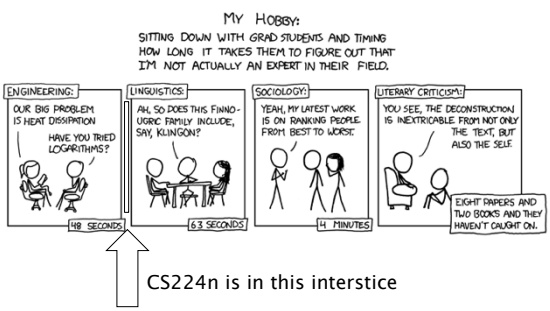


CS224n is in this interstice

cartoon from xkcd.com

Natural Language Processing CS224N/Ling284

Christopher Manning
Lecture 1

Course logistics in brief

- Instructor: Christopher Manning
- TAs: Spence Green, Leon Lin, Richard Socher … and probably more
- Time: MW 11:00–12:15, Skilling Aud … maybe moving to Gates B03
- Programming language: mainly Java
- Other information: see the class2go webpage
  - http://cs224n.stanford.edu/
- "Handouts": online

This class

- Assumes you come with some skills…
  - Some basic linear algebra, probability, and statistics; decent programming skills
  - But not everyone has the same skills
    - Assumes some ability to learn missing knowledge
- Teaches key theory and methods for (statistical) NLP: MT, information extraction, parsing, semantics, etc.
  - Learn techniques which can be used in practical, robust systems that can (partly) understand human language
- But it's something like an "AI Systems" class:
  - A lot of it is hands-on, problem-based learning
  - Often practical issues are as important as theoretical niceties
  - We often combine a bunch of ideas

Goals of the field of NLP

- Computers would be a lot more useful if they could handle our email, do our library research, chat to us …
- But they are fazed by natural human languages.
  - Or at least their programmers are … most people just avoid the problem and get into menus and radio buttons, or XML, or the so-called semantic web, or …
- But someone has to work on the hard problems!
  - How can we tell computers about language?
  - Or help them to learn it as kids do?

Natural language: the earliest UI

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave. I'm afraid I can't do that.

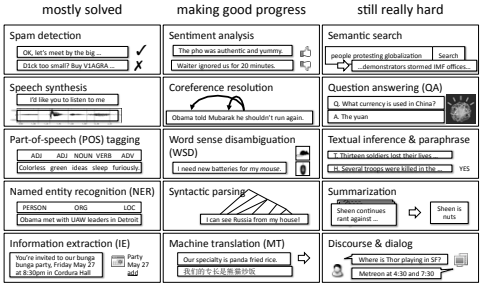

(cf. also false Maria in Metropolis – 1926)

What/where is NLP?

- Goals can be very far reaching …
  - True text understanding and interpretation
  - Real-time participation in spoken dialogs
  - High quality machine translation
- Or very down-to-earth …
  - Finding the price of products on the web
  - Analyzing reading level or authorship statistically
  - Sentiment detection about products or stocks
  - Extracting names, facts or relations from documents
- These days, the latter predominate
  - As NLP becomes increasingly possible, it becomes increasingly engineering-oriented
    - Also related to changes in approach in AI/NLP in general

Commercial world

Language Technology



The hidden structure of language

- We're going beneath the surface…
  - Not just string processing
  - Not just keyword matching in a search engine
    - This is the move that Google has been increasingly engaged in in recent years
    - Moving from matching keywords to satisfying user needs
  - Not just converting a sound stream to a string of words
    - Like Nuance/Google speech recognition
- We want to recover and manipulate at least *some* aspects of language structure and meaning

Is the problem just cycles?

- Bill Gates, Remarks to Gartner Symposium, October 6, 1997:
  - Applications always become more demanding. Until the computer can speak to you in perfect English and understand everything you say to it and learn in the same way that an assistant would learn – until it has the power to do that – we need all the cycles. We need to be optimized to do the best we can. Right now linguistics are right on the edge of what the processor can do. As we get another factor of two, then speech will start to be on the edge of what it can do.

