CS 181: Natural Language Processing

Lecture 1: Introduction

Kim Bruce
Pomona College
Spring 2008

Disclaimer: Slide contents borrowed from many sources on web!

Organization

- Seminar -- looking for lots of discussion, esp. analysis of techniques. Share learning.
- Texts -- on-line preprints.
  - *Speech and Language Processing* by Jurasky & Martin
  - *Natural Language Processing in Python* by Bird et al
- Homework -- written and programming.
  - Groups of 2 or 3 encouraged
  - Up to 3 late days, max 2 at a time
- Programming in Python using nltk library.
- Final project -- using computation.

NLP & Computational Linguistics

- Applications:
  - Web search engines, speech recognition, text to speech synthesizers, document summarizers, machine translation, ...
- Scientific study of language from computational viewpoint.
  - Computational explanation of linguistic phenomena
  - Working programs involving text or speech processing

Computing as an Intellectual Tool

Expressing methodology in a computer language forces it to be unambiguous and computationally effective. The task of formulating a method as a computer-executable program and debugging that program is a powerful exercise in the learning process. The programmer expresses his/her poorly understood or sloppily formulated idea in a precise way, so that it becomes clear what is poorly understood or sloppily formulated. Also, once formalized procedurally, a mathematical idea becomes a tool that can be used directly to compute results.

Gerry Sussman, 2005

Goals of NLP

- Get computers to do more stuff to help us
  - handle e-mail & other communications
  - Find & extract needed information
  - Reason about text
- Computers aren't good at human langs
  - Hard problems
  - Making progress

Course Goals

- Learn basic principles and theoretical underpinnings of NLP.
- Learn to use tools to develop practical & useful robust systems to understand text.
- Gain insight into open research problems.

Two Competing Approaches

- **Empirical**: Statistics, Machine learning, and Stochastic Processes
- **Rationalist**: Logics, Lambda Calculus, Boolean Algebras, and Lattices

What Works?

- Empiricist approach has been very successful
  - Most practical applications based on machine learning (*AI*) and statistics
  - Highly successful on low-level issues, e.g. word recognition, speech-to-text, simple question-answering.
- Now blending with rationalist approach, esp. with semantics.

Why is understanding language hard?

Ambiguity!

- Parts of speech:
  - "Time flies like an arrow."
  - "Fruit flies like a banana."
- Other parsing
  - "I saw the man in the park with the telescope."
- Sometimes it is not parsing at all:
  - "Every student in the class was reading a story."
  - "They enjoyed it/them."

Amusing Headlines

- Iraqi head seeks arms
- Juvenile court to try shooting defendant
- Teacher strikes idle kids
- Kids make nutritious snacks
- British left waffles on Falkland Islands
- Red tape holds up new bridges
- Ban on nude dancing on governor's desk
- Hospitals are sued by *7* foot doctors
- Stolen painting found by tree

More Ambiguity

- "I made her duck"
  - I cooked a bird for her
  - I cooked the bird that belonged to her
  - I created a bird-like object and gave it to her
  - I caused her to lower her head suddenly
  - I magically changed her into a bird-like substance
- How can an automated system handle these distinctions?

Many ways to say same thing

- John ate a frog
- A frog was eaten by John
- It was John who ate the frog
- The one who ate the frog was John
- What was eaten by John was a frog
- What John ate was a frog
- ...

Reference Resolution

A: Where is "Sweeney Todd" playing near Claremont?

B: It is playing at the Edwards Laverne 12

A: When is it playing there?

B: It's playing at 4:20 p.m. and 10:10 p.m.

A: I'd like 2 adult and 2 children for the late show. How much is that?

- Knowledge sources:
  - Domain knowledge
  - Discourse knowledge
  - World knowledge

NLP is hard because:

- Highly ambiguous
- Complex and subtle
- Fuzzy and probabilistic
- Involves reasoning about the world
- A key part of people interaction w/ others
  - Involves persuasion & changing emotions
- But surprisingly, sometimes quite easy

History

- Early Days (1940's - 1950's)
- Two Camps (1957 - 1970)
- Four Paradigms (1970 - 1983)
- Empiricism & Finite State Models (1984 - 1993)
- Together Again (1994 - present)

Early Days (1940's - 1950's)

- Finite State Automata / Regular Languages (Turing, Shannon, Kleene)
- Context-free grammars (Chomsky, Backus, Naur)
- Noisy channel models (Shannon)
- MT basically word substitution
  - Little understanding of syntax, semantics, pragmatics

Two Camps (1957 - 1970)

- Symbolic
  - Generative Transformational Grammar (Harris, Chomsky)
  - Symbolic AI (McCarthy, Minsky, ...)
- Stochastic
  - Emergence of first Computer corpora (Brown corpus) and machine-readable dictionary
  - Speech recognition

Four Paradigms (1970 - 1983)

- Stochastic: speech recognition (IBM, CMU) using HMM, noisy channel models
- Logic-based: Grammars in Prolog (Pereira, Warren), Lexical Functional Grammar (Kay, Bresnan, Kaplan)
- Natural language understanding (focus on meaning)
- Discourse modeling

Empiricism & Finite State Models (1983 -1993)

- Return of finite-state models
- Return of probabilistic approaches
- Natural Language Generation (NLG)

Together Again (1994 - present)

- Statistical Machine translation
- Probabilistic parsing
- Statistical NLG
- Extended use of machine learning
- Information retrieval & text categorization

NLP as AI

- Turing test: Can a computer fool a person into thinking it is human
- What is thinking? What is understanding?
- Eliza (1966)

Layers of Linguistics

- Phonetics & Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Learning Language

- Linguistics rules: How to make plurals
- Memorization: goose -> geese
- Not static: New nouns and verbs added

Probabilistic Models

- Bayesian Classifiers (not rules)
- Hidden Markov Models (not fsa's)
- Probabilistic context free grammars
- ... lots of Machine learning, stats, etc.

Goals of Course

- Intro to NLP problems and solutions
- Focus on probabilistic integration of evidence
- Hands-on experimentation with programs and data
- Start w/words & move up through syntax and semantics

In this class

- Statistical NLP: classification & sequence models, including POS tagging
- Probabilistic parsing
- Semantic representations
- Applications
- *No coverage of speech recognition/generation, natural language generation, phonology/ morphology, ...*

Topics

- Python
- Morphology
- N-grams (multi-word)
- part of speech tagging
- Parsing (including probabilistic)
- Semantics
- Discourse

Course Prerequisites

- Finite automata, Context-free grammars
- Predicate logic, including a discussion of models
- Programming experience
  - Use Python and NLTK
- No linguistics background needed

Assessment

- Programming & Paper and Pencil hmwk
- Take-home midterm, but no final
- Large final project
- Class participation

Any Questions?