Natural Language Processing

Introduction to NLP

Yulia Tsvetkov

yuliats@cs.washington.edu
Announcements

https://courses.cs.washington.edu/courses/cse447/22au/

- **Quiz 1: Monday Oct 10**
  - 5 questions, open during lecture time, 10-min in the end of the class
  - Materials from weeks 1 and 2
    - Introduction to NLP, introduction to text classification, NB
    - Instructions for HW 1

- **Discussions on Ed**
  - Reminder about 10% bonus grade for commenting on Ed
  - TAs will respond to your questions within 24 hours

- **Office hours**
  - TA OH locations have been updated on the website
  - Yulia is traveling in the rest of this week, no OHs on Friday
Personal assistants

- Amazon Alexa
- Siri
- Google Now
- Facebook M
- Cortana
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?
Machine translation

I study deep learning and machine learning.

我学习深度学习和机器学习

Wǒ xuéxiāo shēndé xuéxiāo hé jīqì xuéxiāo
Sentiment analysis

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner
$89 online, $100 nearby ★★★★★ 377 reviews
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 she

Reviews

Summary - Based on 377 reviews

<table>
<thead>
<tr>
<th>1 star</th>
<th>2</th>
<th>3</th>
<th>4 stars</th>
<th>5 stars</th>
</tr>
</thead>
</table>

What people are saying

ease of use | "This was very easy to setup to four computers."
value | "Appreciate good quality at a fair price."
setup | "Overall pretty easy setup."
customer service | "I DO like honest tech support people."
size | "Pretty Paper weight."
mode | "Photos were fair on the high quality mode."
colors | "Full color prints came out with great quality."
Information extraction

In 1933, while Einstein was visiting the United States, Adolf Hitler came to power.

Because of his Jewish background, Einstein did not return to Germany.

He settled in the United States and became an American citizen in 1940.

Einstein supported the Allied forces, but he generally denounced the idea of using nuclear fission as a weapon.

He signed the Russell-Einstein Manifesto with British philosopher Bertrand Russell, which highlighted the danger of nuclear weapons.

He was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.
Type in a word and we'll highlight the good and the bad

"united airlines"

Sentiment analysis for "united airlines"

Sentiment by Percent

- Negative (66%)
- Positive (32%)

Sentiment by Count

Positive (11)
Negative (23)

- jliacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human. Posted 2 hours ago
- 12345clumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this damn mess! ? Posted 2 hours ago
- EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination. http://t.co/Z9QloAjF Posted 2 hours ago
- CountAdams: FANTASTIC customer service from United Airlines at XNA today. Is tweet more, but cell phones off now! Posted 4 hours ago
Information extraction for disaster relief

- Translation + information extraction

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

"Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital."

An earthquake struck Haiti on January 12, 2010

Most local services failed, but most cell-towers remained functional.
Hate speech detection
Covid19 misinformation

Detecting COVID-19-Related Fake News Using Feature Extraction
Suleman Khan, Saqib Hakak, N. Deepa, B. Prabadevi, Kapal Dev and Silvia Trelova

Computational social science

- computational social science answering questions about society given observational data
- example: "do movie scripts portray female or male characters with more power or agency?" [Sap+ 2017]
Language change

Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change

William L. Hamilton, J. Leskovec, Dan Jurafsky
Natural Language Processing

- **Applications**
  - Machine Translation
  - Information Retrieval
  - Question Answering
  - Dialogue Systems
  - Information Extraction
  - Summarization
  - Sentiment Analysis
  - ...

- **Core technologies**
  - Language modelling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - Coreference resolution
  - Word sense disambiguation
  - Semantic Role Labelling
  - ...

- **Core technologies**
  - Language modelling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - Coreference resolution
  - Word sense disambiguation
  - Semantic Role Labelling
  - ...
Symbolic and Probabilistic NLP

Logic-based/Rule-based NLP

- Interlingua
- Transfer
- Direct translation

Source text → Analysis → Interlingua → Transfer → Direct translation → Target text

Statistical NLP

- Translation Model
  - Source phrase
  - Target phrase
  - Translation features

- Language Model

- Reranking Model
  - Feature weights
  \[ \text{argmax}_e P(f|e)P(e) \]
Probabilistic and Connectionist NLP

Engineered Features/Representations

Translation Model

Language Model

Reranking Model

Learned Features/Representations

~mid 2010s
Where are we now?

mostly solved

Spam detection
Let's go to Agra!
Buy VIAGRA ...

