Natural Language Processing (CSE 447/547M): Introduction

Noah Smith

University of Washington
nasmith@cs.washington.edu

January 7, 2019

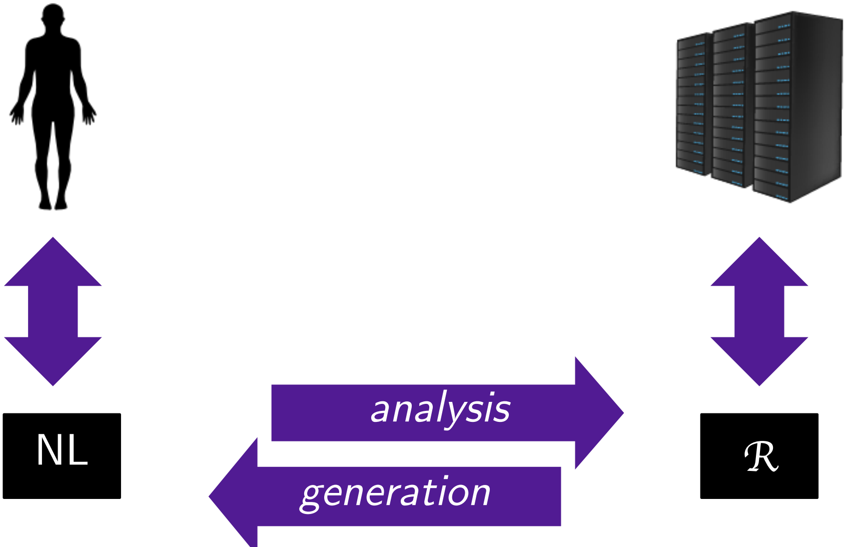
What is NLP?

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

Automation of:

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

What is $\mathcal{R}$?



NL

$\mathcal{R}$

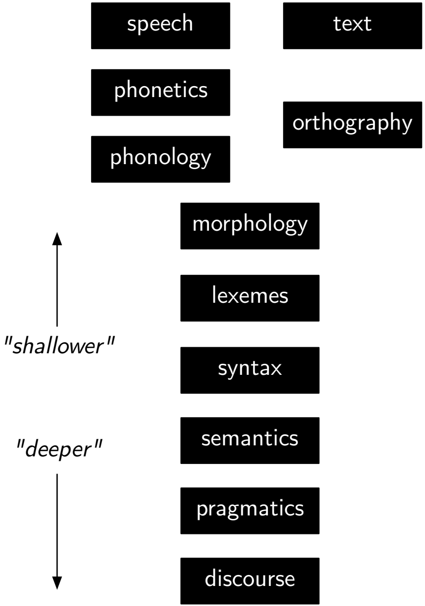
analysis

generation

THINK
ΣΚΕΨΟΥ
PIENSE
TENK
DENKE
DINK
DENK
THINK
ДУМАЙ
考えろ
सोचिए
$24,000
$77,147
$21,600
WATSON

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-YELED BA-GAN
"you will meet the boy in the park"

finsta, demonetize, chillax, unfriend, Frankenfood, Obamacare, Manfuckinghattan, screenager, Twitterati, girther

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 lololol

Example: Part-of-Speech Tagging

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

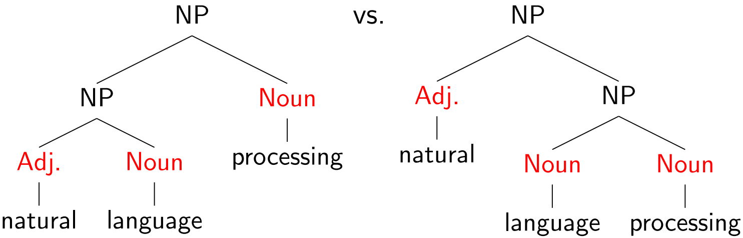
				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol

