6.863J/9.611J Natural language & computers
Lecture 1: Walking the walk, talking the talk

Professor Robert C. Berwick
berwick@csail.mit.edu
Course web page:
http://www.mit.edu/~6.863/spring2011/
The Menu Bar

• The Rules of the Game: goals of the field and the course (yes, Anne Hunter, it is an AUS this term)
• Mark Twain’s problem
• The “WYSIWYG” language problem
• We haven’t had a sale in 40 years
• For next time… (predict what word comes next!)
Goals of the field: the problem of natural language understanding

Computers would be a lot more useful if they could handle our email, do our library research, talk to us …

How can we tell computers about language? (Or help them learn it as kids do?)
Natural language processing (NLP)

- We know what we want from computer software
- “killer applications”
  - those that can make sense of language data
    - retrieve language data: (IR)
    - summarize knowledge contained in language data
    - answer questions (QA), make logical inferences
    - translate from one language into another
    - recognize speech: transcribe medical records,…
- In short: we want computers to be smart about language and pass the Turing test…
- Well, perhaps not *that* smart…
Some examples

• Answering questions:
  – *Bin Laden is too angry to talk to*
  – Q: Can we talk to Bin Laden easily?
  – *Bin Laden is too angry to talk to Cheney*

• Extracting information – from biological papers, can a computer figure out what proteins interact? (‘data mining’)

• And, the usual Holy Grail…
Killer Application: computer translation
Has Google solved this?

How does this work?
Does it work?
Language understanding?

Translate Text

Original text:

Book them, Danno

Automatically translated text:

Libreta de ellos, Danno
Goals of the course

• Introduce you to natural language processing problems
• Gain working knowledge of the two key approaches have been to solving NLP problems: linguistics & statistics

By the end of the course you should:
• Agree that language is subtle & interesting
• Feel some ownership over the formal & statistical models
• Understand research papers in the field
Course organization, I

- Me: Bob Berwick, berwick@csail.mit.edu
- TA: Igor Malioutov igorm@mit.edu
- Course web page for all readings/assignments/code: http://web.mit.edu/6.863/spring2011/
- 2 Lectures/week; office hours & lab hours if needed; ‘official’ lab hours if needed
- Lab oriented, w/ a few Winston-style “reading and responses” (R&R)
- No final exam
- 6-7 Labs, 1 final project lab; final project lab is joint
- Labs typically 2-3 weeks long
- R&R, typically out on Wednesday, due Sunday 5pm/Monday next class for mandatory in-class discussion
- All work can be done jointly, but you must write-up your own reports, identifying who you worked with
- Email pdf/web URLs for Lab write-ups to: 6.863-graders@mit.edu
Course organization, II

• Grade determined by:
  – 60% Lab assignments & Competitive Grammar (CGW)
  – 20% Reading & Response & class participation
  – 20% final project

• Late days to help with time management (7 days) – see web page for details (1 day = 24 consecutive hours)
  • So you won’t have to ask me for ‘extensions’
  • Use your late days wisely!
(Required!) Textbook (Jurafsky & Martin)

New and improved for 2008
(2nd edition)

Nearly 1000 pages
(full year’s worth...)
25 chapters
Divided into 5 parts

I. Words
II. Speech
III. Syntax
IV. Semantics and Pragmatics
V. Applications

For Monday: Read ch 1.; pp.83-94; 114-116. (online; also on Barker reserve P98.J87 2009);
So, not so fast, re language being ‘solved’…?
More Problems…

- *Mark Twain’s problem*: Parents spend…
- *The non-WYSIWYG problem*: language is not ‘WYSIWIG’ (required information is sometimes just not *there* – not what appears ‘on the surface’…and so…how can we learn from what is statistically invisible ‘on the surface’?)
- *Language-as-communication* problem: Not even obviously well-designed for ‘ease of communication’ (ask any diplomat, teacher, student, …); it is *ambiguous*, leading to nondeterminism
Human language is not ‘wysiwig’ and not designed for ‘ease of communication’

• Invisible elements – shared knowledge – a window into the human mind
• *Bin Laden is too angry to talk to*
• *Bin Laden is too angry to talk to Cheney*
• What are the ‘invisible elements’ here?
• Human language is *ambiguous* – sometimes on purpose
  *We haven’t had a sale in forty years*
Language ambiguity makes NLP hard: Especially in news headlines

Iraqi Head Seeks Arms
Juvenile Court to Try Shooting Defendant
Teacher Strikes Idle Kids
Stolen Painting Found by Tree
Kids Make Nutritious Snacks
Local HS Dropouts Cut in Half
Obesity Study Looks for Larger Test Group
Levels of Description in language

• **Phonetics/phonology/morphology**: what words (or subwords) are we dealing with?
• **Syntax**: What phrases are we dealing with? Which words modify one another?
• **Semantics**: What’s the literal meaning?
• **Pragmatics**: What should you conclude from the fact that I said something? How should you react?
And even subtler ambiguity

• Q: Why does my high school give me a suspension for skipping class?

