Introduction to NLP

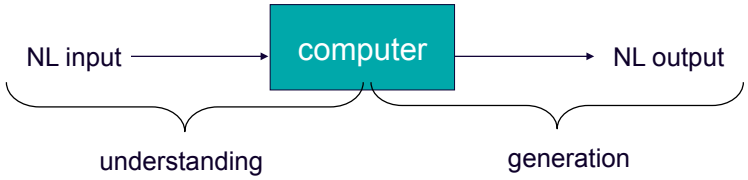
CS 4740 / CS 5740 / LING 4474 / COGST 4740

- 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.

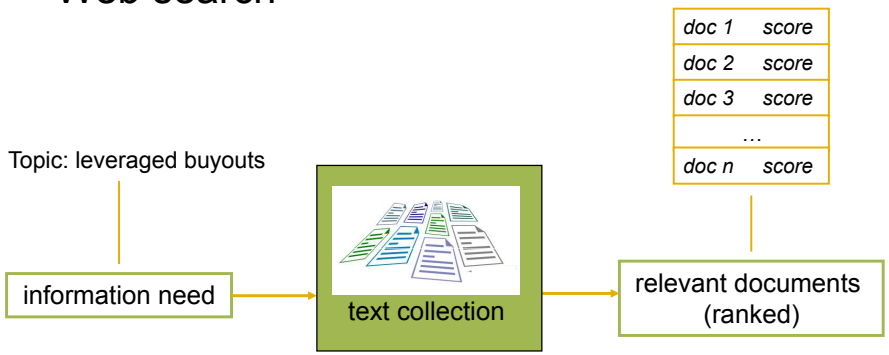
Natural Language Processing (NLP)

- "Natural" language
  - Languages that people use to communicate with one another
- 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



Information retrieval

- Ad-hoc IR
- Web search



IR system

Information retrieval

- 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

IBM's Watson

http://www-03.ibm.com/innovation/us/watson/what-is-watson/why-jeopardy.html

Machine translation

- one of the first applications envisioned for NLP techniques
  - – *The spirit is willing, but the flesh is weak.*
  - – "open"

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."

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 bite victim$]]$
      - ◆ $[_{np}$ squad$]$ $[_{vp}$ helps $[_{np}$ dog$]$ $[_{inf\text{-}clause}$ bite victim$]]$
    - » Helicopter powered by human flies.

Why is dealing with NL hard?

Ambiguity!!!! …at **all** levels of analysis ☹

- 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.

Why is dealing with NL hard?

Ambiguity!!!! …at **all** levels of analysis ☹

- 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)]

Why is dealing with NL hard?

Ambiguity!!!! …at **all** levels of analysis ☹

- Pragmatics
  - Concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

    "I just came from Collegetown Bagels."

    » Do you want to go to Collegetown Bagels?
    » Do you want to go to Gimme Coffee?
    » Boy, you look tired.

What topics can we cover?

Language modeling
Phonetic analysis
Morphological analysis
Word-sense disambiguation
Part-of-speech tagging
Parsing
Grammar induction
Semantic analysis
Pronoun resolution
Coreference analysis
NL Generation
Machine translation
Dialogue systems
Information extraction
QA systems
Topic models

Reference Material

- Required text book:
  - Jurafsky and Martin, *Speech and Language Processing*, Prentice-Hall, **2nd edition**.
- Other useful references:
  - Manning and Schutze. *Foundations of Statistical NLP*, MIT Press, 1999.
  - Others listed on course web page…

Prereqs, Coursework, & Grading

- Prerequisites
  - CS 2110.
- Grading
  - 75%: four programming projects with short (5-6pg) reports
  - 15%: critiques of selected readings and research papers
  - 9%: 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.
  - 1%: course evaluation completion

http://www.cs.cornell.edu/courses/cs4740/