The Future of NLP

A Few Random Remarks

Computational Linguistics

- We can study anything about language …
- 1. Formalize some insights
- 2. Study the formalism mathematically
- 3. Develop & implement algorithms
- 4. Test on real data

The Big Questions

- What are the right formalisms to encode linguistic knowledge?
  - Discrete knowledge: what is possible?
  - Continuous knowledge: what is likely?
- How can we compute efficiently with these formalisms?
  - Or find approximations that work pretty well?

Reprise from Lecture 1: What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

- These ambiguities now look familiar
- You now know how to solve some:
  - Word sense disambiguation
  - PP attachment
- You can imagine how to solve others:
  - Which NP does "it" refer to? (pronoun reference resolution)
  - Could use techniques from word-sense disambig. or language modeling
- Others still seem beyond the state of the art:
  - Anything that requires semantics or reasoning

Some of the Active Research

- Syntax: It's converging, but still messy
  - New: Attach probabilities to "deep structure" of syntax
- Phonology: Formalism under hot development
- Speech:
  - Better language modeling (predict next word)
  - Better models of acoustics, pronunciation
  - Emotional speech, kids/old folks, bad audio, conversation
  - Adaptation to particular speakers and dialects
- Translation models and algorithms
- Semantic theories and connection to AI – use stats?
  - Too many semantic phenomena. Really hard to determine and disambiguate possible meanings.

Some of the Active Research

- All of these areas have learning problems attached.
- We're really interested in **unsupervised** learning.
- How to learn FSTs and their probabilities?
- How to learn CFGs? Deep structure?
- How to learn *good* word classes?
- How to learn translation models?

Semantics Still Tough

- **"The perilously underestimated appeal of Ross Perot has been quietly going up this time."**
  - Underestimated by whom?
  - Perilous to whom, according to whom?
  - "Quiet" = unnoticed; by whom?
  - "Appeal of Perot" ⇐ "Perot appeals ..."
    - a court decision?
    - to someone/something? (actively or passively?)
  - "The" appeal
  - "Go up" as idiom; and refers to amount of subject
  - "This time" : meaning? implied contrast?

Deploying NLP

- Speech recognition and IR have finally gone commercial over the last few years.
- But not much NLP is out in the real world.
- What killer apps should we be working toward?

- Resources:
  - Corpora, with or without annotation
  - WordNet; morphologies; maybe a few grammars
  - Perl, Java, etc. don't come with NLP or speech modules, or statistical training modules.
  - But there are research tools available:
    - Finite-state toolkits
    - Machine learning toolkits (e.g., WEKA)
    - Annotation tools (e.g., GATE)
    - Emerging standards like VoiceXML
    - Dyna – a new programming language being built at JHU

Deploying NLP

- Sneaking NLP in through the back door:
  - Add features to existing interfaces
    - "Click to translate"
    - Spell correction of queries
    - Allow multiple types of queries (phone number lookup, etc.)
    - IR should return document **clusters** and **summaries**
    - From IR to QA (question answering)
    - Machines gradually replace humans @ phone/email helpdesks
  - Back-end processing
    - Information extraction and normalization to build databases: CD Now, New York Times, ...
    - Assemble good text from boilerplate
  - Hand-held devices
    - Translator
    - Personal conversation recorder, with topical search

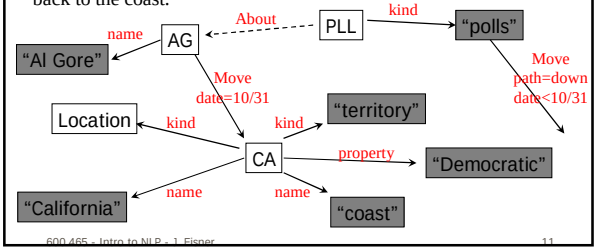
IE for the masses?

"In most presidential elections, Al Gore's detour to California today would be a sure sign of a campaign in trouble. California is solid Democratic territory, but a slip in the polls sent Gore rushing back to the coast."

