Natural Language Processing Introduction

Adam Meyers
New York University

2023

Outline

- Administrative Matters
  - Grades, Exams, Policies, etc.
- Assignment 1: A Manual Annotation Task
- Text Books and Suggested Reading
- Discussion of Background Knowledge
- Defining the Field
- CL Applications
- Types of Text Analysis used in CL
- Summary and Syllabus

Administrative Matters

Covid Policy: Masks

- NYU no longer requires masks
- I may be wearing a mask (N95 or KN95) in class if the COVID rate is high
- If you move to within 5 feet of me to talk before or after class
  - I will put on a mask if you are wearing one
  - If I am wearing a mask, please put a mask on

Covid Policy: sickness

- If you are sick, please stay home
  - Please test for Covid and follow appropriate medical protocols
  - My slides are available on line
  - Video versions of some lectures may be available on request (from the Covid lockdown).
- If I am sick, I will stay home.
- If I am sick, but well enough to teach, I will teach through zoom – If this occcurs, I will email zoom links to the class.

Undergraduate Schedule

- **First Day of Class: Sept 5**
- **Last Day of Class: Dec 14**
- **Tuesday & Thursday 2PM to 3:15 PM**
- **Room: Wavery G08**
- **No Class:**
  - **Oct 10 (Administrative Monday)**
  - **Nov 23 (Holiday)**
- **Professor and TA Office Hours Listed here**
  - **https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/priv/links.html**

Graduate Schedule

- **First Day of Class: Sept 5**
- **Last Day of Class: Dec 12**
- **Tuesday 4:55–6:55 PM**
- **Room: CIWW 101**
- **No Class:**
  - **Oct 10 (Administrative Monday)**
- **Professor and TA Office Hours Listed here**
  - **https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/priv/links.html**

Undergraduate Websites and Contact Info

- Course Website: https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/
  - Updated throughout the term. May include “broken” links for future slides to be presented in class.
- Gradescope, Brightspace, Edstem, Office Hour and other Zoom, etc.:
  - https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/priv/links.html
- Data for Assignments and Final Project
  - https://drive.google.com/drive/u/3/folders/1a7La0Vqq1Q47wXpMG38gl3pMl_IxZ-6L
  - Sign in with your netid and password to gain access
  - Members of class are in a google group with access to the above link
    - fa23_csci-ua_480_1_057_292662
- **My Email:** meyers@cs.nyu.edu

Graduate Websites and Contact Info

- Course Websites: https://cs.nyu.edu/courses/fall23/CSCI-GA.2590-001/
  - Updated throughout the term. May include "broken" links for future slides to be presented in class.
- Gradescope, Brightspace, Edstem, Office Hour and other Zoom, etc.:
  - https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/priv/links.html
- Data for Assignments and Final Project
  - https://drive.google.com/drive/u/3/folders/1a7La0Vqq1Q47wXpMG38gl3pMl_IxZ-6L
  - Sign in with your netid and password to gain access
  - Members of class are in a google group with access to the above link
    - fa23_csci-ga_2590_1_001_292664
- **My Email:** meyers@cs.nyu.edu

Accessing Class BrightSpace, Zoom Google Drive, and other Links

- If you are signed in to Google, make sure you are identified as your NYU user (the same as BrightSpace)
- If it is not working
  - open a private browser window
  - sign in to your NYU user in that windows
  - Access in that window should work

The Class is a Work in Progress

- **Ask me Questions!**
  - **Questions may lead to interesting details**
  - **I might proceed to quickly if you don’t ask :)**
- Goals of Class:
  - Overview of the field of Natural Language Processing
    - Evaluation based on midterm and long & short homework assignments
  - Work on a Collaborative Longer Term Project with a Paper
    - Evaluation based on Final Projects (and long homework assignments)
    - Wider implications: conference papers, CV, grad school & job applications
- Final Projects
  - All final projects will be completed by teams of 3 or 4 students
- Midterm: paper and pen/pencil test in class
- TA help
  - 1st ½ of semester: TA office hours for help with assignments
  - 2nd ½ of semester: Mentors supervise weekly progress on projects

Grade = 30% Long HW + 30% Midterm + 5% Short HW + 35% Final Project

- Long Homework – 6 Assignments (1st half of semester)
  - 1 annotation assignment & 4 shared tasks submitted through gradescope
  - 1 Final project proposal (submitted through Brightspace)
- Midterm: covers Lectures, Reading Material and HW for the first ½ Semester
- Short Homework: based on Assigned Readings & Lectures for second ½ Semester
- Final Project
  - Sample Topics
    - https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/final_project_info.html#Topics
    - Slides for lectures include related final project descriptions
  - Opportunities for feedback before final draft is due
    - Final Project Proposal: Counts as 1 homework
    - 30 Second Progress Reports
    - First Draft (Optional)
    - Short Talks
    - Regular meetings with TA mentors
  - **Will be evaluated based on written report (unusual for undergraduate STEM class)**

Intellectual Integrity

- http://www.cs.nyu.edu/webapps/content/academic/undergrad/academic_integrity
- Midterm – I will help you during the midterm by explaining instructions or fixing errors in the phrasing of questions, but nobody else should help you.
- Homework – discuss with anyone, but your work should be your own.
  - Be prepared to solve it on your own after you submit your answer
- Final Project
  - Advice is fine: professor, mentor, classmates, etc.
  - Collaboration is encouraged.
    - You may need people to do an annotation or an evaluation task that you lead.
    - Multiple teams can "compete" and refer to each other's work
  - All team members should contribute – no freeloading

Homework Restrictions

- Instructions for HW requires specifics
  - Filename guidelines, Output format, Other
- Important for Grading
  - Most assignments submitted via GradeScope
    - Instant feedback for shared-task style assignments, using evaluation programs
    - GradeScope may reject homework in the wrong format
      - **This can be problematic for late homework submissions**
  - Final Project proposal and drafts submitted via BrightSpace
- If in doubt, it should work in linux:
  - All students have a linux account at NYU
  - Using linux (not Apple or Windows) eliminates some compatibility issues.
- Questions about fixing file formats (and other matters) – through edstem

How to Use NYU’s Linux Cluster

- https://cims.nyu.edu/webapps/content/systems/resources/computeservers
  - Website about Linux Servers
- Undergrad: All students were given accounts: log in early in the semester
- Grad students: https://cims.nyu.edu/dynamic/systems/userservices/accounts/obtain/
- Logging in:
  - ssh to access.cims.nyu.edu
    - This is the gateway computer – do not run large programs on access
    - Copy files between your home computer and “access” using scp
  - ssh from access to a compute servers listed on the Linux server website
  - “access” shares file systems with the compute servers
- A useful program for file formats:
  - dos2unix
- Other factors
  - Path variables are set properly
  - Follow specifications provided with homework: filenames, format, etc.
  - Submit individual files, not zip files and directories of files

Late Homework Policies

- **Natural Consequence**
  - Falling behind leads to lower marks on midterm because midterm is based on material learned by doing homework
  - Less time to work on final project
  - Less help from TAs to fix format issues or other grading issues
- **Additional Consequences and Constraints:**
  - Contact me before submitting homework more than 3 weeks late
  - TAs are less available to help with format issues in 2nd ½ of semester
  - Homework submitted after last day of class may not be graded
  - If graded, maximum 1 point may be taken off for lateness (out of 1-10).
- **Pay attention to deadlines on the class website**
  - Gradescope allows late submissions

Final Project

- Details on Final Project Website
- A paper supported by experiments
  - A program that performs a task and scored against an answer key. Discuss task, methods, evaluation, previous work, etc.
  - Annotation defining a task. Similar to program option. Evaluate for annotator consistency
  - Evaluation of Systems – Papers about evaluation metrics, differences between systems, etc.

Purpose of Brightspace

- Submit project proposals, 1st draft & final draft
- Project Grades & Comments
- **Google groups**
  - fa23_csci-ua_480_1_057_292662
  - fa23_csci-ga_2590_1_001_292664
  - Based on class Brightspace accounts
  - **Used for access to resources**
  - **Data for assignments and final projects**

GradeScope

- Online platform for grading
- Facility for automatic grading of programming assignments
  - Low Scores often meaning “formatting issues”
  - Make sure required filenames exist
  - Make sure output file has required number of lines (if applicable)
  - Make sure each line is in the correct format
  - Using linux sometimes fixes format issues
    - For example, it is sometimes a good idea to run “dos2unix” on the output of a Windows system
- Platform for consistent manual grading
  - Midterm
  - Short Homework Assignments

Purpose of edstem

- Programming Assignment Help
  - Format incompatibility with autograder
  - Other questions
  - Topics A1, A2, A3 … correspond to assignments 1, 2, 3, ...
- Other Grading Questions
  - Midterm assignment & “short” Homeworks
- Coordination of Research Group Creation

Purpose of NYU Drive

- Distribute data
  - Including some larger files
  - Including data with licensing restrictions
- https://drive.google.com/drive/u/3/folders/1a7La0Vqq1Q47wXpMG38gl3pMl_IxZ-6L
- Available with your netid and password (group readable, where the group is defined by the class Brightspace account)

