Algorithms for NLP
CS 11-711 · Fall 2020

Lecture 1: Introduction

Emma Strubell
Welcome!

Emma  Yulia  Bob  Sanket  Han  Jiateng
Course website:
http://demo.clab.cs.cmu.edu/11711fa20/

Algorithms for NLP

CMU CS 11711, Fall 2020

Tuesday/Thursday 1:30-2:50pm EDT, on Zoom

Emma Strubell (office hours: By appointment), strubell@cmu.edu
Yulia Tsvetkov (office hours: By appointment), ytsvetko@cs.cmu.edu
Robert Frederking (office hours: By appointment), ref@cs.cmu.edu

Teaching Assistants:

Jiateng Xie (office hours: Wednesday 10:00-11:00pm EDT, Zoom), jiatengx@cs.cmu.edu
Sanket Vaibhav Mehta (office hours: Friday 9:00-10:00am EDT, Zoom), svmeha@cs.cmu.edu
Xiaochuang Han (office hours: Monday 10:30-11:30am EDT, Zoom), xiaochuh@cs.cmu.edu

Forum: Piazza

Note: Sensitive information related to the class (e.g., Zoom links) will be available on Piazza

Piazza:
https://piazza.com/CMU/fall2020/11711/
Communicating with machines

~1950s-1970s:
Communicating with machines

~1980s:

```plaintext
say 'Welcome Coffee drinker.'
```

```plaintext
say "What is the price of your coffee?",
"(e.g. 1.50 = $1.50)"
```

```plaintext
say "How many coffees a week do you have?"
```

```plaintext
say "What annual interest rate would you like to see on that money?,
"(e.g. 8 = 8\%"
```

```plaintext
Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
```
Communicating with machines

Today:

“Play a good song”

Sorry, I couldn’t find ‘a good song’ in your music.

“I need a dinner reservation for Valentine’s Day.”

I’ll see if any restaurants have a table for one.

“No, I need a reservation for two.”

Why? Is your mother in town?
What is natural language processing (NLP)?

- NL ∈ {Mandarin, Hindi, Spanish, Arabic, English, … Inuktitut}

- Automation of NLs:
  - analysis (NL → R)
  - generation (R → NL)
  - acquisition of R from knowledge and data
What is natural language processing (NLP)?
Language technologies
What technologies are required to write such a program?

- A conversational agent contains:
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Information retrieval
  - Text to speech
Language technologies
Machine translation

I study deep learning and machine learning.
Language technologies
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 at the end of the 18th century?
■ What do scientists think about the ethics of human cloning?
## Natural language processing

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

### Core technologies
- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic role labelling
- ...

What does an NLP system need to “know”?

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing and understanding language
- Ideally, so would a computer!
What does it mean to “know” a language?
Levels of linguistic knowledge

"shallower"

- speech
- phonetics
- phonology

"deeper"

- morphology
- lexemes
- syntax
- semantics
- pragmatics
- discourse

Image credit: Noah Smith
Speech, phonetics, phonology

This is a simple sentence.

/ðɪs ɪz ə 'sɪmpl 'sɛntɛns/.
**Words**

- Morphological analysis
- Tokenization
- Lemmatization

**Tokens**

This is a simple sentence.

**Morphology**

be
3sg
present
Syntax

- Part-of-speech tagging

<table>
<thead>
<tr>
<th>Parts of speech</th>
<th>DT</th>
<th>VBZ</th>
<th>DT</th>
<th>JJ</th>
<th>NN</th>
<th>PUNC</th>
</tr>
</thead>
</table>

Tokens

This is a simple sentence.

Morphology

be 3sg present

"shallower"

"deeper"
Syntax

- Part-of-speech tagging
- Syntactic parsing

This is a simple sentence .

Parts of speech

Tokens

Morphology

be 3sg present

"shallow"
This is a simple sentence.
This is a simple sentence.

But an instructive one.
Where are we now?

OpenAI's new language generator GPT-3 is shockingly good — and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

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: Which is heavier, a mouse or a toaster?
A: A toaster is heavier than a mouse.

