CS 4705
Natural Language Processing
Fall 2017

Professor Kathy McKeown

Class Size and In-class Discussion

- Piazza to interact in class. For in-class and video students

- Regular intervals: pause to answer questions

- I will also ask questions of you
    - Answers will be shown on the slide
    - Making your name shown to fellow students
    - On the instructor systems, we will see who posed a question - > assign credit for interaction
    - No penalty for wrong answer except for inappropriate content
    - May use an interesting answer to further discussion
    - In-class participation will count towards your grade

Download Piazza App on your Phone

Class Policy on Electronics

- Cell phone in class OK and needed for Piazza interaction
- Keep laptops closed or don't bring to class

Today

- What is NLP?
- Class Logistics
  - What will we cover
  - Helpful background
  - Class homework and exams

What will we study in this class

- How can machines *understand* and *generate* language?
  - Examples drawn from naturally occurring corpora
  - Theories about language
  - Algorithms
  - Statistical methods
  - Applications

Knowledge Needed

- Morphology: word formation
- Syntax: word order
- Semantics: word meaning and composition
- Pragmatics: Influence of context and situation

Goal: discover what the speaker meant

Morphology

- Important for search, machine translation, summarization

- *Union Activities in New York*
  - Singular/plural
    - Union/unions
    - Activity/activities
  - Other languages are morphologically rich
    - Arabic: definite embedded in the word (clitics): The union (Al+) vs. a union, unions
    - German: case part of the word (subj vs obj)
  - *Are there examples in your language?*

Responses

News article titles

- Stud tires out
- Eye drops off shelf
- Teacher strikes idle kids
- Drunk gets nine months in violin case
- Enraged cow injures farmer with ax
- Ban on nude dancing on Governor’s desk
- Hospitals are sued by seven foot doctors
- Red tape holds up new bridges
- Government head seeks arms
- Patient at death’s door – doctors pull him through
- In America a woman has a baby every 15 minutes

Syntax

- Part of speech tagging: is a word a noun, verb, adverb, adjective, etc?

- Parsing
  - Identifying constituents
    - NP: *Kathy McKeown, a man in the park*
    - VP: *was looking up, had risen*
  - Identifying subjects and objects
    - *Bill hit John* vs *John hit Bill*
  - Modification
    - *John saw the man in the park with a telescope*

Part of Speech tagging

- *Stud tires out*
  - *Tires: a noun or a verb?*
- *Eye drops off shelf*
  - *Drops: a noun or a verb?*
- *Teacher strikes idle kids*
  - *Strikes: a noun or a verb?*

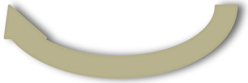
Responses

Constituent Structure and Modification

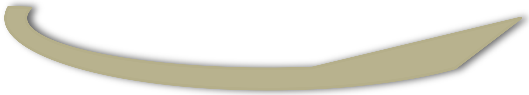
- The problem of PP attachment

  Enraged cow injures farmer with ax

- [Enraged cow] injures farmer [with ax]



- [Enraged cow] injures farmer [with ax]



Representing modification with brackets

- [Enraged cow] [injures [farmer [with ax]]]

- [Enraged cow] [injures [farmer] [with ax]]]

Constituent Structure and Modification

- The problem of PP attachment

  Ban on nude dancing on governor's desk

  NP NP PP

- [Ban] on [nude dancing] [on governor's desk]

- There are two possible modifications? What are they?
- Which one is correct?

Response

Constituent Structure and Modification

- The problem of PP attachment

  Ban on nude dancing on governor's desk

NP PP

- [[Ban] on [nude dancing]] [on governor's desk]

NP PP

- [Ban] on [[nude dancing] [on governor's desk]]

NP

Noun noun modification

- *Water fountain:* a fountain that *supplies* water
- *Water ballet:* a ballet that *takes place* in water
- *Water meter:* a device (called a meter) that *measures* water
- *Water barometer:* a barometer that *uses* water (instead of mercury) to measure air pressure
- *Water glass:* a glass that is *meant to hold* water

Noun noun modification: constituent structure

- *Which constituent structure best represents the meaning of country song platinum album*?