Assignment 1 Specifications

https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/homework1.html

Specifications for Homework Annotation Task

- Adjectives occur in two main positions
  - Attributive
    - Adjectives precede nouns that they modify
    - Ex: ***the big sandwich***
  - Predicative
    - A noun phrase is linked to an adjective by predication, sometimes with arguments of the adjective
    - Ex: ***The sandwich is big***
    - Ex: ***I made the sandwich big***
    - Ex: ***Philosophy may seem difficult to understand***
- Adjectives can have three morphological forms
  - Normal: ***big***
  - Comparative: ***bigger*** (Do not mark multi-word non -er cases as comparative, e.g., "***more angry***")
  - Superlative: ***biggest*** (Similarly, do not mark cases like "***most significant***" as superlative)
  - This is a **morphological** classification of a **word** which pertains to suffixes and prefixes
  - It is not a **semantic** classification and it is not a classification of **word sequences** (more than 1 word)
- Adjectives should not be confused with nouns
  - Ex: ***The truck salesman***
    - ***truck*** is not an adjective
    - Nouns, not adjectives can occur in the plural, e.g., ***trucks***
    - Nouns, not adjectives are modified by determiners like ***the*** or ***a***, e.g., ***a truck***, ***the truck***

Adjective Specifications: Slide 2

- A word can have distinct meanings as an adjective and as a noun
  - ***They are studying for the final.*** (Noun)
  - ***This was their final attempt.*** (Adjective)
- Adjectives may be used as nouns with an adjective + *one(s)* meaning
  - ***They exploit the poor.***
    - ***poor*** means something like poor ones or poor people
    - ***poor*** is used much more frequently as an adjective
    - Compare with predicative position
      - ***They are poor***
      - If it was a noun, poor would be plural (poors)
- Frequency is an issue, e.g., assume color words are adjectives in ambiguous contexts because noun-like uses are rare
  - ***I really love this red*** (noun) vs ***That clown nose is red*** (adj) or ***The red nose***
- Idiomatic Constructions involving verbs and adjectives – if the word occurs elsewhere as an adjective and there is no other obvious part of speech, it is probably a predicative adjective
  - ***Mary fell ill. John went ballistic.***
- Idioms and near idioms with literal interpretations that favor adjective marking – mark as adjectives
  - ***red herring, cold shoulder***, …
- Present participles (***flowing, flying***, …) and past participles (***understood, fixed***, …) of verbs
  - Mark as adjective in attributive position (***flowing hair, fixed position***, …)
  - Only mark as adjective in predicative position if an adjective meaning is clear, e.g., understanding means two different things in these examples – only the first meaning is adjectival. Other clues are modifiers (***very*** vs ***easily***)
    - ***John was very understanding***
    - ***John was understanding the lecture easily.***

Adjective Specifications Slide 3

- Determiners (see list) ARE NOT adjectives (they occur before adjectives)
  - ***such several one most more many less few enough both all your those this these their the that some our no neither my its his her every either each any another an a***
- Cardinal Numbers ARE NOT adjectives: ***one, two, three, …*** (they are determiners)
- Ordinal Numbers ARE adjectives: ***first, second, third, ...***
- Adjectives can be modified by other words:
  - ***light red, very hungry, quite upset***
- In attributive position, they occur after any determiners and before any nouns, e.g.,
  - ***the hairy mountain gorilla***
    - ***the*** = determiner
    - ***hairy*** = adjective
    - ***mountain*** = noun
    - ***gorilla*** = noun

Demo of Mae Annotation Tool

- MAE – Latest version of Mae annotation tool (Amber Stubbs, Keigh Rim)
  - https://github.com/keighrim/mae-annotation/
- Put all files from zip in the same directory:
  - mae-2.2.10-fatjar.jar
  - adjective.dtd
  - state_of_the_union.txt
  - AM_state_of_the_union.xml
    - sample with Instructor annotation of first 2 paragraphs
- java -jar mae-2.2.10-fatjar.jar (or double click)
- File → load → dtd → adjective.dtd (dtd file defines the task – write a different dtd file for a different task)
- File → load → state_of_the_union.txt
  - You can also load your saved xml file
- To mark an adjective:
  - Drag left mouse over adjective and click, only one choice
  - Change attributes in list of adjectives (left click and select on slot)
  - Click yes on done
    - go to repeated cases by double clicking on new row created
    - Change the new row if necessary
    - Or delete the new row by right click and select
- Periodically save:
  - File → Save File as XML – choose a name other than the original filename e.g., alm4_state_of_the_union.xml
  - Rename this as submission.xml before submitting it to gradescope.

Text Books, Readings, Software, Packages, Other Resources Etc.