The early history: 1950s

- Ear
  les
- Fou
  lan
- Fir
- MT
  wo
- Litt
  ser
- Pro

Why NLP is difficult: Newspaper headlines

1. Minister Accused Of Having 8 Wives In Jail
2. Juvenile Court to Try Shooting Defendant
3. Teacher Strikes Idle Kids
4. Miners refuse to work after death
5. Local High School Dropouts Cut in Half
6. Red Tape Holds Up New Bridges
7. Clinton Wins on Budget,
8. Hospitals Are Sued by 7
9. Police: Crack Found in M

Minister Accused Of Having 8 Wives In Jail
May 21, 2007 06:49 AM

ATLANTA (AP) -- A tra
served two years in pri
has been jailed again fo
marry more women.

Bishop Anthony Owens,
Ga., is in a Gwinnett Co
four women claimed he
after being released fro

Why is natural language understanding difficult

Fed raises interest rates 0.5% in effort to control inflation

- *NYT* headline, from better economic times (17 May 2000)

Language: still the ultimate UI

Where is A Bug's Life playing in Mountain View?

A Bug's Life is playing at the Century 16 Theater.

When is it playing there?

It's playing at 2pm, 5pm, and 8pm.

OK. I'd like 1 adult and 2 children for the first show. How much would that cost?

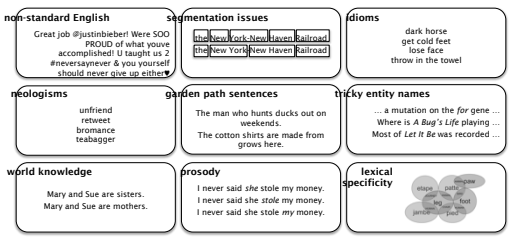
But we need domain knowledge, discourse knowledge, world knowledge, linguistic knowledge.

Why is natural language computing hard?

- Natural language is:
  - highly ambiguous at all levels
  - complex and subtle use of context to convey meaning
  - fuzzy, probabilistic
  - involves reasoning about the world
  - a key part of people interacting with other people (a social system):
    - persuading, insulting and amusing them
- But NLP can also be surprisingly easy sometimes:
  - rough text features can often do half the job

OK, why *else* is NLP hard?

Oh so many reasons!



But that's what makes it fun!

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- The answer that's been getting traction:
  - probabilistic models built from language data
    - P("maison" → "house") high
    - P("L'avocat général" → "the general avocado") low
- Some computer scientists think this is a new "A.I." or "machine learning" idea
  - But really it's an older idea that was taken from the electrical engineers….

Where do we head?

Look at subproblems, approaches, and applications at different levels

- Statistical machine translation
- Statistical NLP: classification and sequence models (part-of-speech tagging, named entity recognition, information extraction)
- Syntactic (probabilistic) parsing
- Building semantic representations from text. QA.

- *(Unfortunately left out: natural language generation, phonology/ morphology, speech dialogue systems, more on natural language understanding, …. There are other classes for some! cs224u/s)*

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

⇨

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

The classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

About $26 billion spent annually on human translation.

Scott Klemmer: I learned a surprising fact at our research group lunch today. Google Sketchup releases a version every 18 months, and the primary difficulty of releasing more often is not the difficulty of producing software, but the cost of internationalizing the user manuals!

Many slides from Kevin Knight (at ISI)

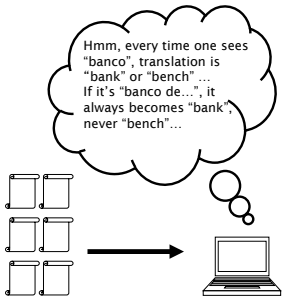
Statistical Solution

- Parallel Texts
  - Rosetta Stone

Hieroglyphs
Demotic
Greek

Statistical Solution

- Parallel Texts
  - Instruction Manuals
  - Hong Kong/Macao Legislation
  - Canadian Parliament Hansards
  - United Nations Reports
  - Official Journal of the European Communities
  - Translated news



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .