Natural Language Processing (CSE 517 & 447): Introduction

Noah Smith

University of Washington
nasmith@cs.washington.edu

January 3, 2022

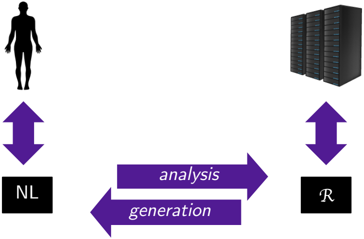
Readings: Eisenstein (2019) chapter 1, lecture notes

What is NLP?

NL $\in$ {Mandarin Chinese, English, American Sign Language, . . ., Lushootseed}

Automation of:

- analysis of ("understanding") what a text means, to some extent
- generation of fluent, meaningful, context-appropriate text



What applications is NLP for?

`https://courses.cs.washington.edu/courses/cse447/22wi/`

→ "Live poll" link on the left

Why do we have a whole class on NLP?

It's really hard.

Decomposition into meaningful parts can be non-trivial.

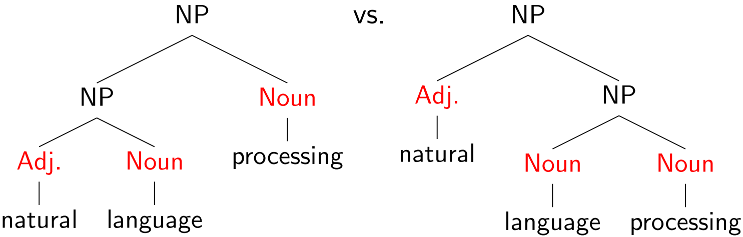
ลูกศิษย์วัดกระทิงยังยื้อปิดถนนทางขึ้นไปนมัสการพระบาทเขาคิชฌกูฏ หวิดปะทะ
กับเจ้าถิ่นที่ออกมาเผชิญหน้าเพราะเดือดร้อนสัญจรไม่ได้ ผวจ.เร่งทุกฝ่ายเจรจา
ก่อนที่ชื่อเสียงของจังหวัดจะเสียหายไปมากกว่านี้ พร้อมเสนอหยุดจัดงาน 15 วัน....

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”

unfriend, Obamacare, Manfuckinghattan, covid-19

Ambiguity is everywhere.



Ambiguity is everywhere.

We saw the woman with the telescope wrapped in paper.

Ambiguity is everywhere.

We saw the woman with the telescope wrapped in paper.

- Who has the telescope?

Ambiguity is everywhere.

We saw the woman with the telescope wrapped in paper.

- Who has the telescope?
- Who or what is wrapped in paper?

Ambiguity is everywhere.

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?

Ambiguity is everywhere.

Every fifteen minutes a woman in this country gives birth.

– Groucho Marx

Ambiguity is everywhere.

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

– Groucho Marx

Desiderata for NLP

(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

Don't expect a silver bullet to "solve language."

Intellectual Connections

- machine learning
- linguistics
- artificial intelligence
- ethics

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

In 2022, I would add:

- Consumer and investor demand
- Emerging ethical questions around deployment

Noah's Approach to Teaching NLP

- Application tasks are difficult to define formally and are always evolving, so I focus on useful **abstractions** (with examples).
- Objective evaluations of performance are always up for debate, so I discuss shortcomings and emphasize **tradeoffs**.
- Different applications require different $\mathcal{R}$, so I encourage **openmindedness**.
- People who succeed at NLP for long periods of time understand many tools and how they relate to each other.
- This class doesn't teach you how to solve any one particular problem; it gives you tools that will help you understand and tackle new problems.

Administrivia

Course Website

`https://courses.cs.washington.edu/courses/cse447/22wi/`

Links to **all the things**.

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

TAs: Gabriel, Ivan, Jerome, Kaiser, Leroy, Suchin, Tobi, Velocity, Xiujun

Learn more about all of us by reading our self-introductions on the Ed discussion board!

Outline of Topics

1. Classification and multinomial logistic regression
2. Language modeling, especially with neural networks
3. Vector embeddings for documents and words
4. Morphology and weighted finite-state transducers
5. Sequence labeling and conditional random fields
6. Syntax, semantics, and linguistic structure prediction
7. Translation and sequence-to-sequence models

Readings

- Main reference text: Eisenstein (2019)
- Occasional course notes from the instructor and others
- Research articles

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

Evaluation

- Nine assignments (A1–A9), roughly one per week, completed individually (50%).
- Project, in a team of three (40%); 447 and 517 projects are quite different.
- Quizzes on Canvas (10%). Not graded for correctness.

Zero credit for late work; for assignments, if you're within one week of the deadline, you will get feedback.

Evaluation

- Nine assignments (A1–A9), roughly one per week, completed individually (50%).
  - Some pencil and paper, some programming.
  - Graded mostly on your writeup (so please take written communication seriously).
  - 517 version will typically be heavier.
- Project, in a team of three (40%); 447 and 517 projects are quite different.
- Quizzes on Canvas (10%). Not graded for correctness.

Zero credit for late work; for assignments, if you're within one week of the deadline, you will get feedback.

About Assignments

We assume that:

- Most students will not complete every assignment. (We count your best two double and zero out your worst two.)
- You will benefit from an opportunity to develop the skill of doing your best in a finite amount of time.
- If you get stuck, you will take advantage of the many resources about NLP that are available online—articles, tutorials, papers, lectures by other professors.

About Lectures

Last year all lectures were prerecorded (captions and transcripts also available).

- You are welcome to use these however you like.
- If you feel sick, definitely use these instead of coming to class!
- The prerecorded videos are deceptively short. They don't have any interruptions or discussion, which is usually about a third of lecture time! Most people will need to slow them down, take breaks, re-watch some parts, etc.

Am I Ready for This Course?

- The course is designed for CSE students.
  - There will be programming, on your own and in small groups.
  - There will be math (e.g., conditional probability, gradient descent, the chain rule from calculus, linear algebra). **Use A0 to gauge how much extra work you'll need to put in.**
- In the past, about a third of 517 students came from outside CSE; they worked *hard*!
- We are here to help, but if you need extreme amounts of help, we'll advise you drop the course.
- It's your call!

COVID-19 Safety

Please do not expose others if you feel sick or have been exposed.

The main lectures have prerecorded videos that suffice (that's all we used in 2021).

There will be regular remote office hours.

More on the website.

To-Do List

- Get the book (Eisenstein, 2019) and start reading chapter 1.
- Print, read, sign, and submit (through Canvas) the academic integrity statement linked on the course web page.
- Complete A0 and use the provided solutions to gauge how much extra work you might need to put in on math.
- Use the Ed discussion board to introduce yourself and find teammates.
- If you do not have a CSE account, request one (instructions on course web page).
- Make sure you can log in to course machines (instructions on course web page).

References I

Jacob Eisenstein. *Introduction to Natural Language Processing*. MIT Press, 2019.

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`.