Text Books

- SPEECH and LANGUAGE PROCESSING **2nd Edition**
  - By Daniel Jurafsky and James H. Martin
  - http://www.cs.colorado.edu/~martin/slp.html
  - Overview of the Field, explanations of techniques, algorithms, etc.
  - Parts of the 3rd edition draft are available online:
    - https://web.stanford.edu/~jurafsky/slp3/
- Natural Language Processing with Python
  - By Steven Bird, Ewan Klein, and Edward Loper
  - http://www.nltk.org/book (look at the rest of the website also)
  - Book is available on line (or you can purchase a paper version)
    - Online version may be more up-to-date then the paper version
  - Mostly Python 3, but there may be some legacy compatibility with Python 2
    - Paper version of the book is available
    - Electronic version is continually being revised by authors
  - Downloadable open source NLP programs to try out, inspect and possibly modify

More Stuff to Read/Download, etc.

- My website: http://nlp.cs.nyu.edu/people/meyers.html
  - Termolator – an open source terminology extraction tool
  - GLARF: processing tool written in Common Lisp (for linux)
  - NomBank: annotation project
  - COMLEX, NOMLEX: lexicon projects
  - Web of Law – current research, including student researchers
    - http://nlp.cs.nyu.edu/meyers/web_of_law.html
- Other useful links:
  - Last term's NLP Class: https://cs.nyu.edu/courses/spring23/CSCI-UA.0480-057/
  - Association for Computational Linguistics: http://aclweb.org/
  - ACL repository of conference papers in NLP: http://aclweb.org/anthology/

Linguistic Data Consortium

- Their catalog: https://catalog.ldc.upenn.edu/
- They have data which may be useful for your final project and NYU has a license
- If you need LDC data, contact me or the nyu library (this is a relatively new thing)
  - If you are logged in to the library, try this link: https://catalog-ldc-upenn-edu.proxy.library.nyu.edu/
- LDC has different types of corpora (text), some of it manually annotated.
  - Some data is already available elsewhere: freely downloadable or via the class Resources page

Some Pointers for Installing NLTK on your own Machine

- **Website:** https://www.nltk.org/install.html
  - Install the programs and the data
- **Linux:** NLTK is easy to install in **linux**
- **NYU Linux servers:** install nltk locally and use my version of the nltk data:
  - Command is: **nltk.data.path.append("/home/meyers/nltk_data/“)**
  - There may not be enough diskpace to install the data yourself
- **Apple:** OK, but not as smooth as linux. To get all the bells and whistles, you may have to register as a developer
- **Windows:** I have not tested recently. Linux recommended if you have trouble.
- **Use Python3: Support has ended for Python 2**
  - It is not impossible to get Python2 to work, but I don’t think it is worthwhile

Background

Computer and Math Background

- UNIX
  - Many NLP resources work better/are easier to install/etc. in UNIX
  - Most common Unix platforms today: Apple (BSD) and Linux (preferred)
  - All software provided in this class will run in Linux. No guarantee for other OSs
    - **Whenever a homework script does not run properly, I suggest running it in linux**
  - **CS linux accounts are available for all students in this class. See**
    - https://cims.nyu.edu/webapps/content/systems/resources/computeservers
  - It is possible to get access to an **HPC** (high performance computing) account if these are not sufficient
    - https://sites.google.com/nyu.edu/nyu-hpc/accessing-hpc/getting-and-renewing-an-account
    - Some of the links will only work from an NYU computer (or an NYU VPN)
- UNIX utilities and script languages: grep, shell scripts, sed, awk, etc.
- Programming Languages
  - Some experience with Python helpful for NLTK and example code discussed in class.
  - Some homework assignments use JAVA, but possible to use as black box
  - Any programming language OK for some assignments, e.g., Final Project
  - Common Lisp makes some NLP data structures easier to process, but few people use this now
- Experience writing large programs helpful
  - Programs that may take minutes or hours (not seconds) to run
- Math: Probability and Statistics are especially useful

Linguistics in NLP

- Words & Tokens: morphology, segmentation, part of speech tagging, lexicography, punctuation (relevant for analyzing text), etc.
- Syntax:
  - NLP:
    - descriptive grammars for indivividual languages
    - Many different "frameworks" adopted depending on available tools and other resources
  - Theoretical linguistics, e.g., Chomskyan linguistics at NYU
    - Theory may not be descriptively adequate for single language
    - Must handle all languages
- Phonetics, Phonology: covered in NLP, but omitted here due to time
- Discourse, Pragmatics: discourse arguments, anaphora
- Semantics: Sense Disambiguation, Predicate Argument Structure, Word Similarity and other areas.