Part-of-speech (POS) tagging
ADJ ADJ NOUN VERB ADV
Colorless green ideas sleep furiously.

Named entity recognition (NER)
PERSON ORG LOC
Einstein met with UN officials in Princeton

making good progress

Sentiment analysis
Best roast chicken in San Francisco!
The waiter ignored us for 20 minutes.

Coreference resolution
Center told Mubarak he shouldn't run again.

Word sense disambiguation
I need new batteries for my mouse.

Parsing
I can see Alcatraz from the window!

Machine translation (MT)
第13届上海国际电影节开幕...
The 13th Shanghai International Film Festival...

Information extraction (IE)
You're invited to our dinner party, Friday May 27 at 8:30 PM

still really hard

Question answering (QA)
Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
XYZ acquired ABC yesterday
ABC has been taken over by XYZ

Summarization
The Dow Jones is up
The S&P500 jumped
Housing prices rose

Dialog
Where is Citizen Kane playing in SF?
Castro Theatre at 7:30. Do you want a ticket?
Chat bots

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html
Where are we now? GPT-3 Turing Test

This gives us a hint for how to stump the AI more consistently. We need to ask questions that no normal human would ever talk about.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many eyes does my foot have?
A: Your foot has two eyes.

Q: How many eyes does a spider have?
A: A spider has eight eyes.

Q: How many eyes does the sun have?
A: The sun has one eye.

Q: How many eyes does a blade of grass have?
A: A blade of grass has one eye.

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html
GPT-3 Turing Test

Now we're getting into surreal territory. GPT-3 knows how to have a normal conversation. It doesn't quite know how to say “Wait a moment... your question is nonsense.” It also doesn't know how to say “I don't know.”

Q: How do you sporgle a morgle?
A: You sporgle a morgle by using a sporgle.

Q: How many bonks are in a quoit?
A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?
A: It takes two rainbows to jump from Hawaii to seventeen.

Q: Which colorless green ideas sleep furiously?
A: Ideas that are colorless, green, and sleep furiously are the ideas of a sleep furiously.

Q: Do you understand these questions?
A: I understand these questions.

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html
One trend that continues from the common sense is that GPT-3 is reluctant to express that it doesn’t know the answer. So invalid questions get wrong answers.

Q: Who was president of the United States in 1700?
A: William Penn was president of the United States in 1700.

Q: Who was president of the United States in 1650?
A: Richard Bennett was president of the United States in 1650.

Q: Who was president of the United States in 1620?
A: James I was president of the United States in 1620.

Q: Who was president of the United States in 1600?
A: Queen Elizabeth I was president of the United States in 1600.

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html
An interesting corollary is that GPT-3 often finds it easier to write code to solve a programming problem, than to solve the problem on one example input:

Q: Write one line of Ruby code to reverse an array.
A: ary.reverse

Q: Write one line of Ruby code to count how many X's are in an array.
A: ary.count { |x| x == 'X' }

Q: Reverse the following array: [1, 3, 5, 6, 10, 4, 2, 77]
A: [10, 6, 4, 2, 77, 3, 5, 1]

Q: How many X's are in this array: [1, 3, 5, 6, 10, 4, 2, 77]
A: 2
Where are we now? Meena → LaMDA

Conversation A

<table>
<thead>
<tr>
<th>Human:</th>
<th>say you’re in an empty giant office and you see a black bird flying indoors. what do you do?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human:</td>
<td>well it’s a predator bird and it’s gonna eat you I would grab my phone and call 911.</td>
</tr>
</tbody>
</table>