Q: Which is heavier, a toaster or a pencil?
A: A pencil is heavier than a toaster.
**Where are we now?**

<table>
<thead>
<tr>
<th>Conversation A</th>
<th>Cross-turn Repetition Example 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human:</strong> say you're in an empty giant office and you see a black bird flying indoors. what do you do?</td>
<td><strong>Human:</strong> how about Japan?</td>
</tr>
<tr>
<td><strong>Meena:</strong> what kind of bird? is it friendly? how large? How big is the office? Is it indoors or outdoors?</td>
<td><strong>Meena:</strong> I'd love to go to Japan too.</td>
</tr>
<tr>
<td><strong>Human:</strong> well it's a predator bird and it's gonna eat you</td>
<td><strong>Human:</strong> are you lying to me?</td>
</tr>
<tr>
<td><strong>Meena:</strong> I would grab my phone and call 911.</td>
<td><strong>Meena:</strong> I'd love to go to Japan too.</td>
</tr>
</tbody>
</table>

**Human:** I've been using Python for 15 years
**Meena:** Ah I see, so I'm guessing you're pretty pro at it by now then?
**Human:** probably
**Meena:** What's your current project?
**Human:** NLP
**Meena:** Neuro-linguistic Programming?
**Human:** correct
Where are we now?

Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day

By James Vinson | Mar 24, 2016, 6:43am EDT
Via The Guardian | Source: TayandYou (Twitter)

Human: I don't like you!
Meena: I like you, though!
Human: You were insensitive to me being a Vegan!
Meena: I didn't mean for it to come off that way! I'm sorry! You can still love me?
Human: No. I love chicken more
Meena: Oh. I'm sorry. I love you still.
Human: No need
Meena: Oh, but I need.
Human: That's way too much flirty
Meena: No, it's just right.
Human: Bye Meena! Sorry it couldn't work out between us!
Meena: It's okay, you know I'm still in love with you.
Human: It's not me. It's about you
Meena: It's about us.
Why is NLP hard?

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

- Ambiguity at multiple levels:
  - Word sense: **bank** (finance or river)
  - Part of speech: **chair** (noun or verb?)
  - Syntactic structure: **I saw the man with the telescope**
  - Multiple: **I saw her duck**
Ambiguity: Tokenization

这是一个简单的句子

WORDS

This is a simple sentence

הזא משבט הפרט
Ambiguity: Word sense disambiguation

in tea
her daughter

- most of the vowels unspecified
Ambiguity: Tokenization + disambiguation

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

in tea בטה
in the tea בטה
that in tea שבטה
that in the tea שבטה
and that in the tea שבטה
and her saturday כשבתה
and that in tea כשבתה
and that her daughter כשבתה
Ambiguity: Morphology

- unfriend
- Obamacare
- Manfuckinghattan
Ambiguity: Semantic analysis

■ Every language sees the world in a different way

■ For example, could depend on historical conditions

■ Russian has very few words for colors; Japanese has hundreds

■ Idioms: **happy as a clam, it’s raining cats and dogs, les carottes sont cuites**
Ambiguity: Semantics

Every fifteen minutes a woman in this country gives birth.
Ambiguity: Semantics

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

— Groucho Marx
Ambiguity: Syntax + semantics

- I saw the woman with the telescope wrapped in paper.
  - Who has the telescope?
  - Who or what is wrapped in paper?
  - Event of perception or assault?
Dealing with ambiguity

- How can we model ambiguity and choose correct analysis in context?
  - Non-probabilistic methods return all possible analyses.
  - Probabilistic models return best possible analysis, i.e. most probable one according to the model.

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 raw text

- Examples
  - Penn Treebank: 1M words of parsed Wall Street Journal
  - Canadian Hansards: 10M+ words of aligned French/English sentences
  - Yelp reviews
  - The Web / Common Crawl: billions of words of who knows what
Corpus-based methods

- Give us statistical information

All NPs

NPs under S

NPs under VP

<table>
<thead>
<tr>
<th></th>
<th>NP PP</th>
<th>DT NN</th>
<th>PRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>All NPs</td>
<td>11%</td>
<td>9%</td>
<td>6%</td>
</tr>
<tr>
<td>NPs under S</td>
<td>9%</td>
<td>9%</td>
<td>21%</td>
</tr>
<tr>
<td>NPs under VP</td>
<td>23%</td>
<td>7%</td>
<td>4%</td>
</tr>
</tbody>
</table>
Statistical NLP

- Like most other parts of AI, NLP is dominated by statistical methods
  - Typically more robust than earlier rule-based methods
  - Relevant statistics/probabilities **learned from data**
  - Normally requires lots of data about any particular phenomenon
Why is NLP hard?