Role of Linguistic Theory in Computational Linguistics

- Framework = Language for Expressing Theory
- Theory = Set of Statements in Framework
- Different Theories/Frameworks are typically designed with different interests/biases/etc.
  - Chomskian Linguistics: Meta Grammar for all languages, set of primatives,
- Computational Linguistics is Applied Field of Study
  - Theories/Frameworks are important to the extent that they help make a successful application
  - Descriptive Adequacy is more important than Explanatory Adequacy
  - **The authors of the answer key determine the framework**
  - Some systems handle multiple theories/frameworks
- Frameworks that are popular in CL: Statistics-based Analysis (various), Dependency Grammar, Penn Treebank (based on 1980s Chomskian Linguistics), PropBank/Nombank (~ Relational Grammar), Frame Semantics (based on FrameNet), ...
- Only Broad Coverage Grammars are suitable, e.g., old theories with descriptive track records
- Proviso: there is a small niche within CL, in which researchers implement new theories

Defining Computational Linguistics

- AKA, Natural Language Processing (NLP), Language Engineering, ...
- **Domain**: The set of problems involving the interpretation and generation of human language text and speech
- **Properties**
  - As with applied science: the proof is in the pudding
  - Sometimes at odds with theoretical linguistics
    - Need not model human abilities and human methods
    - Need not correspond to published linguistic theories
    - Sometimes draws on linguistic theories and/or studies of human processing
  - Broad and changing domain influenced by available funding

Computational Linguistics Applications

CL Applications: Slide 1

- **Machine translation**
  - Methods are not at all based on how humans translate
  - Effective for gisting text, generating 1st draft translations, but not for high-level translation
  - Works better for “controlled languages” – technical manuals (Microsoft, Catterpiller, etc.)
  - Systran: https://translate.systran.net/?lang=en
  - Google: http://www.google.com/language_tools?hl=en
- **Spoken Language**
  - dictation (IBM ViaVoice, Dragon Naturally Speaking)
  - Telephone-based customer support (phone mazes)
- **Information Retrieval**
  - Finding documents based on a query, e.g., Web Searches

CL Applications Slide 2

- **Information Extraction**
  - Dealtime, Google Products, Monster.com (job search)
  - Some open source tools:

    https://opennlp.apache.org/

    http://alias-i.com/lingpipe/
  - Tools on NYU website include:
    - http://nlp.cs.nyu.edu/projects/index.shtml#t-r-i
    - http://cs.nyu.edu/grishman/jet/jet.html
    - http://nlp.cs.nyu.edu/ice/
    - http://nlp.cs.nyu.edu/termolator/
  - Example from disease domain http://nlp.cs.nyu.edu/info-extr/biomedical-snapshot.jpg
- **Question Answering**
  - ask.com, Wolfram Alpha, MIT start: http://start.csail.mit.edu/
- **Summarization**: http://textsummarization.net/text-summarizer
- **Spelling/Grammar Checking, etc.** https://languagetool.org/
- **Other NLP demos:** **https://towardsdatascience.com/the-best-nlp-tools-of-early-2020-live-demos-b6f507b17b0a**

Levels of Analysis

Lowest Level Syntactic Processing (text)

- **Tokenization and Segmentation**
  - Given a sentence, determine the words or word-like units that it consists of:
    - ***They announced in unison, “We don't agree with each other.”***
    - Tokenization: ***They*** | ***announced*** | ***in*** | ***unison*** | ***,*** | ***“***| ***We*** | ***do*** | ***n't*** | ***agree*** | ***with*** | ***each*** | ***other*** | ***.*** |***”***
      - Controversial parts: ***n't***, ***each other***
  - NLTK command: ***nltk.word_tokenize('this is a sentence')***
- **Part of Speech Tagging** (modified PTB)
  - Apply a set of part of speech tags to a set of tokens
    - ***They***/PRP ***announced***/VBD ***in***/IN ***unison***/NN ***,***/PU “/PU ***We***/PRP ***do***/VBP ***n't***/RB ***agree***/VB ***with***/IN ***each***/DT ***other***/JJ ***.***/PU ”/PU
  - NLTK command: ***nltk.pos_tag(tokens)***

Low Level Syntactic Processing

- **Named Entity Tagging** (with a little semantics)
  - Mark boundaries of names of type PERSON, ORGANIZATION, FACILITY, GPE, LOCATION, …
  - <ENAMEX TYPE=”PERSON”> Adam Meyers</ENAMEX> works for <ENAMEX TYPE=”ORGANIZATION”>New York University</ENAMEX>
  - test_sentence = 'Adam Meyers works for New York University.'
  - NLTK command: ***nltk.chunk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(test_sentence)***
- **Chunking -**
  - mark verb groups and/or noun groups, convenient approximations of syntactic units
  - [$_{NG}$ ***The book***] ***with*** [$_{NG}$ **the *blue cover***] [$_{VG}$***will end up]*** ***on*** [$_{NG}$ **the *shelf***]**.**
  - do not include “right modifiers”, like constituents derived in parsing (next slide)
  - NLTK:
    - sentence = 'The book with the blue cover will end up on the shelf.'
    - chunks = r"""