• A: Administrative error. They’re supposed to give you a suspension for auto shop, and a jump rope for skipping class.
What makes NLP hard? The ‘Road Touring Test’ for NLP…

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.

To get a donut (spare tire) for his car?
Why is this story hard?

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.

store where donuts shop? or is run by donuts? or looks like a big donut? or made of donut? or has an emptiness at its core?
Why is this story hard?

I stopped smoking freshman year, but 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.
Why is this story hard?

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.

Describes where the store is? Or when he stopped?
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.

Well, actually, he stopped there from hunger and exhaustion, not just from work.
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.

At that moment, or habitually?

(Similarly: Mozart composed music.)
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.

But actually, a coffee only stays good for about 10 minutes before it gets cold.
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.

the particular coffee that was good every few hours? the donut store? the situation?
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.

too expensive for what? what are we supposed to conclude about what John did?
how do we connect “it” to “expensive”? 
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.

That’s how often he thought it?
The secret to success

• “Science is the art of reducing complex visibles to simple invisibles” – Jacques Perrin, Nobel Prize in Chemistry for atomic theory, 1920

• So: what are the ‘simple invisibles’ – the ‘atoms of language’?
Problem: Deep Language Analysis

- Sentence/Question Understanding

"Which target did you say was too distant to consider without retargeting?"
There can be even more complex ‘invisible’ element examples…

- Consider the question:
  - Which report did you file without reading?
  - Now there are lots of ‘missing elements’
  - Which report did you file [that report] without [you] reading [that report]

- And there are no other possible interpretations
  - meaning (for example) that we cannot be asking about some report that you filed but someone else read
Example continued

• Consider:
  – The report was filed without reading
  – The report was filed after Bill read
  – The report was filed without being read
  – These papers are easy to file without reading
  – This report is not worth reading without attempting to analyze deeply

• What are these rules? Were you taught them? Can you find them in grammar books?

• A very subtle interacting pattern of constraints
Q: How do existing ‘statistically based’ systems do?

Ans: Not very well

- semantic role labeler (Koomen, Punyakanok, Roth & Yih 2005)
  - uses Statistical Parser + Machine Learning Techniques for labeling

perhaps state-of-the-art performance but performance on this task is poor
which report did you file without reading
Lots of constraints…!

• The constraints appear to be modular and sometimes reflexive:
  *The CIA said that Bin Laden died yesterday*
  *The CIA said that Bin Laden will die yesterday*

• What’s the difference between these two?
  *He chuckled his way through the meeting*
  *He arrived his way through the meeting*
  *He arrived drunk*

“I saw one car flattened down to about one foot high … And my mechanics friend told me that the driver who got out of that cab that was squashed down by accident got out by a narrow escape”

How did you acquire all of this knowledge?
Can a computer acquire all of this knowledge?
So what makes a language a natural language?

- And why do we care about this?
- Infinite & digital (no 2.5 long words), hence recursive
- Constraints: at all ‘levels of representation’
- Language uses **hierarchical structure**
- Language uses **constraints**
  
  $\text{fox} + s \rightarrow \text{foxiz}; \text{cat} + s \rightarrow \text{cats}; \text{dog} + s \rightarrow \text{dogz}; \text{Bach} + s \rightarrow ??$

  Schank killed the spider; the spider killed Schank;
  the rabbit appeared
  the magician appeared the rabbit

- Babies can **learn** natural languages (or rather: acquire them)
- We can figure out the constraints
- We can exploit the constraints for computation
- But the brain also seems to use statistics (read: frequency)
Ingredient for language: chimps can categorize

But they can’t sing worth spit...

Ingredient for language: sophisticated vocal control

(bird song)
Which bird is this?

Juan ordéñame las cabras
‘John milks the goats’

(whistling)

Domingo está enfermo
‘Domingo was sick’

Silbo Gomero
What's the missing ingredient?