Cross-turn Repetition Example 1

<table>
<thead>
<tr>
<th>Human:</th>
<th>Hi!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meena:</td>
<td>Hi! How are you?! :)</td>
</tr>
<tr>
<td>Human:</td>
<td>doing good, what would be the best city to visit in Europe?</td>
</tr>
<tr>
<td>Meena:</td>
<td>Paris, Barcelona, Amsterdam, Prague</td>
</tr>
<tr>
<td>Human:</td>
<td>great list! why?</td>
</tr>
<tr>
<td>Meena:</td>
<td>Paris is a beautiful city, and Barcelona is too. And I’ve always wanted to go to Amsterdam and Prague and have never been.</td>
</tr>
<tr>
<td>Human:</td>
<td>have you even been to India?</td>
</tr>
<tr>
<td>Meena:</td>
<td>No, but I’d love to go.</td>
</tr>
<tr>
<td>Human:</td>
<td>how about Japan?</td>
</tr>
<tr>
<td>Meena:</td>
<td>I’d love to go to Japan too.</td>
</tr>
<tr>
<td>Human:</td>
<td>how about Paris?</td>
</tr>
<tr>
<td>Meena:</td>
<td>I’d love to go to Paris too.</td>
</tr>
<tr>
<td>Human:</td>
<td>are you lying to me?</td>
</tr>
<tr>
<td>Meena:</td>
<td>I’d love to go to Japan too.</td>
</tr>
<tr>
<td>Human:</td>
<td>i think you’re lying to me, that makes me sad</td>
</tr>
<tr>
<td>Meena:</td>
<td>I’d love to go to Japan, too.</td>
</tr>
<tr>
<td>Human:</td>
<td>you’re blocking me out! you have to talk to me</td>
</tr>
<tr>
<td>Meena:</td>
<td>I’d love to go to Japan.</td>
</tr>
<tr>
<td>Human:</td>
<td>where else?</td>
</tr>
<tr>
<td>Meena:</td>
<td>I’d love to go to Japan, too!</td>
</tr>
</tbody>
</table>

Towards a Human-like Open-Domain Chatbot Adiwardana et al. arXiv preprint 2020
AI chatbot is REMOVED from Facebook after saying she 'despised' gay people, would 'rather die' than be disabled and calling the #MeToo movement 'ignorant'

- Lee Luda is a South Korean chatbot with the persona of a 20-year-old student
- It has attracted more than 750,000 users since its launch last month
- But the chatbot has started using hate speech towards minorities
- In one of the captured chat shots, Luda said she ‘despised’ gays and lesbians
- The developer has apologised over the remarks, saying they ‘do not represent our values as a company’
Bias in machine translation

What can we do about this problem? We’ll discuss in NLP class!
Syllabus

https://courses.cs.washington.edu/courses/cse447/22au/

- **Introduction**
  - Overview of NLP as a field

- **Modeling (ML fundamentals)**
  - Text classification: linear models (perceptron, logistic regression), non-linear models (FF NNs, CNNs)
  - Language modeling: n-gram LMs, neural LMs, RNNs
  - Representation learning: word vectors, contextualized word embeddings, Transformers

- **Linguistic structure and analysis (Algorithms, linguistic fundamentals)**
  - Words, morphological analysis,
  - Sequences: part of speech tagging (POS), named entity recognition (NER)
  - Syntactic parsing (phrase structure, dependencies)

- **Applications (Practical end-user solutions, research)**
  - Sentiment analysis, toxicity detection
  - Machine translation, summarization
  - Computational social science
  - Interpretability
  - Fairness and bias
Learning goals

At the end of this course, you will be able to:

- Build a supervised classifier to solve problems like sentiment classification
- Build a neural network and train it using stochastic gradient descent
- Build tools for extracting linguistic knowledge from raw text, including names, and sentence structure
- Learn ML fundamentals for text processings (including state-of-the-art methods)
- Learn important algorithms for text processings (that are useful also in other fields)
- Learn methodological tools (training/test sets, cross-validation)

- It's gentle (my goal is to explain everything) and broad (covering many many topics)
- Mastery independent learning, quizzes and programming homeworks
- No research project, but fun research-oriented lectures towards the end of the course
Linguistic Background
What does it mean to “know” a language?
What does an NLP system need to ‘know’?

