CS 4650/7650: Natural Language Processing

Introduction to NLP

Diyi Yang

Some slides borrowed from Yulia Tsvetkov at CMU and Noah Smith at UW
Welcome!

Diyi Yang

Ian Stewart

Jiaao Chen

Nihal Singh
Course Website

https://www.cc.gatech.edu/classes/AY2020/cs7650_spring

CS-4650/7650: Natural Language Processing, Spring 2020

NLP: CS-4650/7650
MW 3:00-4:15pm, Kendeda 215
Communication With Machines

~50-70s

~80s

today
Conversational agents contain:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech
In early 2011, an IBM computing system named Watson competed against the world's best Jeopardy! champions.
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?
I study deep learning and machine learning.
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
- Word sense disambiguation
- Semantic role labeling
- ...

NLP lies at the intersection of computational linguistics and machine learning.
Level Of Linguistic Knowledge
Phonetics, Phonology

- Pronunciation Modeling
Words

- Language Modeling
- Tokenization
- Spelling correction

WORDS This is a simple sentence
Morphology

- Morphology analysis
- Tokenization
- Lemmatization
Part of Speech

- Part of speech tagging

**PART OF SPEECH**

**WORDS**

This is a simple sentence

**MORPHOLOGY**

be 3sg present
Syntax

- Syntactic parsing

This is a simple sentence

be 3sg present
Semantics

- Named entity recognition
- Word sense disambiguation
- Semantic role labeling
Discourse

This is a simple sentence

But it is an instructive one.
Where Are We Now?
Where Are We Now?

Baseline mutual information model (Li et al. 2015)

<table>
<thead>
<tr>
<th>A: Where are you going? (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B: I’m going to the restroom. (2)</td>
</tr>
<tr>
<td>A: See you later. (3)</td>
</tr>
<tr>
<td>B: See you later. (4)</td>
</tr>
<tr>
<td>A: See you later. (5)</td>
</tr>
<tr>
<td>B: See you later. (6)</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A: how old are you? (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B: I’m 16. (2)</td>
</tr>
<tr>
<td>A: 16? (3)</td>
</tr>
<tr>
<td>B: I don’t know what you are talking about. (4)</td>
</tr>
<tr>
<td>A: You don’t know what you are saying. (5)</td>
</tr>
<tr>
<td>B: I don’t know what you are talking about. (6)</td>
</tr>
<tr>
<td>A: You don’t know what you are saying. (7)</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

VS

3 GOLDEN GLOBE NOMINATIONS
BEST PICTURE
WINNER
THE BEST FILM OF THE YEAR
HER
JASON MRAZ
SNOW WHITE
SINGING IN THE RAIN
THE BEST PERFORMANCE OF THE YEAR
THE BEST CINEMATOGRAPHY OF THE YEAR
"I LOVED HER: A GORGEOUSLY UGLY PERFORMANCE. SCARLETT JOHANSSON IS SEDUCTIVE AND WINNING.

HER
A SPIKE JONZE LOVE STORY
JASON MRAZ
RACHEL McADAMS
SIGUR ROY
BRIAN NADON
BECKY JACOBS
EVERYWHERE JANUARY 10
Where Are We Now?


Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations
Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations
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 made her duck**
“One morning I shot an elephant in my pajamas”

I made her duck
[SLP2 ch. 1]

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- …
Ambiguity and Scale
The Challenges of “Words”

- Segmenting text into words
- Morphological variation
- Words with multiple meanings: bank, mean
- Domain-specific meanings: latex
- Multiword expressions: make a decision, take out, make up
Part of Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol
Part of Speech Tagging

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

so  he  can  add  u  on  fb  lololol
Part of Speech Tagging

I know, right  shake my head
  ikr    smh    he    asked    for    your
  !      G      O      V      P      D      A    N
  interjection    acronym    pronoun    verb    prep.    det.    adj.    noun

so    he    can    add    u    on    Facebook    laugh out loud
  P      O      V      V      O      P    ^    !
  preposition    proper noun
Syntax

```
NP
  NP
    Adj.  Noun
      natural  processing
NP
  Adj.
    natural
NP
  Noun
    language
NP
  Adj.
    natural
NP
  Noun
    language  Noun
      processing
```
A ship-shipping ship, shipping ships
Every fifteen minutes a woman in this country gives birth.
Every fifteen minutes a woman in this country gives birth. Our job is to find this woman, and stop her!

– Groucho Marx
We saw the woman with the telescope wrapped in paper.
- We saw the woman with the telescope wrapped in paper.
  - Who has the telescope?
  - Who or what is wrapped in paper?
  - An even of perception, or an assault?
Dealing with Ambiguity

- How can we model ambiguity?
  - Non-probabilistic methods (CKY parsers for syntax) return all possible analyses.
  - Probabilistic models (HMMs for POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return the best possible analyses, i.e., the most probable one.
- 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 French/English sentences
  - Yelp reviews
  - The Web!

