CS6320
Natural Language Processing

Lecture 1: Introduction

Yang Liu

Acknowledgement: some slides from Jason Eisner, Mary Harper, Bonnie Dorr

Language ambiguity

- What has four wheels and flies?
- Time flies like an arrow. Fruit flies like a banana.

Course Info

- Course webpage:
  - http://www.hlt.utdallas.edu/~yangl/cs6320
  - announcement, homework, slides, etc.
  - Will also use elearning
- Office hour (ECSS 3.402)
  - T/Th 2:15-3pm
- TA (TBA)

- Prerequisite:
  - CS5343: Algorithm analysis and data structure

Textbooks

- J&M: Speech and Language Processing
- M&S: Foundations of Statistical Language Processing (recommended)
  - Available online, and other useful material

on reserve in library

- Others:
  - E. Charniak, Statistical Language Learning
  - F. Jelinek, Statistical Methods for Speech Recognition

Conferences/Journals

- Many papers online at ACL anthology
  - http://aclweb.org/anthology-new/
- Conferences:
  - ACL, NAACL, EMNLP, AAAI, SIGIR, DUC/TAC, …
- Journals:
  - Computational Linguistics, Natural Language Engineering, Computer Speech and Language, Transactions of Association for Computational Linguistics, …

Resources

- LDC
  - Large corpora of text and speech, with various annotation
- A lot of data and tools online

Logistics (subject to change)

- 1 midterm (20%)
- 5 homework (45%) (tentative)
  - Intro, regex, sentence segmentation, word tokenization
  - N-gram LM
  - POS tagging
  - text categorization
  - information extraction or others
- 1 final project/exam (30%). Possibly a quiz or final exam.
- Participation in class (5%)
- Letter grade, A, A-, B+, B, B-, C+, C, C-

Suggestions

- Get hands-on with data
- Understand theoretical foundations
- Often practical issues are even more important than theoretical niceties

NLP Introduction

Goals of the Field

Computers would be a lot more useful if they could handle our emails, talk to us, identify relevant information online, recommend good products, …

But they are fazed by natural human language.

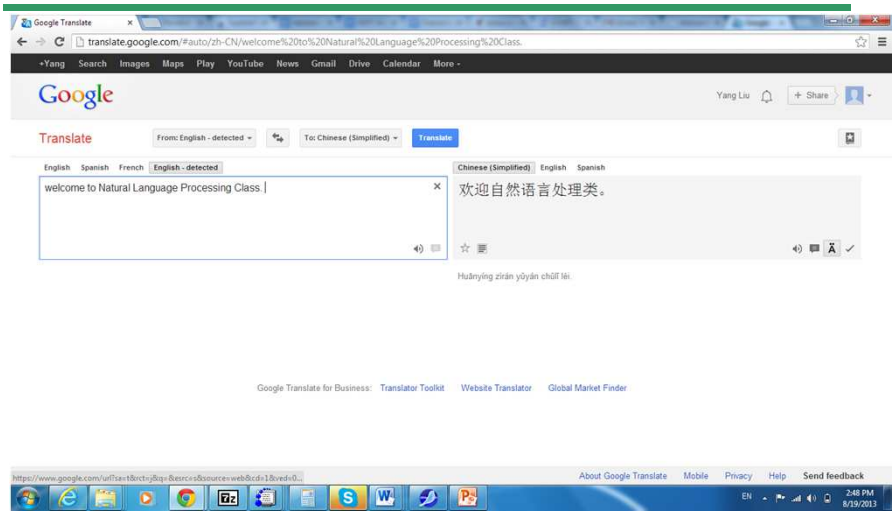
How can we tell computers about language? (Or help them learn it as kids do?)

Why Is NLP Important?

