Introduction to Computational Linguistics: Problems and Methods

Adam Meyers
New York University
1/23/2012
Outline

• Grades, Exams, Policies, etc.
• Text Books and Suggested Reading
• Defining the Field
• CL Applications
• Types of Text Analysis used in CL
• A Practice Manual Annotation Task
• Summary and Syllabus
• Homework No. 1
Grades, Homework, Exams, Final Projects

• Grade Breakdown:
  – 1/3 Homework + 1/3 Final Exam + 1/3 Final Project

• Homework
  – Homework due every 1 to 1.5 weeks (10 total)
  – Submit Homework via Blackboard (nyu.home.edu/academics)

• Final Project
  – Due May 7, 2012 (last class)
  – Sample topics on the website by March 1

• Final Exam
  – May 14, 2012, 7:10-9PM
  – Sample Exam and Specifics on website by April 1.
Policies

• Late/Missing Homework
  – Natural Consequence: You fall behind and have trouble keeping up, leading to lower mark on final exam and final project
  – Absolution (no penalty):
    • One missing homework per term
    • The grader will mark at most 2 of your homeworks/week
      – If there are homeworks left over at end of term, that could be problematic
  – Other Late Homeworks
    • Depends on how busy the grader is
    • Do not let it pile up: there are 33 people in the class
  – Final Homework Deadline: May 14

• Intellectual Integrity:
  – http://www.cs.nyu.edu/webapps/content/academic/undergrad/academic_integrity
  – You may discuss work with anyone (including Ang and myself), but your work should be your own.
Contact Info

• Adam Meyers
  – 719 Broadway, Room 702
  – meyers@cs.nyu.edu
  – 212 998 3482
  – http://nlp.cs.nyu.edu/people/meyers.html
  – Office Hours: Thursday 11:00AM—12:00PM or by appointment

• Ang Sun
  – 715 Broadway, Room 709
  – asun@cs.nyu.edu
  – 212 998 3491
  – http://cs.nyu.edu/~asun/
  – Office Hours: By Appointment

• Class Website: http://cs.nyu.edu/courses/spring12/CSCI-GA.2590-001/
• Class Mailing List: csci_ga_2590_001_sp12@cs.nyu.edu
  – http://www.cs.nyu.edu/mailman/listinfo/csci_ga_2590_001_sp12
Text Books

• SPEECH and LANGUAGE PROCESSING 2nd Edition
  – By Daniel Jurafsky and James H. Martin
  – http://www.cs.colorado.edu/~martin/slp.html
  – Great Overview of the Field, explanations of techniques, algorithms, etc.

• 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 it)
  – Downloadable open source programs to try out various computational linguistics tools and inspect their code
  – Written in Python 2
More Stuff to Read/Download, etc.

• Look at projects currently going on at NYU:
  – The Proteus website: http://nlp.cs.nyu.edu/
  – My website: http://nlp.cs.nyu.edu/people/meyers.html
    • GLARF: processing tool written in Common Lisp (for linux, but will soon be available for MAC)
    • NomBank: annotation project
    • COMLEX, NOMLEX: lexicon projects

• Other useful links (I will put more on a website):
  – http://cs.nyu.edu/courses/spring10/G22.2590-001/index.html
  – http://aclweb.org/
  – http://www.cs.vassar.edu/sigann/
Defining Computational Linguistics

• AKA, Natural Language Processing (NLP), Language Engineering, ...

• Definition: Study of how to solve 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
    • But sometimes draws on one or both
  – Broad and changing domain influenced by available funding
CL Applications: Slide 1

• Machine translation
  – Methods are not at all based on how humans translate
  – Effective for gisting text, generating 1\textsuperscript{st} draft translations, but not for high-level translation
  – Works better for “controlled languages” – technical manuals (Microsoft, Catterpillar, etc.)

• Spoken Language
  – dictation (IBM ViaVoice, Dragon Naturally Speaking)
  – Telephone-based customer support (phone mazes)

• Information Retrieval (not like in the movie \textit{Brazil})
  – Web Searches (mostly statistics)
CL Applications Slide 2

- Information Extraction
  - Dealtime, Google Products, Monster.com (job search)
  - Some open source tools:
    http://cs.nyu.edu/courses/spring10/G22.2590-001/index.html
    http://alias-i.com/lingpipe/
  - NYU
    - Some tools on website
    - 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://newsblaster.cs.columbia.edu/

- Summarization, Spelling/Grammar Checking, etc.
Types of Analysis

• Phonetics/Phonology: speech recognition and speech synthesis (not in this class)
  – **We will focus on text analysis**
  – Text does not represent some phonological features
  – Text has punctuation

• Syntactic/Semantic: sentence splitting, tokenization, part-of-speech tagging, chunking, parsing, predicate/argument structure, sense disambiguation

• Discourse: anaphora, discourse argument structure, sentiment analysis

• Other: multi-lingual processing (including MT), summarization, IE, etc.
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