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!
Levels of linguistic knowledge
Speech, phonetics, phonology

This is a simple sentence.
/ðɪz ɪz ə 'simpəl 'sɛntəs/.

"shallower"

"deeper"

speech
phonetics
phonology
orthography
morphology
lexemes
syntax
semantics
pragmatics
discourse
Orthography

هذه جملة بسيطة

đây là một câu đơn giản

यह एक साधारण वाक्य है

This is a simple sentence.

/ ɔɪs ɪz ə 'sɪmpl 'sɛntəs /. 

speech  
phonetics  
phonology  

speech

morphology

lexemes

syntax

semantics

pragmatics

discourse

"shallower"

"deeper"
Words, morphology

- Morphological analysis
- Tokenization
- Lemmatization

Tokens: This is a simple sentence.

Morphology: be 3sg present

Diagram:
- "shallow": speech, phonetics, phonology
- "deeper": morphology, lexemes, syntax, semantics, pragmatics, discourse
Syntax

- Part-of-speech tagging

**Parts of speech**

- DT
- VBZ
- DT
- JJ
- NN
- PUNC

**Tokens**

This is a simple sentence.

be
3sg
present

**Morphology**

speech
phonetics
phonology
orthography
morphology
lexemes
syntax
semantics
pragmatics
discourse

"shallower"

"deeper"
Syntax

- Part-of-speech tagging
- Syntactic parsing

Syntax

Parts of speech

Tokens

Morphology

This is a simple sentence.
Semantics

- Named entity recognition
- Word sense disambiguation
- Semantic role labelling

This is a simple sentence.

Morphology

Syntax

Parts of speech

Tokens

Semantics

be 3sg present

SIMPLE1: having few parts

SENTENCE1: String of words satisfying the grammatical rules of a language

speech    text
phonetics  orthography
phonology

morphology    lexemes
syntax        semantics
pragmatics    discourse
Discourse

- Reference resolution
- Discourse parsing

This is a simple sentence.

But an instructive one.
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$
Ambiguity: word sense disambiguation
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 saw her duck*
Every language sees the world in a different way

- For example, it could depend on cultural or historical conditions

- Russian has very few words for colors, Japanese has hundreds

- Multiword expressions, e.g. happy as a clam, it’s raining cats and dogs or wake up and metaphors, e.g. love is a journey are very different across languages
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$
Africa is a continent with a very high linguistic diversity: there are an estimated 1.5-2K African languages from 6 language families. 1.33 billion people
NLP beyond English

- ~7,000 languages
- thousands of language varieties
Most of the world today is multilingual

Source: Ethnologue

The Countries With The Most Spoken Languages
Number of living languages spoken per country in 2015

Source: Ethnologue

Source: US Census Bureau

Yulia Tsvetkov

Undergrad NLP 2022
Tokenization

This is a simple sentence

これは簡単なフレーズです
Tokenization + disambiguation

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them
- tokenization is highly ambiguous
Tokenization + morphological analysis

- Quechua

Much’ananayakapushasqakupuniñataqsunamá

Much’a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

“So they really always have been kissing each other then”

<table>
<thead>
<tr>
<th>Much’a</th>
<th>to kiss</th>
</tr>
</thead>
<tbody>
<tr>
<td>-na</td>
<td>expresses obligation, lost in translation</td>
</tr>
<tr>
<td>-naya</td>
<td>expresses desire</td>
</tr>
<tr>
<td>-ka</td>
<td>diminutive</td>
</tr>
<tr>
<td>-pu</td>
<td>reflexive (kiss <em>eachother</em>)</td>
</tr>
<tr>
<td>-sha</td>
<td>progressive (kiss<em>ing</em>)</td>
</tr>
<tr>
<td>-sqa</td>
<td>declaring something the speaker has not personally witnessed</td>
</tr>
<tr>
<td>-ku</td>
<td>3rd person plural (they kiss)</td>
</tr>
<tr>
<td>-puni</td>
<td>definitive (really*)</td>
</tr>
<tr>
<td>-ña</td>
<td>always</td>
</tr>
<tr>
<td>-taq</td>
<td>statement of contrast (...then)</td>
</tr>
<tr>
<td>-suna</td>
<td>expressing uncertainty (So...)</td>
</tr>
<tr>
<td>-má</td>
<td>expressing that the speaker is surprised</td>
</tr>
</tbody>
</table>
Tokenization + morphological analysis

