Introduction to NLP

- Instructor: Claire Cardie
  - Professor in CS and IS (and CogSci)

Computationally oriented introduction to natural language processing, the goal of which is to enable computers to use human languages as input, output, or both. Possible topics include parsing, grammar induction, information retrieval, and machine translation.

Natural Language Processing (NLP)

- "Natural" language
  - Languages that people use to communicate

- Ultimate goal
  - To create computational models that perform as well at using natural language as humans do

- Immediate goal
  - To build computer systems that can process text and speech more intelligently

![Diagram of NL input and output]

Information retrieval

- Web search engines

![Diagram of information retrieval process]

Query: (articles on) leveraged buyouts

- Query: (articles on) leveraged buyouts involving more than 100 million dollars that were attempted but failed during 1986 and 1990

- I see what I eat = I eat what I see
  [Mad Hatter, Alice in Wonderland]
Question answering (QA)

- Task
  - How many calories are there in a Big Mac?
  - Who is the voice of Miss Piggy?
  - Who was the first American in space?
  - Retrieve not just relevant documents, but return the answer

Machine translation

- one of the first applications envisioned for NLP techniques
  - The spirit is willing, but the flesh is weak.
  - “open”
  - Certainly see a need for it...
    - The extension of the coverage of the health services to the underserved or not served population of the countries of the region was the central goal of the Ten-Year Plan and probably that of greater scope and transcendence.
    - Welcome to Chinese Restaurant. Please try your Nice chinese Food With chopsticks. the traditional and typical of Chinese glorious history and cultural. PRODUCT OF CHINA

Dialogue-based systems

- Assistant: Can I help you?
- Customer: I was wondering whether you have any switched brass lampholders.
- Assistant: The brass lampholders are out of stock, but they should be in on Wednesday. The plastic ones are over here...

Why is dealing with NL hard?

Ambiguity!!!! …at all levels of analysis 😒

- Phonetics and phonology
  - Concerns how words are related to the sounds that realize them. Important for speech-based systems.
    - "I scream" vs. "ice cream"
    - "nominal egg"
    - "It's very hard to recognize speech." vs. "It's very hard to wreck a nice beach."
- Morphology
  - Concerns how words are constructed from morphemes.
    - "unionized"
      - union + ized
      - un + ion + ized
Why is dealing with NL hard?

Ambiguity!!!! …at all levels of analysis 😞

- Syntax
  - Concerns sentence structure
  - Different syntactic structure implies different interpretation
    » Squad helps dog bite victim.
      ♦ [np squad] [vp helps [np dog] [inf-clause bite victim]]
    » Helicopter powered by human flies.

- Semantics
  - Concerns what words mean and how these meanings combine to form sentence meanings.
    » Red-hot star to wed astronomer.
    » The once-sagging cloth diaper industry was saved by full dumps.

- Discourse
  - Concerns how the immediately preceding sentences affect the interpretation of the next sentence
    » Jack drank the wine on the table. *It* was brown and round.
    » Jack saw Sam at the party. *He* went back to the bar to get another drink.
    » Jack saw Sam at the party. *He* clearly had drunk too much.

[Adapted from Wilks (1975)]
What topics will we cover?

- Language modeling
- Phonetic analysis
- Morphological analysis
- Lexical semantics and word-sense disambiguation
- Part-of-speech tagging
- Parsing
- Grammar induction
- Semantic analysis
- Pronoun resolution
- Coreference analysis
- NL Generation
- Machine translation
- Dialogue systems
- Information extraction

Reference Material

- Required text book:

- Other useful references:
  - Others listed on course web page...

Prereqs, Coursework, & Grading

- Prerequisites
  - CS 2110.

- Grading
  - 15%: critiques of selected readings and research papers
  - 10%: the daily quiz
  - 65%: programming projects and reports
  - 8%: participation
  - You'll be expected to participate in class discussion and class exercises or otherwise demonstrate an interest in the material studied in the course.
  - 2%: course evaluation completion

www.cs.cornell.edu/courses/cs4740/20010sp/