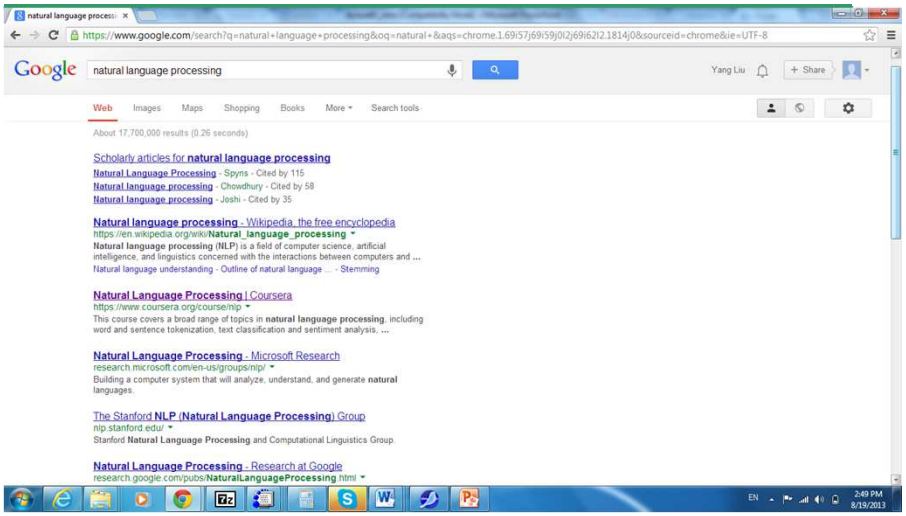
- Easy for everyone to use language
  - Natural human interface for a variety of applications
- Is getting more important to deal with information overload nowadays
  - Google indexed tens of billions of webpages
  - 400 millions tweets per day (March 2013)

NLP Applications/tasks:
Question Answering (e.g., IBM Watson)

NLP Applications/tasks: Machine Translation



NLP Applications/tasks: Information Retrieval



NLP Applications/tasks: Spell Check

NLP Applications/tasks: Text Categorization

- Topic classification
- Sentiment analysis
- Authorship attribution
- Spam filtering

Sentiment Analysis

- Tell me a movie that is more famous than this. Tell me a movie that has had more parodies spinned off its storyline than this. Tell me one movie that has been as quoted as a much as this. The answer is you can't. No movie has had as much of an impact as The Godfather has had ever since it was released.
  The acting was simply amazing, what else could you say.

- First things first: disregard the rating above because they don't allow for rating low enough for this movie. I would honestly say 0 for real numbers although negative would be more appropriate since I left the theater feeling worse than when i entered.
  Next, I have to preface this by saying I only watched the first half of the movie. I'll also add that I have never walked out of a movie theater and i would have walked out far earlier if I were by myself. Luckily i saw it for free so those idiots who made it will get no profit from me.

Spam Filtering

- Dear User I am Jackson, Beth from customer service department. Your mailbox has exceeded its Web limit for this reason it will be very slow when sending massages, With time your mail box may not be able to send or receive new e-mails.login to our services system, click here<http://securemailpage.webs.com/> to enable us reset the size and speed of your mail box when sending messages.

NLP Applications/tasks

- Automatic scoring
- Summarization (e.g., http://freesummarizer.com/)
- Spoken dialog systems (Siri, etc.)

and many more ….

Goals of This Course

- Introduce you to NLP problems & solutions
- Relation to linguistics & statistics

- At the end you should:
  - Agree that language is subtle & interesting
  - Feel some ownership over the formal & statistical models
  - Understand research papers in the field

