CS 581/Ling 581: Computational Linguistics

TuTh 9:30–10:45 / PSFA-113

Rob Malouf
619.594.7111
rmalouf@mail.sdsu.edu

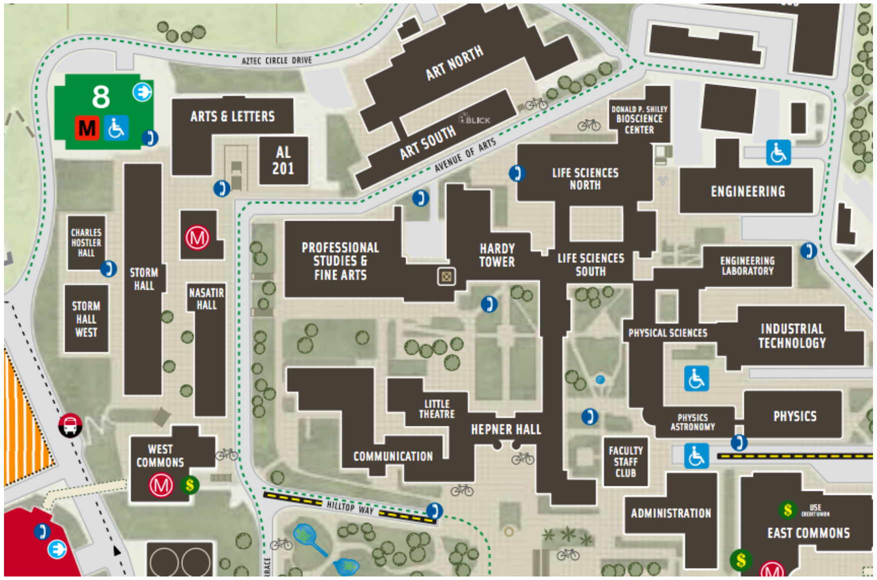
Office hours
M 11:00–12:00, Tu 1:00–2:00, or by appt (SHW 244)

Course basics

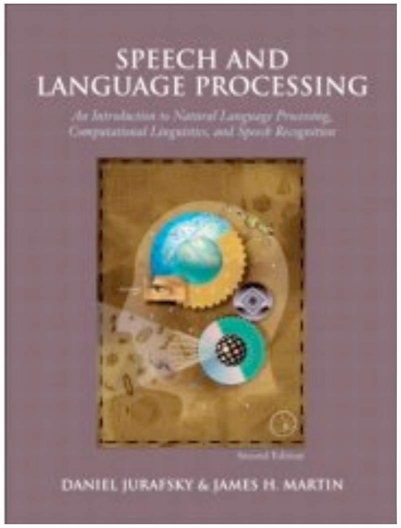
- Introduction to computational linguistics and natural language processing
- Topics to be covered:
  - regular expressions and finite state machines
  - computational morphology
  - probability and information theory
  - Markov models
  - information retrieval

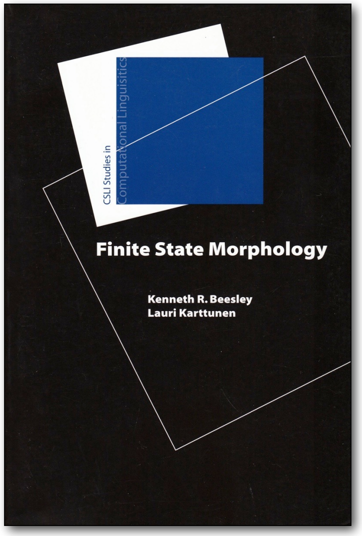
Course basics

- Prerequisites
  - At least two linguistics and/or programming classes
- Things you should know (or be ready to get caught up on):
  - Regular expressions
  - Unix
  - Python
  - Linguistics
- Safari Books Online



Computational Linguistics Lab (SHW 243)


SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN



Software

- Enthought Canopy Academic

  https://store.enthought.com/#canopy-academic

- Free if you register with an academic email address
- Don't use Canopy Express or Canopy Essentials!
- Enthought Training on Demand
  - Python Development Tools
  - Python Essentials / Advanced Python
  - NumPy / SciPy

Homework

- Check blackboard
- Read pages 1–26 in Jurafsky and Martin
- Create an account on github.com
- Add to spreadsheet at:

  https://docs.google.com/spreadsheets/d/1ZK9umxSTGo6M-jHVbKjWvU0n6_EO3DwbX976wwGbBHY/edit?usp=sharing

Homework

- Github tutorials:
  - http://git-scm.com/videos
  - https://www.youtube.com/watch?v=0fKg7e37bQE