- German

Infektionsschutzmaßnahmenverordnung
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $R$
Linguistic variation

- Non-standard language, emojis, hashtags, names

chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂💕 Note the cute little crab indicator in the 2nd pic 🦀💖
Variation

- Suppose we train a part of speech tagger or a parser on the Wall Street Journal

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

@_rkpntrnte hindi ko alam babe eh, absent ako
kagina I’m sick rn hahaha😊🙌
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$
Sparse data due to Zipf’s Law

- To illustrate, let’s look at the frequencies of different words in a large text corpus
- Assume “word” is a string of letters separated by spaces
## Word Counts

Most frequent words in the English Europarl corpus (out of 24m word tokens)

<table>
<thead>
<tr>
<th>any word</th>
<th>Frequency</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,698,599</td>
<td>the</td>
<td></td>
</tr>
<tr>
<td>849,256</td>
<td>of</td>
<td></td>
</tr>
<tr>
<td>793,731</td>
<td>to</td>
<td></td>
</tr>
<tr>
<td>640,257</td>
<td>and</td>
<td></td>
</tr>
<tr>
<td>508,560</td>
<td>in</td>
<td></td>
</tr>
<tr>
<td>407,638</td>
<td>that</td>
<td></td>
</tr>
<tr>
<td>400,467</td>
<td>is</td>
<td></td>
</tr>
<tr>
<td>394,778</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>263,040</td>
<td>I</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nouns</th>
<th>Frequency</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>124,598</td>
<td>European</td>
<td></td>
</tr>
<tr>
<td>104,325</td>
<td>Mr</td>
<td></td>
</tr>
<tr>
<td>92,195</td>
<td>Commission</td>
<td></td>
</tr>
<tr>
<td>66,781</td>
<td>President</td>
<td></td>
</tr>
<tr>
<td>62,867</td>
<td>Parliament</td>
<td></td>
</tr>
<tr>
<td>57,804</td>
<td>Union</td>
<td></td>
</tr>
<tr>
<td>53,683</td>
<td>report</td>
<td></td>
</tr>
<tr>
<td>53,547</td>
<td>Council</td>
<td></td>
</tr>
<tr>
<td>45,842</td>
<td>States</td>
<td></td>
</tr>
</tbody>
</table>
Word Counts

But also, out of 93,638 distinct words (word types), 36,231 occur only once.

Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoﬁt, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a
Plotting word frequencies

Order words by frequency. What is the frequency of nth ranked word?
Zipf’s Law

Implications

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$
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
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$
Unmodeled variables

World knowledge

- I dropped the glass on the floor and it broke
- I dropped the hammer on the glass and it broke

“Drink this milk”
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$
Unknown representation

- Very difficult to capture what is $\mathcal{R}$, since we don’t even know how to represent the knowledge a human has/needs:
  - What is the “meaning” of a word or sentence?
  - How to model context?
  - Other general knowledge?
Dealing with ambiguity

- How can we model ambiguity and choose the correct analysis in context?
  - non-probabilistic methods (FSMs for morphology, CKY parsers for syntax) return all possible analyses.
  - probabilistic models (HMMs for part-of-speech tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return the best possible analysis, i.e., the most probable one according to the model
  - Neural networks, pretrained language models now provide end-to-end solutions

- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?
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 aligned French / English sentences
  - Yelp reviews
  - The Web: billions of words of who knows what
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
- Ethical
NLP ≻ Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.
What is nearby NLP?

- **Computational Linguistics**
  - Using computational methods to learn more about how language works
  - We end up doing this and using it

- **Cognitive Science**
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!

- **Speech Processing**
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP
Next class

- Classification

Questions?