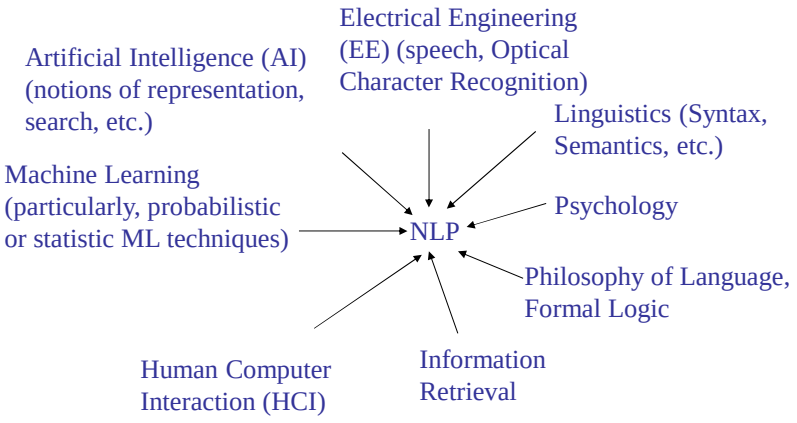
Course Topics

- Some linguistics, regular expression
- Probability and information theory
- Language modeling
- Hidden Markov model
- Part-of-speech tagging
- Syntactic parsing
- Machine learning methods
- Lexical semantics
- Some applications
  - Text categorization, information extraction, speech recognition, discourse segmentation, machine translation, information retrieval, summarization

Disclaimer

- Cannot cover all the details or state-of-the-art systems
- You will need to read research papers

Relation of NLP to Other Disciplines



A Sampling of "Other Disciplines"

- Linguistics: formal grammars, abstract characterization of what is to be learned.
- Computer Science: algorithms for efficient learning or online deployment of these systems.
- Engineering: stochastic techniques for characterizing regular patterns for learning and ambiguity resolution.
- Psychology: insights into what linguistic constructions are easy or difficult for people to learn or to use.

NLP Issues

- Why is NLP difficult?
  - Many "words", many "phenomena", many "rules"
    - OED: 400k words; Finnish lexicon (of forms): ~2 . 10^7
    - sentences, clauses, phrases, constituents, coordination, negation, imperatives/questions, inflections, parts of speech, pronunciation, topic/focus, and much more!
    - irregularity (exceptions, exceptions to the exceptions)
      - potato → potato es (tomato, hero,...); photo → photo s, and even: both mango → mango s or → mango es

Difficulties in NLP (cont.)

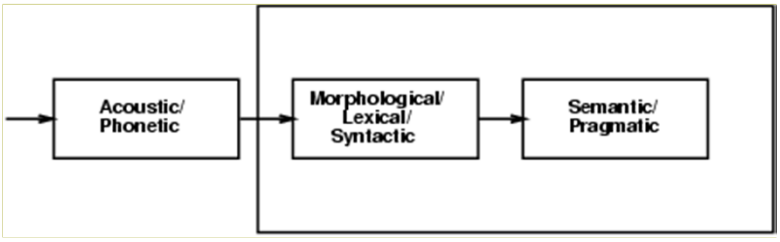
- Ambiguity in language
  - books: NOUN **or** VERB?
    - you **need** many books *vs.* she **books** her flights online
  - Thank you for not smoking, drinking, eating or playing radios without earphones. ***(MTA bus)***
    - Thank you for not eating without earphones??
    - Thank you for drinking?? …
  - Fred's hat was blown off by the wind. He tried to catch it.
    - ...catch the wind *or* ...catch the hat ?

More Examples of Ambiguity

- Iraqi Head Seeks Arms
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Local HS Dropouts Cut in Half
- Hospitals Are Sued by 7 Foot Doctors
- Fred saw the dog with his binoculars.
- I saw the Golden Gate Bridge flying into San Francisco.
- Every man saw the boy with his binoculars
- I made her duck

Knowledge Sources

- NLP is different from other data processing in its use of knowledge about language
- Humans use a lot of knowledge sources to understand or disambiguate language



Types of Knowledge (levels of language)

- **Acoustic/Phonetic Knowledge:** How words are related to their sounds.
- **Morphological Knowledge:** How words are constructed out of basic meaning units.
  - un + friend + ly → unfriendly
  - love + past tense → loved
  - object + oriented → object-oriented
  - Could have spelling changes: dropping, flies

More Types of Knowledge

- **Lexical Knowledge (or Dictionary):** This should include information on parts of speech, features (e.g., number, case), typical usage, and word meaning.
- **Syntactic Knowledge:** How words are put together to make legal sentences (or constituents of sentences).
  - Sentence -> subject verb object

More Types of Knowledge

- **Semantic Knowledge:** Word meanings, how words combine into sentence meaning. Combining words into a sentence affects sentence and word meaning.