Readings

- Online resources
  - Speech and Language Processing

    www.cs.colorado.edu/~martin/slp.html
  - Finite State Morphology: www.fsmbook.com
  - Main Python site: www.python.org
  - Software Carpentry: software-carpentry.org
  - Natural Language Toolkit: www.nltk.org
  - Safari Technical Books Online

Requirements

- The final grade will be based on:
  - homework assignments (30%)
  - a take-home midterm exam (30%)
  - a take-home final exam (40%)
- Work together on homework, but list all names
- Do not work together on exams
- Homework assignments and projects can be done in the computational linguistics lab

Computational linguistics

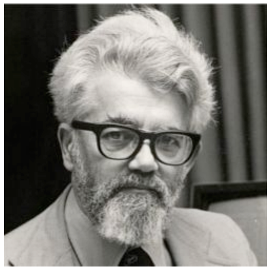
- Practical applications (NLP)
  - Artificial intelligence
  - Task-oriented speech or language systems
- Theoretical applications (linguistics)
  - Theory specification tools
  - Theoretical modeling

Foundational insights (1940–50s)

- McCulloch-Pitts neuron: a simplified computational model of a neuron
- Shannon (1948): automata for language, incorporating probabilistic models
- Chomsky (1956): formal language theory
- Sound spectrograph (Koenig et al. 1946): foundation for instrumental phonetics.
- First machine speech recognizers (Bell Labs, Davis et al. 1952)
- Shannon-Weaver information theory: Noisy channel model

Artificial intelligence

- **Machine translation** was born in 1949 with Warren Weaver's "translation memorandum"
- It was soon absorbed into **artificial intelligence**, the field co-founded by John McCarthy in the 1950's as "the science and engineering of making intelligent machines"



Machine translation

- Early translation systems based on word-for-word correspondences
- Warren Weaver's *Translation* memorandum (1949)
  - context dependent translation ("fast")
  - logical basis for language
  - cryptographic methods
  - language universals
- Georgetown-IBM demonstration (1954)
- ALPAC report (1966)

Two camps (1957–1970)

- Beginnings of Artificial Intelligence as a field (John McCarthy, Marvin Minsky, Claude Shannon, Nathaniel Rochester)
- General Problem Solver (1959), Newell and Simon's Logic Theorist: computable models of reasoning and logic
  - Subjects speak aloud as they solve problems
  - Problem solving modeled with a rule-based system where the rules (or reasoning steps) correspond to the steps human reasoners took for those kinds of problems.

Artificial intelligence

- The strong AI program aimed to write programs which could **reason**, **learn**, **plan**, and **communicate** like humans do
- Herbert Simon wrote in 1965 that "machines will be capable, within twenty years, of doing any work a man can do"
  - Simon et al.'s *General Problem Solver* (1959)
  - Weizenbaum's *ELIZA* (1966)
  - Winograd's *SHRDLU* (1970)
  - Atkin and Slate's *Chess 4.5* (1976)
- Early work looked good at first, but the field grossly overpromised and underdelivered

Two camps (1957–1970)

- George Miller and Donald Broadbent: importing computational ideas into psychology.
- Zellig Harris's Transformations and Discourse Analysis Project (1957–1958)
- Bayesian methods and optical character recognition: Using probabilistic methods on recognition problems (1950's; see Chapter 5)
- Bayesian authorship identification (Mosteller and Wallace 1964)
- Brown Corpus (1964)

Four paradigms (1970–1983)

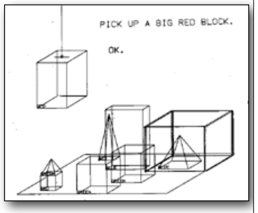
- Stochastic
  - HMMs in speech recognition (Jelinek, Bahl and Mercer at IBM)
- Logic-based programming (Prolog)
  - Q-systems and metamorphosis grammars (Colmerauer)
  - Definite Clause Grammars (Pereira and Warren 1980)
  - Unification grammar (Kay, Bresnan and Kaplan)

Four paradigms (1970–1983)

- Natural language understanding (serious attention to semantics)
  - Yale school
    - Scripts, plans, goals (Schank and Abelson, Wilensky, Lehnert). Story and text understanding.
- Discourse-modeling (Grosz, Sidner, Perrault, Allen, Cohen). Discourse as plans guided by intentions and beliefs. Communicative acts as steps in plans.

