Thought

“Language is an anonymous, collective and unconscious art; the result of the creativity of thousands of generations.”

Edward Sapir
anthropologist and linguist
(1884-1939)

Thought

“Language is a city to the building of which every human being brought a stone.”

Ralph Waldo Emerson
writer and philosopher
(1803-1882)

Quick Course Info.

Office Hours:
- TBD
- OR: By appointment

Communication
- Wiki: http://wiki.cs.byu.edu/cs-679/
  - Syllabus and Schedule will be located there
  - Accessible via Learning Suite
- Email: ringger AT cs.byu.edu
- Announcements and Discussion: on our CS 679 Google Group (see the wiki for a link)

Assignments
- Readings:
  - Optional: Manning & Schuetze (in bookstore, etc.)
  - Optional: Jurafsky & Martin (on reserve soon)
  - Papers (see course web page for hyperlinks)
- Weekly Reading Reports:
  - % completed
  - Your questions
  - Projects
  - Final Project
- Policies
The Dream

It would be great if machines could ... But they can’t because ...

What is NLP?

What is nearby NLP?

What is this Class?

Class Requirements

- Class requirements
  - Probability theory and some graphical models
    - CS 401R (future 477?), CS 677
  - Basic linguistics background or experience with another language
  - Good coding skills & ability with Java

Learning Outcomes

Mathematical Models
- The student will grow in confidence in their mathematical and statistical abilities. In particular, the student will understand the models, methods, and algorithms of statistical Natural Language Processing (NLP) for common NLP tasks, such as speech recognition, machine translation, spam filtering, text classification, and spell checking.
- The student will be able to implement probabilistic models in code, estimate parameters for such models, and run meaningful experiments to validate such models.

Algorithms
- The student will apply core computer science concepts and algorithms, such as dynamic programming.

Linguistics
- The student will gain understanding of linguistic phenomena and will explore the linguistic features relevant to each NLP task.
Learning Outcomes (2)

**Generalization**
- The student will apply the methods to new NLP problems and will be able to apply the methods to problems outside NLP.

**Analysis**
- The student will be familiar with some of the NLP literature and will read and suggest improvements to published work.
- The student will see where opportunities for research await and prepare to conduct research in NLP or related fields.
- The student will also analyze experimental results, write reports for each course project to develop scientific writing skills.

Language is Ambiguous

- **Headlines:**
  - Iraqi Head Seeks Arms
  - Ban on Dancing on Governor’s Desk
  - Juvenile Court to Try Shooting Defendant
  - Teacher Strikes Idle Kids
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
  - British Left Waffles on Falkland Islands
  - Clinton Wins on Budget, but More Lies Ahead
  - Hospitals Are Sued by 7 Foot Doctors

- Why are these funny?

Ambiguities Everywhere!

- Maybe we’re sunk on funny headlines, but normal, boring sentences are unambiguous, right?

Fed raises interest rates 0.5% in a measure against inflation

Semantic Ambiguities

- Even correct tree-structured syntactic analyses don’t always nail down the meaning

Every morning someone’s alarm clock wakes me up

John’s boss said he was doing better

More Attachment Ambiguities

Semantic Ambiguities

- Even correct tree-structured syntactic analyses don’t always nail down the meaning

Every morning someone’s alarm clock wakes me up

John’s boss said he was doing better
Semantic Ambiguities

- Even correct tree-structured syntactic analyses don’t always nail down the meaning

*Every morning someone’s alarm clock wakes me up*

John’s boss said he was doing better

Lower: Ambiguities at Word Level

- *sara* andata
- be+fut+3sg go+ppt+fem

Lower: Ambiguities at Word Level

Higher: Ambiguities Beyond the Sentence Level

- **Discourse**: how do sentences relate?
- **Pragmatics**: what intent is expressed by the literal meaning, how to react?

Disambiguation for Applications

- Sometimes life is “easy”
  - Text classification
  - Authorship attribution
  - Web search
- *jaguar habitat*
- Sometimes only certain ambiguities are relevant
  - Text-to-Speech (TTS)
- *he hoped to record a world record*
- Other times, many levels of ambiguity are relevant
  - Translation

Some Early NLP History

- **1950s: Optimism**
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military (imagine that)
    - Toy models: MT using basically word-substitution
- **1960s and 1970s: NLP Winter**
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
  - Work shifts to deeper models, syntax
  - ... but toy domains / grammars (SHRDLU, LUNAR)
History (2)

- 1980s: The Empirical Revolution
  - Expectations get reset
  - Corpus-based methods
  - Deep analysis often traded for robust and simple approximations
  - Evaluate everything

- Late 1990s: The Web

- 2000s:
  - Deluge of Statistical Machine Learning

- 2010s
  - Awesome language apps
  - Unsupervised methods

Assignment

- Reading for Tuesday:
  - Manning & Schuetze ch. 1
  - Kay: CL and NLP

- Submit reading report, typed and printed, at beginning of class

- Homework
  - Discuss scope

What’s Next?

- Why Data?
- Lectures on probability theory and probabilistic models
  - Fundamental ideas we’ll use and re-use!
- Increasingly complex problems and models
  - Dealing with scale and sparsity
  - Sequence tasks (POS tagging, entity recognition)
  - Tree tasks (parsing, semantic interpretation)
  - Alignment (translation)