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

- 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
Sparsity

- 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

- Out of 93,638 distinct words (types), 36,231 (~40%) occur only once.

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

- Order words by frequency. What is the frequency of $n$th 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 NLP hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$
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 on social media?
Why is NLP Hard?
Why is NLP hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathbb{R}$
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 Deni. vs. She gave Deni the book.

Some kids popped by. vs. A few children visited.

Is that window still open? vs. Please close the window.
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.”

skater eats pavement
Unknown representation

- Very difficult to capture what is $R$, since we don’t even know how to represent the knowledge a human has/needs:
  - What is the “meaning” of a word, sentence, utterance?
  - How to model context?
  - Other general knowledge?
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
- Efficient
- Ethical
Symbolic and probabilistic NLP

Logic/Rule-based NLP

Statistical NLP

~1990s
Probabilistic and Connectionist NLP

Engineered features

\[
\text{Reranking Model: } \arg\max_{x \in \mathcal{X}} P(f | e) P(e)
\]

Learned features

~mid 2010s
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 better understand how language works
  - We end up doing this and using it

- Cognitive science
  - How does the human brain work?
  - 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 statistical NLP
Course logistics
What is this class?

Three aspects to the course:

- **Linguistic issues**
  - What are the range of language phenomena?
  - What are the knowledge sources that let us disambiguate?
  - What representations are appropriate?
  - How do you know what to model and what not to model?

- **Statistical modeling methods**
  - Increasingly complex model structures
  - Learning and parameter estimation
  - Efficient inference: dynamic programming, search, sampling

- **Engineering methods**
  - Issues of scale. We’ll focus on what makes problems hard, and what works in practice.
What is this class?

Models and algorithms:

- **Models**
  - State machines (finite state automata / transducers)
  - Logic (first-order logic)
  - Probabilistic models (WFST, language models, HMM, SVM, CRF, …)
  - Vector space models (embeddings, Transformers)

- **Algorithms**
  - State space search (DFS, BFS, A*, dynamic programming — Viterbi, CKY)
  - Supervised learning
  - Unsupervised learning

- **Methodological tools**
  - Training/test sets, evaluation, cross-validation
What is this class?

Outline of topics:

- Words and sequences
  - Classification
  - Language modeling
  - Vector semantics and word embeddings
  - Sequence labeling
- Structured classification
  - Parsing
- Morphology
- Semantics
- Discourse
- Applications
  - Machine translation, summarization, sentiment analysis, computational ethics
What is this class?

Grading

- This is a project-based course. Grading will be done based on 5 homework assignments, as well as class participation.
  - Four projects each worth 20% of your final grade: Classification, sequence labeling, dependency parsing, and coreference.
  - One review/presentation on a paper from this year’s EMNLP: 10%
  - Participation on Piazza: 10%
What is this class?

Requirements and goals

- Class requirements
  - Uses a variety of skills/knowledge
    - Probability and statistics, graphical models
    - Basic linguistics background
    - Strong coding skills (Python)
  - Most people are probably missing one of the above. You will have to work on your own to fill the gaps.

- Class goals
  - Learn the issues and techniques of statistical NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
What is this class?

Readings

- Primary texts
  - Jurafsky and Martin, Speech and Language Processing, 2nd and 3rd Edition (not 1st)
  - Eisenstein, Introduction to Natural Language Processing

- Prerequisites
  - Mastery of basic probability
  - Strong skills in Python or equivalent
  - Deep interest in language
Other announcements

- Course contacts:
  - Webpage: materials and announcements
  - Piazza: discussion forum
  - Canvas: project submissions
  - Homework questions: Recitation, Piazza, TA office hours

- Enrollment: We’ll try to take everyone who meets the requirements

- Computing resources:
  - Homeworks should run on relatively modern hardware just fine. If you need access to more compute for extra credit experimentation, we may be able to arrange for cloud credits.

- Questions?