Four paradigms (1970–1983)

- Natural language understanding (serious attention to semantics)
  - Winograd's SHRDLU, blocks world (1972)



```
Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.
Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
Computer: YES, THE RED CUBE.
Person: IS IT SUPPORTED?
Computer: YES, BY THE TABLE.
Person: CAN THE TABLE PICK UP BLOCKS?
Computer: NO.
Person: CAN A PYRAMID BE SUPPORTED BY A BLOCK?
Computer: YES.
Person: CAN A PYRAMID SUPPORT A PYRAMID?
Computer: I DON'T KNOW.
Person: STACK UP TWO PYRAMIDS.
Computer: I CAN'T.
Person: WILL YOU PLEASE STACK UP BOTH OF THE RED BLOCKS AND EITHER A GREEN CUBE OR A
PYRAMID?
Computer: OK. (does it)
Person: WHICH CUBE IS SITTING ON THE TABLE?
```

Empiricism and FSMs (1983–1999)

- Finite-state models
  - phonology and morphology (Kaplan and Kay 1981, Koskenniemi, Karttunen)
  - syntax (Church 1980)
- Probabilistic models
  - Speech recognition work at IBM
  - Part of speech tagging (chapter 8)

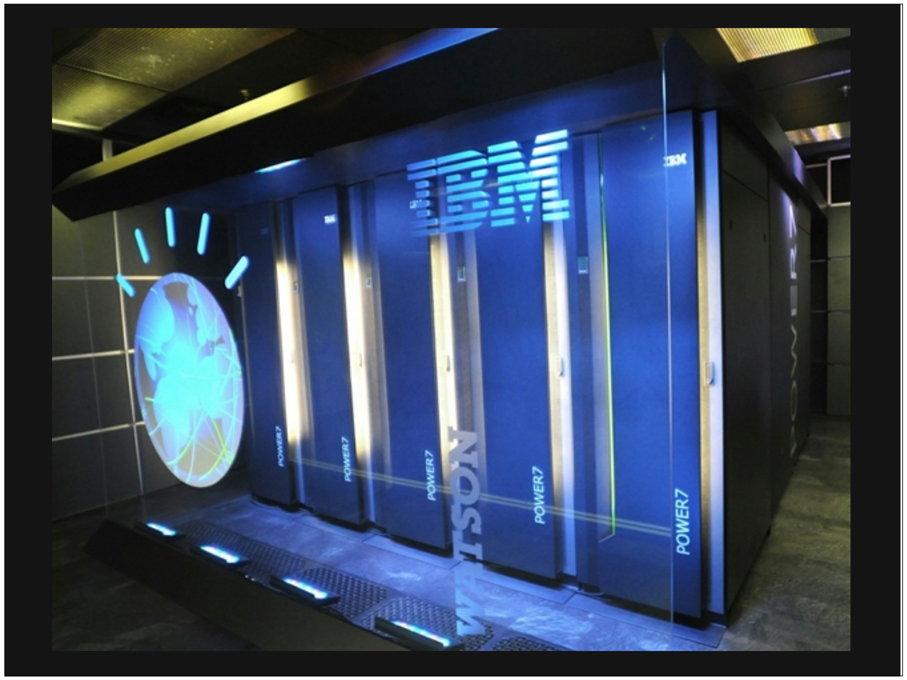
    utter, direct: Adj, V
    walk, pilot, sneer, help: N, V
    hard: Adj, Adv
  - Probabilistic parsing (chapter 12)

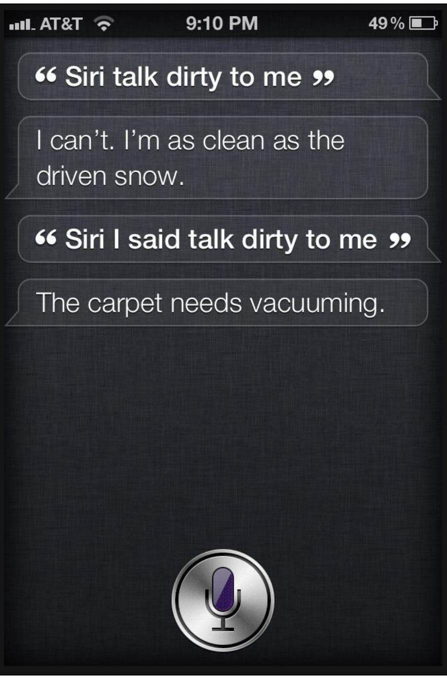
Field comes together (1994–1999)

- Spread of probabilistic methods to all kinds of problems
- Commercial ventures using speech, some NLP
- The web
- Some lessened emphasis on theoretical work

Machine Learning (2000–2008)

- General purpose statistical learning algorithms construct models of patterns given lots of examples
  - More and more data becomes available (LDC, Google)
  - More powerful computers and methods (maxent, support vector machines)
  - Focus on commercial applications in industrial settings
  - Growing interest in unsupervised learning





“ Siri talk dirty to me ”

I can’t. I’m as clean as the driven snow.

“ Siri I said talk dirty to me ”

The carpet needs vacuuming.

Have we finally achieved true natural language understanding?

No.