1. *[country [song [platinum album]]]*
2. *[country [[song platinum] album]]*
3. *[[country song] [platinum album]]*
4. *[[country [song platinum]] album]*
5. *[[[country song] platinum] album]*

Response

Noun noun modification and headlines

- *Hospitals are sued by seven foot doctors*

- *Hospitals are sued by [[seven foot] doctors]*

- *Hospitals are sued by [seven [foot doctors]]*

Word Meaning

- *Red tape **holds up** new bridges*
  - *Holds up*:
    1. [TRANSITIVE] to support someone or something so that they do not fall down
       - *Her legs were almost too shaky to hold her up*.
    2. [TRANSITIVE] [OFTEN PASSIVE] to cause a delay for someone or something, or to make them late
       - *Sorry I'm late, but my flight was held up.*
- *Government **head** seeks **arms***
  - *Head:*
  - *Arms:*

Pragmatics

- Discourse context
  - *John went to the store.* ***He*** *bought bread and butter.*
- Situational (real world) context
  - Day after the Charlottesville, VA riots
  - ***His*** *irresponsible actions took the life of a young woman who was just beginning her adult life.*
- Commonsense knowledge
  - ***Boston*** *called and left a message for Joe*.

The Language of Genres:
news, journal, social media, novel?

- Devastating, but yet amazing storm. IMBY just some branches, ton of leaves, …This surpasses any Big daddy storm, should be called Big Mama. (Computer Guy)
- Produced by a team of 26 scientists led by the University of New South Wales Climate Research Centre, the Diagnosis convincingly proves that the effects of global warming have gotten worse in the last three years. (Somerville et al 2011)
- Hurricane Sandy churned about 290 miles off the Mid-Atlantic coast Sunday night, with the National Hurricane Center reporting that the monster storm was expected to come ashore with near-hurricane-force winds and potentially "life-threatening" storm surge flooding.
- *What is the matter?' I cried. 'A wreck! Close by!'*

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

What data is available for learning?

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

What features yield good predictions?

Machine Learning Methods

- Supervised
  - Support vector machine, Naïve Bayes, Logistic regression
  - Sequence labeling: Hidden Markov Modeling (HMM), Conditional Random Fields (CRF)
  - Neural networks
- Unsupervised
  - Clustering
- Semi-supervised
  - Boot-strapping, self-training, co-training
  - Distant learning

Where does the data come from?

- Manually labeled
- Naturally occurring
- A noisy, but plentiful substitute

Core NLP

- Morphological analysis
- Part of speech tagging
- Parsing

Applications

- Searching very large text and speech corpora
  - E.g., the web
- Question answering over the web
- Translating between one language and another: e.g., Chinese and English
- Summarizing text: e.g., your email, the news, reviews
- Sentiment analysis
- Generating texts
- Dialog systems: Amtrak's Julie

Logistics

Instructor

- Kathy McKeown
  - Office: 722 CEPSR
  - NLP Group
  - 35 years at Columbia, Founding Director of the Data Science Institute (just stepped down)
- Research
  - Summarization
  - Question Answering
  - Language Generation
  - Sentiment analysis
  - Multilingual applications

TAs

- Elsbeth Turcan (head TA)
- Dheeraj Kalmekolan
- Apoorv Kulshreshtha
- Robert Kwiatkowski
- Fei-Tzin Lee
- Bhavana Ramachandra
- Samarth Tripathi

Background

- Programming. We will use Python

- In addition, at least one:
  - Artificial Intelligence
  - Machine learning
  - Programming Languages and Translators
  - Statistics

Syllabus

- Available at: http://www.cs.columbia.edu/~kathy/NLP/2017

Textbooks

- Speech and Language Processing, 2nd Edition, by Jurafsky and Martin. It will be available from Book Culture, as well as from Amazon and other online providers. It is also on reserve in the Science Library.

- Neural Network Methods for Natural Language Processing by Yoav Goldberg. It is available online but you can also purchase hard copy from the publisher.

Assignments

- 4 homework assignments: 3 programming, 1 written
  - We will be using Google Cloud
  - HW0:
    - Worth 2 points, but if you do not/can not do it, this is not the class for you.
    - Sets up your google cloud account properly
  - Four free late days
  - After that 10% off for each day late
- Midterm and final
- Evaluation: 50% homework + 40% exams + 10% class participation (via Piazza)

Academic Integrity

- Copying or paraphrasing someone's work (code included), or permitting your own work to be copied or paraphrased, even if only in part, is forbidden, and will result in an automatic grade of 0 for the entire assignment or exam in which the copying or paraphrasing was done. Your grade should reflect your own work. If you are going to have trouble completing an assignment, talk to the instructor or TA in advance of the due date please. Everyone: Read/write protect your homework files at all times.

For Next Class

- Read Chapters 1-2 of J&M, Chapter 1 of NN
- Questions? (Use Piazza)

Questions