Statistical NLP

Spring 2010

Lecture 1: Introduction

Dan Klein – UC Berkeley

Administrivia

http://www.cs.berkeley.edu/~klein/cs288

CS 288: Statistical Natural Language Processing, Spring 2010

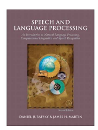
Instructor: Dan Klein
Lecture: Monday and Wednesday, 2:30pm-4:00pm, 405 Soda Hall
Office Hours: Monday and Wednesday 4pm-5pm in 775 Soda Hall.

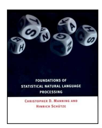
Announcements

1/19/09: The course newsgroup is ucb.class.cs288. If you use it, I'll use it!
1/19/09: The previous website has been archived.
1/19/09: Assignment 1 is posted.

Course Details

- Books:
  - Jurafsky and Martin, Speech and Language Processing, 2 Ed
  - Manning and Schuetze, Foundations of Statistical NLP
- Prerequisites:
  - CS 188 or CS 281 (grade of A, or see me)
  - Strong skills in Java or equivalent
  - Deep interest in language
  - There will be a lot of math and programming
- Work and Grading:
  - Four assignments (individual, write-ups)
  - Final project (group)





Announcements

- Computing Resources
  - You will want more compute power than the instructional labs
  - Experiments will take minutes to hours, with efficient code
  - Recommendation: start assignments early
- Communication:
  - Announcements: webpage
  - Public discussion: newsgroup
  - My email: klein@cs
- Enrollment:
  - Undergrads stay after and see me
- Questions?

AI: Where Do We Stand?

Hollywood

R2D2

KITT

Wall-E

'80 '90 '00 '10

Rule based approaches

Early statistical approaches

Modern statistical approaches

Reality

Nimbro Soccer Robots '04

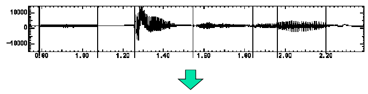
Stanford Racing Team '05

What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!
- End systems that we want to build:
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Unknown: human-level comprehension (is this just NLP?)

Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV



"Speech Lab"

- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)

Information Extraction

- Unstructured text to database entries

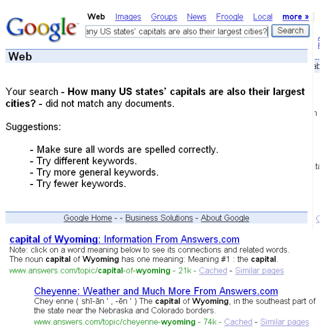
New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: perhaps 80% accuracy for multi-sentence temples, 90%+ for single easy fields
- But remember: information is redundant!

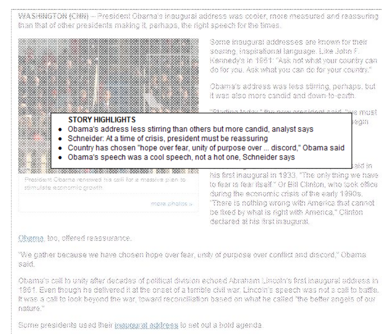
Question Answering

- Question Answering:
  - More than search
  - Ask general comprehension questions of a document collection
  - Can be really easy: "What's the capital of Wyoming?"
  - Can be harder: "How many US states' capitals are also their largest cities?"
  - Can be open ended: "What are the main issues in the global warming debate?"
- SOTA: Can do factoids, even when text isn't a perfect match



Summarization

- Condensing documents
  - Single or multiple
  - Extractive or synthetic
  - Aggregative or representative
  - Even just shortening sentences
- Very context-dependent!
- An example of analysis with generation

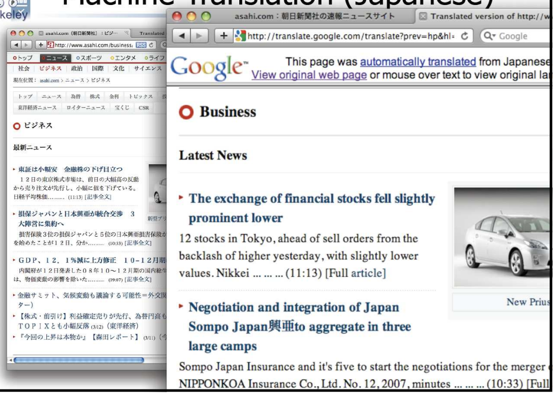


Machine Translation

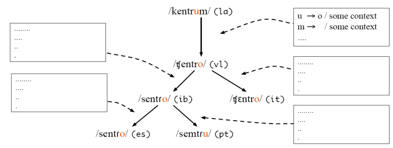
- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (next class) vs fidelity (later)

Machine Translation (French)

Machine Translation (Japanese)



Etc: Historical Change



Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Center	centrum	centro	centro	centro

- Change in form over time, reconstruct ancient forms, phylogenies
- … just an example of the many other kinds of models we can build

Language Comprehension?