• 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
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>`

• Chunking -
  – mark verb groups and/or noun groups, convenient approximations of syntactic units (questionable theoretically).
  – `[NG The book] with [NG the blue cover] [VG was falling] off [NG the shelf].`
Parsing

• \((S \ (NP \ (DT \ the) \ (NN \ book))\)\n  \((VP \ (VBZ \ is))\)
    \((PP \ (IN \ on))\)
      \((NP \ (DT \ the) \ (NN \ shelf)))))\)
Predicate/Argument Structure

• For thousands of years, linguists have employed systems to characterize predictable paraphrases, e.g., Pāṇini, a Sanskrit linguist from the 4th Century BC

• In 21st Century CL, semantic role labeling is popular

They were eaten by a giant clam

John took a walk to the store
Sense Disambiguation

• For interesting characterizations of word senses (and relation between senses), use WordNet (online or download it)
  – wordnet.princeton.edu/

• 2 obviously distinct senses of bank
  – They took money out of the bank.
  – The water flooded over the bank of the river.

• Difficult sense disambiguation
  – Senses 2, 3 and 5 on the next slide are arguably not distinct
  – Lexicographers are acutely aware of the merging vs. splitting problem of enumerating senses
  – CL systems usually collapse some WordNet distinctions
WordNet Noun entry for *table*

1. (52) table, tabular array -- (a set of data arranged in rows and columns; "see table 1")

2. (25) table -- (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs; "it was a sturdy table")

3. (5) table -- (a piece of furniture with tableware for a meal laid out on it; "I reserved a table at my favorite restaurant")

4. mesa, table -- (flat tableland with steep edges; "the tribe was relatively safe on the mesa but they had to descend into the valley for water")

5. table -- (a company of people assembled at a table for a meal or game; "he entertained the whole table with his witty remarks")

6. board, table -- (food or meals in general; "she sets a fine table"; "room and board")
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.
  – **The amusement park** is very dangerous. **The gate** has sharp edges. **The rides** have not been inspected for years.
  – **This book** is valuable, but **the other book** is not.
Discourse Argument Structure

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

They wanted to steal the diamonds.

ARG1

However, they did not possess the necessary skills.

ARG2
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:
    • \[ Kappa = \frac{\text{Percent (Actual Agreement)} - \text{Prob (Chance Agreement)}}{1 - \text{Prob (Chance Agreement)}} \]

• Used to create answer keys to score system output
  – One set of measures are: recall, precision and F-score
  
  \[ \text{Recall} = \frac{|\text{Correct}|}{|\text{Answer Key}|} \quad \text{Precision} = \frac{|\text{Correct}|}{|\text{System Output}|} \quad F - \text{Score} = \frac{1}{\frac{1}{2} \left( \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \right)} \]
Role of Manual Annotation in Supervised Statistical CL

• 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 used infrequently and 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
A Short Hypothetical Task

• Jargon:
  – a word or sequence of words forming a unit
  – specific to a technical sublanguage: higher frequency, special usage, special meaning, etc.

• Let's each mark all the instances of jargon on the following slide – please put square brackets around these phrases in your hand out.
Tuber induction in potato (Solanum tuberosum L.) is a complex, multilevel process, which integrates environmental and internal signals to ensure optimal life strategy during the growing season (reviewed in [34, 37]). Environmental requirements for tuberization vary among potato subspecies and varieties. S. tuberosum ssp. andigena strictly requires short-day photoperiod for tuber formation, however, andigena plants with inactivated phytochrome B gene (involved in photoperiod sensing) tuberize even under continuous light [17].
Hypothetical Task Slide 2

• We now should make a list of inconsistencies
• For each inconsistency, we must make a decision of how these should be marked in the future
• Sometimes, it is necessary to do many rounds of annotation of a single text by multiple people to develop good specifications.
• Remember, the goal is to create an annotation that:
  – Characterizes some phenomenon well
  – Is possible to reproduce by automatic means
Summary and What's Next

- 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
- In the next lecture we will introduce some models of the patterns these produce and programs for identifying these patterns
Syllabus

• Introduction
• Formal Languages and Transducers
• Natural Language 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: Coreference and Similar Phenomena
• Feature Structures and Representing Multiple Phenomena
• Machine Translation
Homework and Readings

• Jurafsky and Martin, Chapter 1
• NLTK Book – Install NLTK, read Chapter 1 and follow along with their examples.
• Find about 200 words of news text on the web and identify the adjectives by inserting a “/JJ” after each adjective.
• Use the following criteria, but please make notes about any problems with these specifications:
  – An adjective must be able to fill in the blanks
    • The ___ noun ... --- occur between “the” and a noun
    • The noun is ____ --- play the role of a predicate after “is”
  – An adjective CANNOT (comfortably) fill the following blank
    • The ___ is --- be the subject of “is”
    • ___s --- become a plural
• NLTK: Chapter 2 (Optional)