  Examples:

  Fred broke the window with the block.

  Fred broke the window with Mary.

More Types of Knowledge

- **Pragmatic:** How do you conclude from what I said? How do you react?
- **Discourse:** structure of language, turn taking
- **World knowledge**: How does your mind work? what do you know or believe?
  - **Examples:**
    - Will you pass the salt?
    - I'm sorry. I'm afraid I can't do that.
    - I read an article about the war in the paper.
    - Fred saw the bird with his binoculars.
    - Tim was invited to Tom's birthday party. He went to the store to buy him a present.

Types of Ambiguity

- **Lexical:**
  - you **need** many books *vs.* she **books** her flights online
- **Syntactic:**
  - Fred saw the bird in the nest with the binoculars.
- **Semantic:**
  - Thank you for not smoking, drinking, eating or playing radios without earphones
  - Fred's hat was blown off by the wind. He tried to catch it.

Rules or Statistics to Disambigute?

- Rules
  - context clues: she books → books is a verb
    - rule: if an ambiguous word (verb/nonverb) is preceded by a matching personal pronoun → word is a verb
  - pronoun reference:
    - she/he/it often refers to the most recent noun or pronoun (but there are certainly exceptions)
  - semantics:
    - We thank people for doing helpful things or not doing annoying things

Statistical NLP

- Learn a statistical model from real data
- Simple example for now: N-gram LM for word/character prediction
  - Letter or word frequencies: 1-grams
  - If you know the previous letter: 2-grams
    - "h" is rare in English (4%)
    - but "h" is common after "t" (20%)
  - If you know the previous 2 letters: 3-grams
    - "h" is really common after "(space) t"

NLP History: 1940-1950's

- Development of formal language theory (Chomsky, Kleene, Backus)
  - Formal characterization of classes of grammar (context-free, regular)
  - Association with relevant automata
- Probability theory: language understanding as decoding through noisy channel (Shannon)
  - Use of information theoretic concepts like entropy to measure success of language models

1957-1983
Symbolic vs. Stochastic

- Symbolic
  - Use of formal grammars as basis for natural language processing and learning systems. (Chomsky, Harris)
  - Use of logic and logic based programming for characterizing syntactic or semantic inference (Kaplan, Kay, Pereira)
- Stochastic Modeling
  - Probabilistic methods for early speech recognition, OCR (Bledsoe and Browning, Jelinek, Black, Mercer)

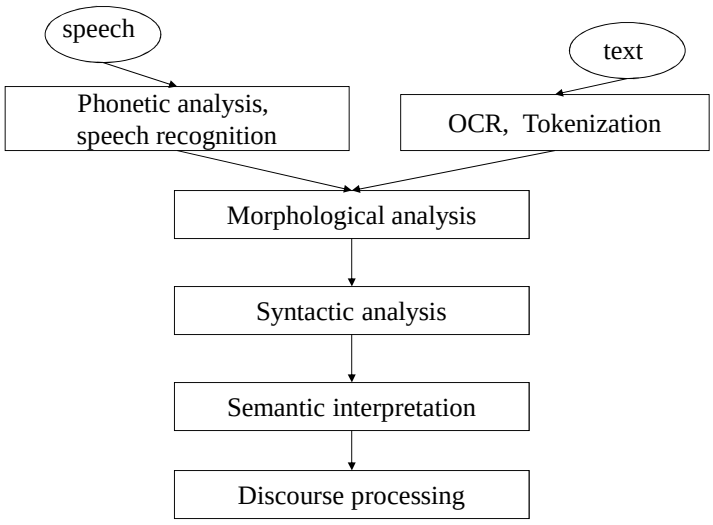
1983-1993:
Return of Empiricism

- Use of stochastic techniques for part of speech tagging, parsing, word sense disambiguation, etc.