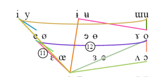
"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake", because he was:

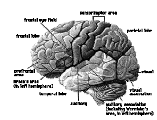
(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery

What is Nearby NLP?

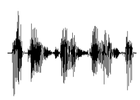
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!



- Speech?
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP



What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Statistical Modeling Methods
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We'll focus on what makes the problems hard, and what works in practice…

Class Requirements and Goals

- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models (parts of cs281)
    - Basic linguistics background (ling101)
    - Decent coding skills (Java) well beyond cs61b
  - Most people are probably missing one of the above
  - You will often have to work on your own to fill the gaps
- Class goals
  - Learn the issues and techniques of statistical NLP
  - Build realistic tools used in NLP (language models, taggers, parsers, translation systems)
  - Be able to read current research papers in the field
  - See where the holes in the field still are!

Some BIG Disclaimers

- The purpose of this class is to train NLP researchers
  - Some people will put in a LOT of time
  - There will be a LOT of reading, some required, some not – you will have to be strategic about what reading enables your goals
  - There will be a LOT of coding and running systems on substantial amounts of real data
  - There will be a LOT of statistical modeling (though we do use a few basic techniques very heavily)
  - There will be discussion and questions in class that will push past what I present in lecture, and I'll answer them
  - Not everything will be spelled out for you in the projects
- Don't say I didn't warn you!

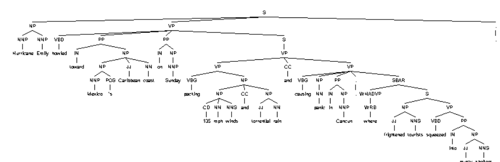
Some Early NLP History

- 1950's:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military
    - Toy models: MT using basically word-substitution
  - Optimism!
- 1960's and 1970's: NLP Winter
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
  - Work shifts to deeper models, syntax
  - … but toy domains / grammars (SHRDLU, LUNAR)
- 1980's and 1990's: The Empirical Revolution
  - Expectations get reset
  - Corpus-based methods become central
  - Deep analysis often traded for robust and simple approximations
  - *Evaluate everything*
- 2000+: Richer Statistical Methods
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up
  - *Begin to get both breadth and depth*

Problem: Ambiguities

- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Ban on Nude Dancing on Governor's Desk
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Why are these funny?

Syntactic Analysis



Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun , where frightened tourists squeezed into musty shelters .

- SOTA: ~90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples

Semantic Ambiguity

- NLP is much more than syntax!
- Even correct tree structured syntactic analyses don't fully nail down the meaning

Every morning someone's alarm clock wakes me up

John's boss said he was doing better

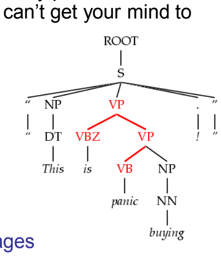
- In general, every level of linguistic structure comes with its own ambiguities…

Dark Ambiguities

- *Dark ambiguities*: most structurally permitted analyses are so bad that you can't get your mind to produce them

This analysis corresponds to the correct parse of

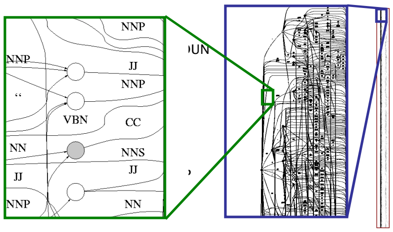
"This will panic buyers ! "



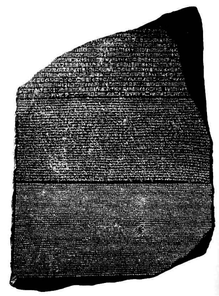
- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

Problem: Scale

- People *did* know that language was ambiguous!
  - …but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
  - …they didn’t realize how bad it would be



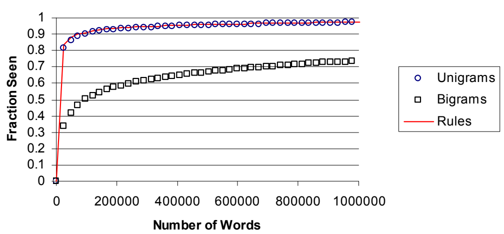
Corpora



- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what

Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire



Outline of Topics

- Words
  - N-gram models and smoothing
  - Classification and clustering
- Sequences
  - Part-of-speech tagging
  - Information extraction
  - Speech recognition / synthesis
- Trees
  - Syntax and semantics
  - Machine translation
  - Question answering
- Discourse
  - Reference resolution
  - Dialog systems

A Puzzle

- You have already seen N words of text, containing a bunch of different word types (some once, some twice…)
- What is the chance that the N+1st word is a new one?