learning

- To be successful, a machine learner needs bias/assumptions, e.g., 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 $\overset{?}{=}$ 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

- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Dialog systems
- Image-to-text
- Scoring exams

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/cse517/16wi/`

Your Instructors

Noah (instructor):

- NLPer since 1998
- Teaching NLP since 2006
- Research interests: machine learning for NLP; translation $\overset{2001}{\rightarrow}$ syntax $\overset{2009}{\rightarrow}$ semantics; NLP for social science

Jesse (TA):

- NLPer since 2011
- Research interests: machine learning for NLP; semantics

Outline of CSE 517

1. **Probabilistic language models**, which define probability distributions over text passages. (5 lectures)
2. **Text classifiers**, which infer attributes of a piece of text by "reading" it. (2)
3. **Analyzers**, which map texts into **linguistic representations** that in turn enable various kinds of understanding. (8)
4. **Generators**, which produce natural language as output. (3)

Outline of CSE 517

1. **Probabilistic language models**, which define probability distributions over text passages. (5 lectures)
   $\mathcal{V}^* \rightarrow \mathcal{V}$
2. **Text classifiers**, which infer attributes of a piece of text by "reading" it. (2)
   $\mathcal{V}^* \rightarrow \Lambda$
3. **Analyzers**, which map texts into **linguistic representations** that in turn enable various kinds of understanding. (8)
   $\mathcal{V}^* \rightarrow \mathcal{Y}$
4. **Generators**, which produce natural language as output. (3)
   $\mathcal{X} \rightarrow \mathcal{V}^*$

Readings

- Course notes from others
- Chapters from books (mostly Jurafsky and Martin, forthcoming)
- Research articles

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

Evaluation

- Approximately four assignments, completed individually (40%)
- A project, completed in teams of 1–3 (35%)
- An oral exam (15%)
- Participation (10%)

Evaluation

- Approximately four assignments, completed individually (40%)
  - Some pencil and paper, some programming
  - Graded mostly on attempt, not correctness
  - Solutions will be made available where appropriate
- A project, completed in teams of 1–3 (35%)
- An oral exam (15%)
- Participation (10%)

Evaluation

- Approximately four assignments, completed individually (40%)
- A project, completed in teams of 1–3 (35%)
  - Probabilistic language model
  - Simple command-line interface
  - Teams of 1–3
  - Due March 9
  - Details in syllabus
- An oral exam (15%)
- Participation (10%)

Evaluation

- Approximately four assignments, completed individually (40%)
- A project, completed in teams of 1–3 (35%)
- An oral exam (15%)
  - Toward the end of the quarter
  - Do not panic
- Participation (10%)

Evaluation

- Approximately four assignments, completed individually (40%)
- A project, completed in teams of 1–3 (35%)
- An oral exam (15%)
- Participation (10%)
  - Proposed oral exam questions

Readings

Hirschberg and Manning (2015)

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, third edition, forthcoming. URL `https://web.stanford.edu/~jurafsky/slp3/`.