Rosetta Stone
Like most other parts of AI, NLP is dominated by statistical methods:

- Typically more robust than rule-based methods
- Relevant statistics/probabilities are learned from data
- Normally requires lots of data about any particular phenomenon
Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations
Sparsity

- Sparse data due to Zipf’s Law
- Example: the frequency of different words in a large text corpus

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Token</th>
<th>Frequency</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,698,599</td>
<td>the</td>
<td>124,598</td>
<td>European</td>
</tr>
<tr>
<td>849,256</td>
<td>of</td>
<td>104,325</td>
<td>Mr</td>
</tr>
<tr>
<td>793,731</td>
<td>to</td>
<td>92,195</td>
<td>Commission</td>
</tr>
<tr>
<td>640,257</td>
<td>and</td>
<td>66,781</td>
<td>President</td>
</tr>
<tr>
<td>508,560</td>
<td>in</td>
<td>62,867</td>
<td>Parliament</td>
</tr>
<tr>
<td>407,638</td>
<td>that</td>
<td>57,804</td>
<td>Union</td>
</tr>
<tr>
<td>400,467</td>
<td>is</td>
<td>53,683</td>
<td>report</td>
</tr>
<tr>
<td>394,778</td>
<td>a</td>
<td>53,547</td>
<td>Council</td>
</tr>
<tr>
<td>263,040</td>
<td>I</td>
<td>45,842</td>
<td>States</td>
</tr>
</tbody>
</table>
Sparsity

- Order words by frequency. What is the frequency of nth ranked word?
Regardless of how large our corpus is, there will be a lot of infrequent words. This means we need to find clever ways to estimate probabilities for things we have rarely or never seen.
Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations
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?
  - “ikr smh he asked fir yo last name so he can add u on fb lololol”
Variation
Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations
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*
Unmodeled Variables

“Drink this milk”

World knowledge

I dropped the glass on the floor and it broke

I dropped the hammer on the glass and it broke
Unmodeled 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?
Desiderate 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
Symbolic and Probabilistic NLP

Logic-based/Rule-based NLP

- Analysis
- Transfer
- Generation
- Interlingua
- Direct translation

- Source text
- Target text

~ 90s

Statistical NLP

Translation Model
- Source phrase
- Target phrase
- Translation features

Reranking Model
- Language Model
- Feature weights

$$\text{argmax}_{f} P(f|e) P(e)$$
Probabilistic and Connectionist NLP
To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.

\( R \) is not directly observable.

Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.
NLP vs. 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
- Data science
- Political science
- Psychology
- Economics
- Education
Today’s Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature
Factors Changing NLP Landscape

1. Increases in computing power
2. The rise of the web, then the social web
3. Advances in machine learning
4. Advances in understanding of language in social context
Logistics
What is this Class?

- **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
  - Deep neural networks for NLP: LSTM, CNN, Seq2seq
Outline of Topics

- Words and Sequences
  - Text classifications
  - Probabilistic language models
  - Vector semantics and word embeddings
  - Sequence labeling: POS tagging, NER
  - HMMs, Speech recognition
- Parsers
- Semantics
- Applications
  - Machine translation, Question Answering, Dialog Systems
Readings

- **Books:**
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd or 3rd Edition
  - Also: Eisenstein, Natural Language Processing
Course Website & Piazza

www.cc.gatech.edu/classes/AY2020/cs7650_spring/
piazza.com/gatech/spring2020/cs7650cs4650
Your Instructors

- **Instructor:**
  - Diyi Yang
    - Assistant professor
    - Research interests: NLP, Computational Social Science

- **TAs:**
  - Ian Stewart: PhD, Computational Sociolinguistics
  - Jiaao Chen: PhD, NLP/ML
  - Nihal Singh: MSCS, NLP
TA Office Hours

- Ian Stewart: Tuesdays, 2-4pm, CODA C1106
- Jiaao Chen: Thursdays, 2-4pm
- Nihal Singh: Fridays, 9-11am
Grading

- 4 Homework Assignments (45%)
- 1 Midterm (15%)
- 1 Course Project (40%)

- 45% Homework Assignments
  - Homework 1: 6%
  - Homework 2: 13%
  - Homework 3: 13%
  - Homework 4: 13%

- 15% Midterm Exam
  - No make-up exam unless under emergency situation

- 40% Course Project
  - Project proposal (2 pages): 5%
  - Midway report (4 pages): 10%
  - Final report (8 pages): 20%
  - Presentation (in class presentation): 5%
Late Polices

- Late Policy
  - 4 late days to use over the duration of the semester for homework assignments only. There are no restrictions on how the late days can be used (e.g., all 4 can be used on one homework). Using late days will not affect your grade. But homework submitted late after all 4 late days have been used will receive no credit.

- No make-up exam
  - Unless under emergency situation
Course Project

- Semester-long project (2-3 students) involving natural language processing – either focusing on core NLP methods or using NLP in support of an empirical research question
  - 2-page Project proposal (5%)
  - 4-page Midway report (10%)
  - 8-page Final report (20%)
  - Project presentation (5%)
    - 10-min in-class presentation (tentative)
Other Announcements

- **Course Contacts:**
  - Webpage: materials and announcements
  - Piazza: discussion forum
  - Homework questions: Piazza, TAs’ office hours

- **Computing Resources:**
  - Experiments can take up to hours, even with efficient computation
  - Recommendation: *start assignments early*
What’s Next?

- Text Classification