      NG: {(<DT|JJ|NN>)*(<NN|NNS>)}

      VG: {<MD|VB|VBD|VBN|VBZ|VBP|VBG>*<VB|VBD|VBN|VBZ|VBP|VBG><RP>?}

      """
    - chunks_grammar = nltk.RegexpParser(chunks)
    - chunks_grammar.parse(nltk.pos_tag(nltk.word_tokenize(sentence)))

Parsing: High Level Syntatic Processing

-
  ```
  (S (NP (DT the) (NN book)
             (PP (IN with)
                   (NP (DT the)
                         (JJ blue)
                         (NN cover))))
        (VP (VBZ is)
             (PP (IN on)
                   (NP (DT the) (NN shelf)))))
  ```

Semantics – ish

- Semantics – A wide range of topics loosely referring to “meaning”
- Some Example Topics which may be part of Semantics (Next Few Slides)
  - Word Sense Disambiguation
  - Predicate Argument Structure
  - Anaphora
  - Discourse Argument Structure
  - “Semantic Parsing”

WordNet Noun entry for *bank*

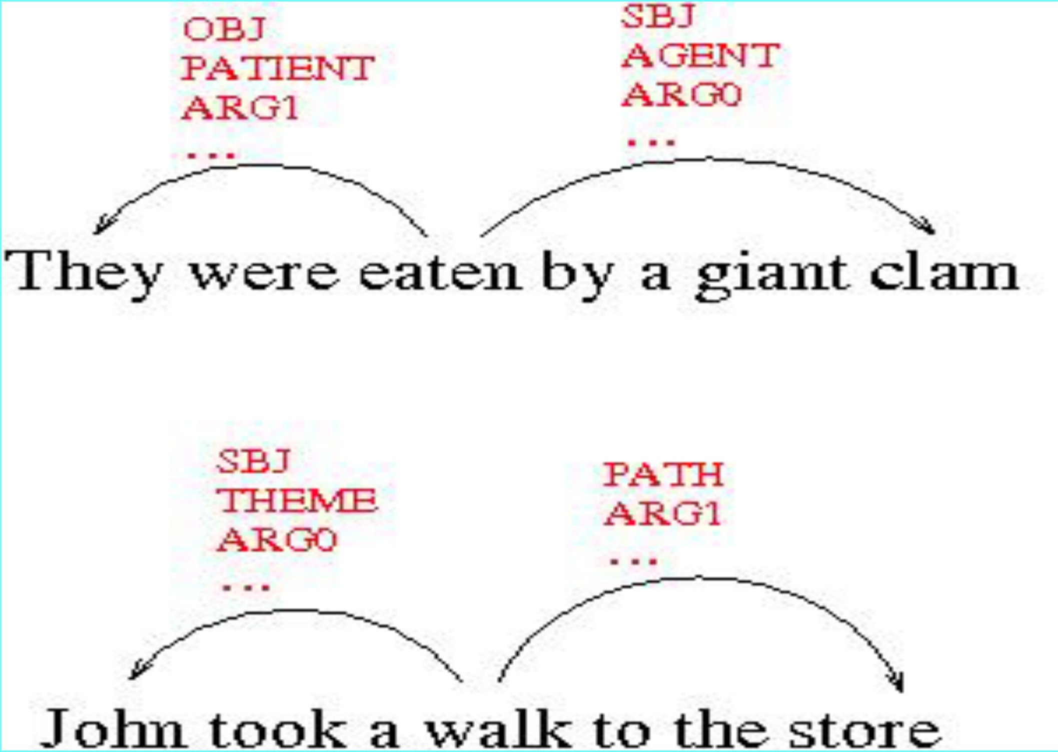
1. S: (n) bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"
2. S: (n) depository financial institution, bank, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) "he cashed a check at the bank"; "that bank holds the mortgage on my home"
3. S: (n) bank (a long ridge or pile) "a huge bank of earth"
4. S: (n) bank (an arrangement of similar objects in a row or in tiers) "he operated a bank of switches"
5. S: (n) bank (a supply or stock held in reserve for future use (especially in emergencies))
6. S: (n) bank (the funds held by a gambling house or the dealer in some gambling games) "he tried to break the bank at Monte Carlo"
7. S: (n) bank, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
8. S: (n) savings bank, coin bank, money box, bank (a container (usually with a slot in the top) for keeping money at home) "the coin bank was empty"
9. S: (n) bank, bank building (a building in which the business of banking transacted) "the bank is on the corner of Nassau and Witherspoon"
10. S: (n) bank (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) "the plane went into a steep bank"

Word Sense Disambiguation

- For interesting characterizations of word senses (and relation between senses), use WordNet (online or download it)
  - wordnet.princeton.edu/
- Fewer than 10 obviously distinct senses of ***bank***, e.g.,
  - *They took money out of the* ***bank****.*
  - *The water flooded over the* ***bank*** *of the river.*
- Difficult sense disambiguation
  - Example: senses 2, 6 and 9 are arguably not distinct
  - Lexicographers are acutely aware of the merging vs. splitting problem of enumerating senses
  - CL systems usually collapse some WordNet distinctions