1993-Present

- Advances in software and hardware create NLP needs for information retrieval (web), machine translation, spelling and grammar checking, speech recognition and synthesis, summarization, sentiment analysis, etc.
- Stochastic and symbolic methods combine for real world applications.
- Rise of machine learning in NLP
- See J&M for more history in the field

NLP Pipeline



Tokenization

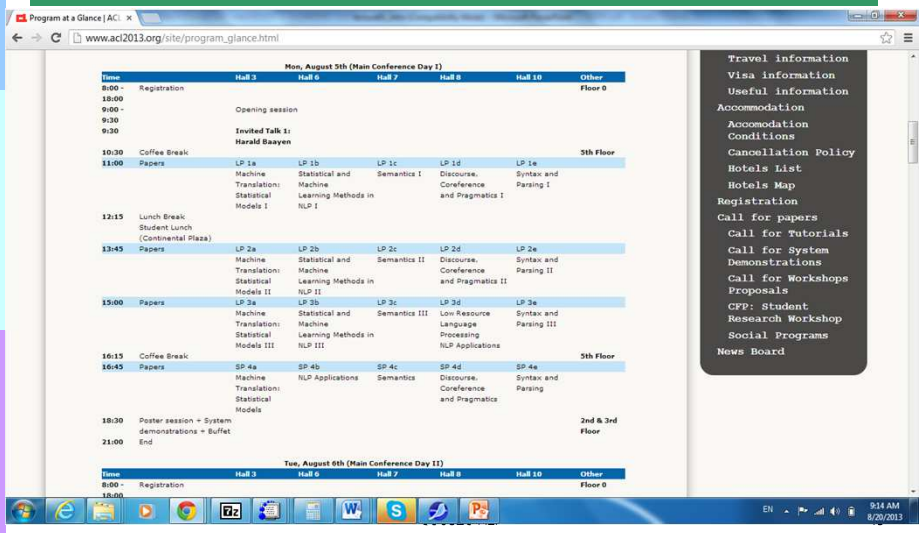
- Split text into sentences and words
  - Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing.
- Need to deal with punctuation, apostrophe, hyphen, email, url, etc.
- What is a token? How to handle compound words?
- Is language specific, some languages don't have space as word delimiters

Things for you to do

- Ambiguity: examples of language ambiguity
  - Either email me, or present in class
- Homework 1 is out

ACL 2013

(http://www.acl2013.org/site/program_glance.html)



Mon, August 5th (Main Conference Day I)

Time		Hall 3	Hall 6	Hall 7	Hall 8	Hall 10	Other
8:00 - 18:00	Registration						Floor 0
9:00 - 9:30		Opening session					
9:30		**Invited Talk 1: Harald Baayen**					
10:30	Coffee Break						5th Floor
11:00	Papers	LP 1a Machine Translation: Statistical Models I	LP 1b Statistical and Machine Learning Methods in NLP I	LP 1c Semantics I	LP 1d Discourse, Coreference and Pragmatics I	LP 1e Syntax and Parsing I	
12:15	Lunch Break Student Lunch (Continental Plaza)						
13:45	Papers	LP 2a Machine Translation: Statistical Models II	LP 2b Statistical and Machine Learning Methods in NLP II	LP 2c Semantics II	LP 2d Discourse, Coreference and Pragmatics II	LP 2e Syntax and Parsing II	
15:00	Papers	LP 3a Machine Translation: Statistical Models III	LP 3b Statistical and Machine Learning Methods in NLP III	LP 3c Semantics III	LP 3d Low Resource Language Processing NLP Applications	LP 3e Syntax and Parsing III	
16:15	Coffee Break						5th Floor
16:45	Papers	SP 4a Machine Translation: Statistical Models	SP 4b NLP Applications	SP 4c Semantics	SP 4d Discourse, Coreference and Pragmatics	SP 4e Syntax and Parsing	
18:30	Poster session + System demonstrations + Buffet						2nd & 3rd Floor
21:00	End						

Tue, August 6th (Main Conference Day II)

Time		Hall 3	Hall 6	Hall 7	Hall 8	Hall 10	Other
8:00 - 18:00	Registration						Floor 0