CS 6740 / INFO 6300
Advanced Language Technologies

Graduate-level introduction to technologies for the computational treatment of information in human-language form,

covering natural-language processing (NLP) and/or information retrieval (IR).

Possible topics include text categorization and clustering, information extraction, latent semantic analysis (LSI), click-through data for web search, language modeling, computational syntactic and semantic formalisms, grammar induction, and machine translation.

Natural Language Processing (NLP)

- “Natural” language
  - Languages that people use to communicate

- Ultimate goal
  - To build computer systems 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 (IR)

- Ad-hoc IR

Topic: Advantages and disadvantages of using potassium hydroxide in any aspect of organic farming, especially...
Text categorization

Is the document about politics? sports? economics? fashion?

Sentiment categorization

Is the document positive? negative?

Summarization

[White et al., 2002]

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 + supporting text
Machine translation

- MT systems would clearly facilitate human-human communication
  - 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 cultual. PRODUCT OF CHINA

Bill Gates, 1997 “...now we’re betting the company on these natural interface technologies”

Dialogue systems

- Require both understanding and generation
  - Dave: Open the pod bay doors, HAL.
  - HAL: I'm sorry Dave, I'm afraid I can't do that.
  - Dave: What's the problem?
  - HAL: I think you know what the problem is just as well as I do.

Umm…there WILL be complications

- NLP
  - AI-complete
    - To “solve” NLP, you’d need to solve all of the problems in AI
  - Turing test
    - Posits that engaging effectively in linguistic behavior is a sufficient condition for having achieved intelligence.

...But little kids can “do” NLP…
  - Why is understanding language hard?

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]]
      - [exp squad] [vp helps [exp dog] [inf-clause bite victim]]
  - Visiting relatives can be trying.
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.
    » We steered the sailboat away from the bank.
    » River bank vs. $$$ bank?
    » Visiting relatives can be trying.
    » Visiting museums can be trying.
      » Same set of possible syntactic structures for this sentence
      » But the meaning of museums makes only one of them plausible

- Discourse
  - Concerns how the immediately preceding sentences affect the interpretation of the next sentence
    » Merck & Co. formed a joint venture with Ache Group, of Brazil. It will be called Prodome Ltd.
    » Merck & Co. formed a joint venture with Ache Group, of Brazil. It will own 50% of the new company to be called Prodome Ltd.
    » Merck & Co. formed a joint venture with Ache Group, of Brazil. It had previously teamed up with Merck in two unsuccessful pharmaceutical ventures.

What topics will we cover?

- Language modeling
- Lexical semantics and word-sense disambiguation
- Part-of-speech tagging and HMMs
- Parsing
- Semantic analysis
- Discourse processing
- Coreference analysis
- NL Generation
- Machine translation
- Information extraction
- Information retrieval models
- Sentiment analysis
- Text categorization
- Question answering
- Summarization

GOAL: learn to evaluate and execute research in NLP/IR.
Reference Material

- Required text book:

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

Prereqs, Coursework, & Grading

- Prerequisites
  - Permission of instructor. Neither INFO/CS4300 nor CS4740 are prerequisites.

- Grading
  - 25%: participation
    - You'll be expected to participate in class discussion and class exercises.
  - 40%: semester project
  - 33%: presentation of readings and research papers
  - 2%: course evaluation completion