Predicate/Argument Structure

- For thousands of years, linguists have employed systems to characterize predictable paraphrases, e.g., Pāṇini, a Sanskrit linguist from the 4rth Century BC
- In 21st Century CL, semantic role labeling is popular

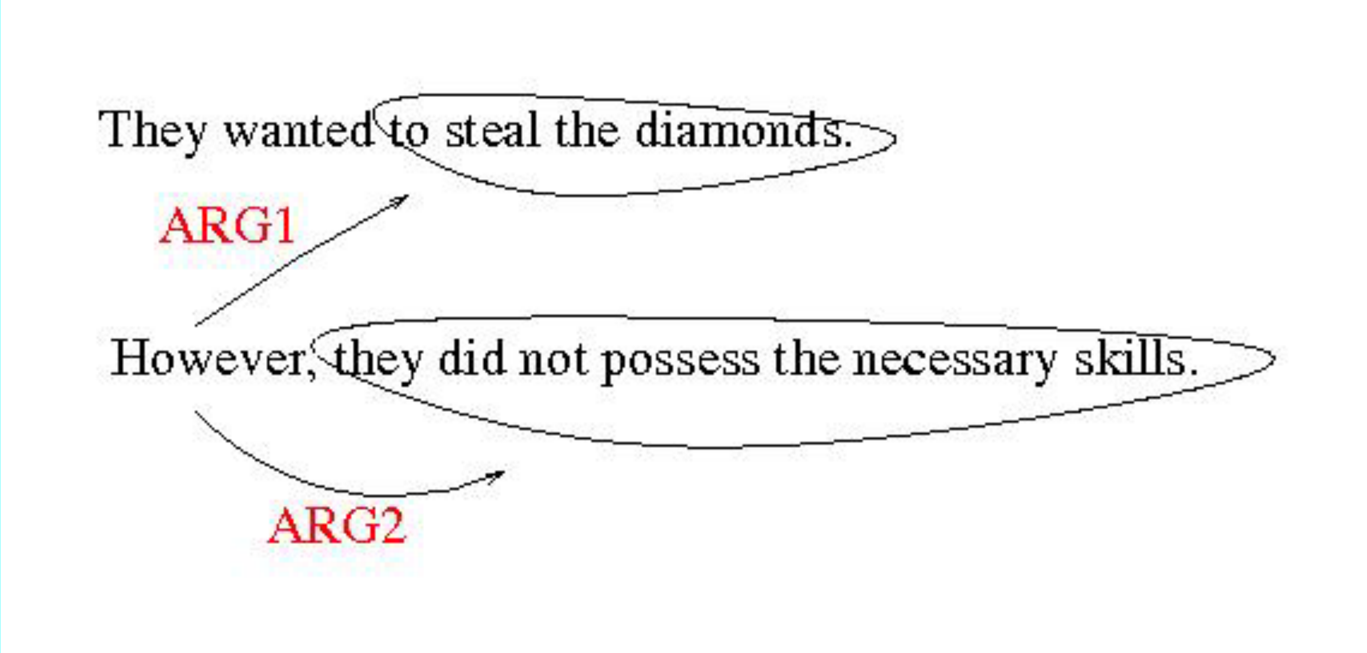

OBJ
PATIENT
ARG1
...
SBJ
AGENT
ARG0
...
They were eaten by a giant clam
SBJ
THEME
ARG0
...
PATH
ARG1
...
John took a walk to the store

Anaphora

- Coreference
  - Though **Big Blue** won the contract, this official is suspicious of **IBM**.
  - **Mary** could not believe what **she** heard.
- Other Varieties
  - John ate **a sandwich** and Mary ate **one** also. [type coref]
  - **The amusement park** is very dangerous. **The gate** has sharp edges. **The rides** have not been inspected for years. [Bridging Anaphora]
  - **This book** is valuable, but **the other book** is not. [Other coref]

Discourse Argument Structure

- Adverbs, Subordinate/Coordinate Conjunctions, among other words link clauses



They wanted to steal the diamonds.

However, they did not possess the necessary skills.