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

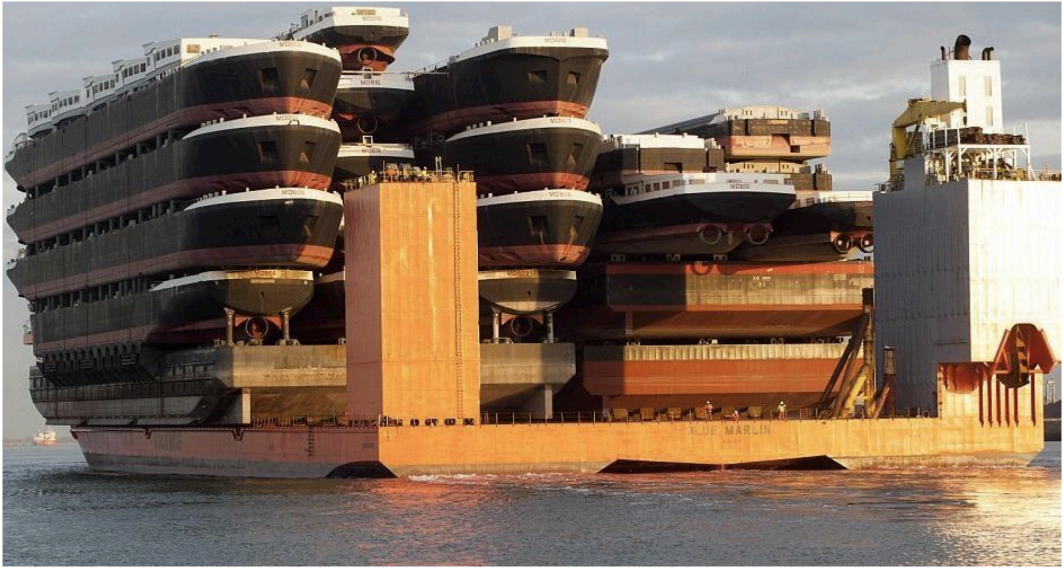
				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol
P	O	V	V	O	P	∧	!
preposition						proper noun	

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.



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

Pragmatics

Noah likes some children

If the speaker meant that Noah likes *all* children, they would have said that.

So we are likely to infer that Noah also *doesn't* like some children.

Discourse

Allen purchased the Portland Trail Blazers NBA team in 1988 from California real estate developer Larry Weinberg for $70 million. He was instrumental in the development and funding of the Moda Center, the arena where they play.

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
6. Explainable to human users (added in 2019)

NLP $\overset{?}{=}$ Machine Learning

- Many NLP problems are reduced to ML problems, and this works better than anything that came before.
- However, $\mathcal{R}$ 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

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

How I Teach NLP

There's quite a lot to cover!

I've selected building blocks that give you a sense of the challenges and problems in the field, so you can learn and do more on your own.

I will often take a few steps in some direction and then tell you where you can find out more. It's up to you!

This year, we're making an effort to update the assignments so you get to work with the latest tools.

Administrivia

Course Website

`http://courses.cs.washington.edu/courses/cse447/19wi/`

There's a link on the website to a spreadsheet showing the course plan, readings, deadlines, etc.

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
- Second gig: research manager for AllenNLP, an open-source NLP research library, built on PyTorch, at AI2

TAs: Elizabeth Clark, Lucy Lin, Nelson Liu, Deric Pang, Kaidi Pei

Outline of CSE 447/547M

1. **Probabilistic language models**, which define probability distributions over text passages. (about 1.5 weeks)
2. **Text classifiers**, which infer attributes of a piece of text by “reading” it. (about 1 week)
3. **Words in context** (about 1.5 weeks)
4. **Sequence models** (about 1.5 weeks)
5. **Syntax** (about 1.5 weeks)
6. **Machine translation** (about 0.5 week)
7. **Semantics** (about 2 weeks)

Readings

- Main reference text: Eisenstein (2018) Download it now!
- Useful reference on neural nets for NLP: Goldberg (2017) Download it now!
- Course notes from the instructor and others
- Research articles

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

Evaluation

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

Evaluation

- Five assignments (A1–5), completed individually (50%).
  - Some pencil and paper, mostly programming
  - Graded mostly on your writeup (so please take written communication seriously!)
  - Effort matters more than correctness
  - Late day policy: 3 late days
- Final exam (30%), to take place at the end of the quarter
- Quizzes (15%), given without warning in class or in quiz sections
- Participation (5%)

Am I Ready for CSE 447?

- The course is designed for CSE majors.
  - There will be programming
  - There will be math (e.g., conditional probability, gradient descent, the chain rule from calculus)
  - There will be linguistics (ideas from syntax, lexical semantics, frame semantics, and compositional semantics)
- We are here to help, but if you need extreme amounts of help, we'll advise you drop the course.
- It's your call!

To-Do List

- Download the book: Eisenstein (2018)
- Print, sign, and upload through Canvas the academic integrity statement on the course web page, `http://courses.cs.washington.edu/courses/cse517/18sp/academic-integrity.pdf`

References I

Jacob Eisenstein. *Natural Language Processing*. 2018. URL https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf.

Yoav Goldberg. *Neural Network Methods for Natural Language Processing*. Morgan Claypool, 2017. URL https://www.morganclaypool.com/doi/abs/10.2200/S00762ED1V01Y201703HLT037.

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.