NAME	AG	"Al Gore"	
NAME	CA	"California"	
NAME	CO	"coast"	
MOVE	AG	CA	TIME=Oct. 31
MOVE	AG	CO	TIME=Oct. 31
KIND	CA	Location	
KIND	CA	"territory"	
PROPRTY	CA	"Democratic"	
KIND	PLL	"polls"	
MOVE	PLL	?	PATH=down, TIME<Oct. 31
ABOUT	PLL	AG	

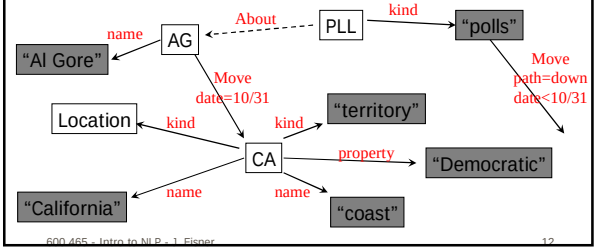
IE for the masses?

"In most presidential elections, Al Gore's detour to California today would be a sure sign of a campaign in trouble. California is solid Democratic territory, but a slip in the polls sent Gore rushing back to the coast."



IE for the masses?

- "Where did Al Gore go?"
- "What are some Democratic locations?"
- "How have different polls moved in October?"



IE for the masses?

- Allow queries over meanings, not sentences
- Big semantic network extracted from the web
- Simple entities and relationships among them
- Not complete, but linked to original text
- Allow inexact queries
  - Learn generalizations from a few tagged examples
- Redundant; collapse for browsability or space

Dialogue Systems

- Games
- Command-and-control applications
- "Practical dialogue" (computer as assistant)
- The Turing Test

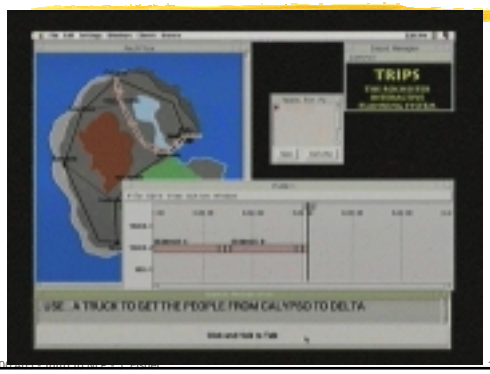
Turing Test

Q: Please write me a sonnet on the subject of the Forth Bridge.
A [either a human or a computer]: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give an answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have my K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.

Turing Test

Q: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?
A: It wouldn't scan.
Q: How about "a winter's day"? That would scan all right.
A: Yes, but nobody wants to be compared to a winter's day.
Q: Would you say Mr. Pickwick reminded you of Christmas?
A: In a way.
Q: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
A: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

TRIPS System



TRIPS System

Dialogue Links (click!)

- Turing's article (1950)
- Eliza (the original chatterbot)
  - Weizenbaum's article (1966)
  - Eliza on the web - try it!
- Loebner Prize (1991-2001), with transcripts
  - Shieber: "One aspect of progress in research on NLP is appreciation for its complexity, which led to the dearth of entrants from the artificial intelligence community - the realization that time spent on winning the Loebner prize is not time spent furthering the field."
- TRIPS Demo Movies (1998)
- Gideon Mann's short course next term

JHU's Center for Language and Speech Processing (CLSP)

- One of the biggest centers for NLP/speech research
- Core faculty:
  - Jason Eisner & David Yarowsky (CS)
  - Bill Byrne, Fred Jelinek, & Sanjeev Khudanpur (ECE)
  - Bob Frank & Paul Smolensky (Cognitive Science)
  - Others loosely associated – machine learning, linguistics, etc.
- Lots of grad students
- Focus is on core grammatical and statistical approaches
  - Many current areas of interest, including multi-faculty projects on machine translation, speech recognition, optimality theory
- More coursework, reading groups
- Speaker series: Tuesday 4:30 when classes are in session