Semantic Parsing, e.g., GLARF

- Means different things to different researchers, but my version of semantic parsing is called GLARF:
  - http://nlp.cs.nyu.edu/meyers/GLARF.html
- One representation of the sentence that includes as much information as possible: lexical categories, predicate argument structure, discourse annotation, etc.
- Next slide is a representation of the sentence:
  - ***Afterwards, she decided to perform the operation.***
  - When it occurs after the sentence: ***The doctor ran some tests***

GLARF

```
(S (ADV (ADVP (HEAD (ADVX (HEAD (RB Afterwards 0))
                            (P-ARG1 (S (EC-TYPE PB)  (INDEX 0+0))
                            (P-ARG2 (S (EC-TYPE PB)  (INDEX 0))
                            (RELATION-TYPE AFTER)))
              (INDEX 1)  (POINTER 0:1))))
   (PUNCTUATION (, , 1))
   (SBJ (NP (HEAD (PRP she 2))  (INDEX 2) (POINTER 2:1))))
   (PRD (VP (HEAD (VG (HEAD (VBD decided 3))
                      (P-ARG0  (NP (EC-TYPE PB) (INDEX 2)))
                      (P-ARG1 (S (EC-TYPE PB) (INDEX 5)))
                      (P-ARGM-TMP (ADVP (EC-TYPE PB) (INDEX 1)))
                      (SEM-TENSE PAST)))
           (COMP  (S (L-SBJ  (NP (EC-TYPE INF) (INDEX 2)))
                     (PRD (VP  (HEAD (VG  (AUX (TO to 4))
                                          (HEAD (VB perform 5))
                                          (P-ARG0 (NP (EC-TYPE PB) (INDEX 2)))
                                          (P-ARG1 (NP (EC-TYPE PB) (INDEX 4)))
                                          (INDEX 3)))
                               (OBJ (NP  (Q-POS (DT the 6))
                                         (HEAD (NX  (HEAD (NN operation 7)))
                                                    (P-SUPPORT (VG  (EC-TYPE PB) (INDEX 3)))
                                                    (P-ARG0 (NP (EC-TYPE PB) (INDEX 2)))))
                                    (INDEX 4)  (POINTER 6:1)))
                            (PB-POINTER 4:1)))
                    (POINTER 4:2)  (INDEX 5)))
            (POINTER 3:1)))
   (PUNCTUATION (. . 8))  (POINTER 0:2) (TREE-NUM 1) (INDEX 0))
```

Role of Manual Annotation

- Used to create, test and fine-tune task definitions/guidelines.
  - For a task to be well-defined, several annotators must agree on classification most of the time.
  - If humans cannot agree, it is unlikely that a computer can do the task at all
  - Popular, but imperfect measurement of agreement:
    - $$Kappa=\frac{Percent(Actual\ Agreement)-Prob(Chance\ Agreement)}{1-Prob(Chance\ Agreement)}$$
- Used to create answer keys to score system output
  - One set of measures are: recall, precision and f-score
  - $$Recall=\frac{|Correct|}{|Answer\ Key|} \quad Precision=\frac{|Correct|}{|System\ Output|} \quad F-Score=\frac{1}{\frac{1}{2}*(\frac{1}{Precision}+\frac{1}{Recall})}$$

Manual Annotation in Supervised Statistical ML

- Divide the corpus into sub-corpora
  - A training corpus is used to acquire statistical patterns
  - A test corpus is used to measure system performance
  - A development corpus is similar to a test corpus
    - Systems are “tuned” to get better results on the dev corpus
    - Test corpora are only used infrequently to insure accuracy/fairness
      - The system should not be tuned to get better results
- More annotated text often yield better results
- Different genres may have different properties
  - Systems can “train” separately on different genres
  - Systems can “train” on one diverse corpus

Why do We Care About Levels?

- Different phenomena require different levels
- Example: ChatGPT generates sequences of words/characters based on previously seen sequences of words/characters
- Possible case names are predicted for legal documents
  - Jones vs. Smith; Roe vs. Smith, Wade vs. Jones, etc.
- Generated case names may not be actual case names
  - https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/
- For the arguments to be valid
  - Fake cases are not appropriate
  - Cited Cases should support specific claims

Syllabus: Subset of these Topics

- Introduction (today)
- Formal Languages and Transducers
- Corpus Annotation
- English Syntax and Parsing
- POS Tagging and Hidden Markov Models
- Named Entities and Machine Learning
- Lexical Semantics and Semantic Role Labeling
- Information Extraction: Entities, Relations, Events, Time
- Anaphora and Coreference Resolution
- Feature Structures and Representing Multiple Phenomena
- Machine Translation

Summary

- Computational Linguistics is an applied discipline with an increasingly large inventory of applications.
- A wide variety of levels of analysis are used to implement these applications.
  - Many, but not all of these levels are derived from or inspired by theoretical linguistics
- One popular paradigm for producing an analysis automatically involves manually annotating text

Homework and Readings

- Readings and Self-Study
  - Chapter 1 in Jurafsky and Martin
  - Install NLTK, Read Chapter 1 and follow examples
  - Optional: Read through the full Penn Treebank Part of Speech tagset description: https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf
- Homework Assignment 1:
  - https://cs.nyu.edu/courses/fall23/CSCI-UA.0480-057/homework1.html