Patrick + saw Marvin

\[ X \quad Y \]

Patrick + saw Marvin

\[ Z \]

Patrick saw Marvin

I said + Patrick saw Marvin

\[ X \quad Y \]

\[ Z \]

I said Patrick saw Marvin

I said + Patrick saw Marvin

\[ Cons \ X,Y \]
Now we can do things like this

“Almost inconceivably, the gun into which she was now staring was clutched in the pale hand of an enormous albino with long white hair”

Dan Brown, *The Da Vinci Code*
The Ingredients

Birds  Chimps  Humans

got rhythm  got categories  got cons

Who could ask for anything more?
How is *cons* implemented?

What makes a language a *natural* language?

- Language computes with **hierarchical structure** *not* by ‘linear order’
  - Example:
    
    *Obama likes him; Obama thinks that Palin likes him*

- Babies (and savants) can learn it easily (Mark Twain’s problem)
- How do they do this?
- Apparently, they combine *constraints* that we can *leverage computationally* along with simple statistical analysis…
Challenge: language learning & language structure – people do not ‘count’

- The man went yesterday
- Emphatic: The man went-\textit{nog} yesterday
- Ordinary Passive:
  \textit{By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed}
- Modified ‘counting’ passive:
  \textit{By 1997, almost WOW! all remaining uses of cancer-causing asbestos will outlawed}

What do people do?
The fMRI envelope please…

Better learning structure
Worse learning counting

Ref: Musso et al, Nature Neuroscience, July 2003
The two kinds of computations

• (Linear) Association: beads on a string (‘bread and…’)

• (Recursively) Structural: nested boxes
…Correspond to two distinct brain areas…

Ref: Frederici, PNAS, 2006
…Correspond to two distinct brain areas…

Structure learning area activated when trying to (surprise) learn structure
And that’s what we’ve got that other animals do not...
How do babies learn language?

well later perhaps some little sea creature was in there once your ball player?
Challenge: segmentation
twasbrilligandtheslithytovesdiddlygyre

\{pabiku, tibudo, daropi, golatu\}

pabikutibudodaropipabiku
tibudodaropitibudodaropi
pabikudaropipabikugolatu
tibudogolatutibudogolatu
golatudaropipabikutibudo
daropigolatudaropipabiku
tibudogolatudaropigolatu
daropigolatupabikutibudo
pabikutibudodaropigolatudaropi
...

(sounds)
Challenge: Combining Inference with Cognitive Constraints
(How real people solve real problems can help real computers)

Problem:  

twasbrilligandtheslithytovesdidgyreandgimble

“Standard” solution: prettybaby  pre-ty-ba-by

Graph of transition probabilities: $\Pr(x_{i+1}|x_i)$ & look for local minima

“Standard” claim: works great; “stats is all you need” (Science, 1996)

$pabikutibudodaropigolatu$…

Pr(bi|pa)=1.0;  Pr(ku|bi)=1.0;  Pr(ti|ku)=0.3,  Pr(bu|ti)=1;  Pr(do|bu)=1.0;  Pr(da|do)=0.3
Pr(ro|da)=1;  Pr(pi|ro)=1.0;  Pr(go|pi)=0.3
Pr(la|go)=1.0;  Pr(tu|la)=1.0  …

Pigola→ pi gola  Works great? NO!!!
Actual results on actual speech to children: works lousy
What’s the answer? But, add a single, universal constraint and it works GREAT!

Precision and Recall, Pure Stat
Interference vs. Stat Inference + UG,
250,000 child-directed examples

Only statistics

\[ \text{precision} = \frac{\text{true positives}}{\text{true positives + false positives}} \]

\[ \text{recall} = \frac{\text{true positives}}{\text{true positives + false negatives}} \]

What IS this ONE universal constraint????

Input Data = $2500/baby

Q: How can we fix this?
A: add universal linguistic constraint
Modeling language
Let’s start with ‘beads on a string’

*It was a bright cold day in April and the clock was striking...*
What is a language model?

- Single letter or word frequencies: 1-grams ("unigrams")
- Useful in solving cryptograms: ETAOINSHRDLU…
- If you know the previous letter: 2-grams (bigrams)
  - "h" is rare in English (4%; 4 points in Scrabble)
  - *but* "h" is common after "t" (20%)
- If you know the previous 2 letters: 3-grams (trigrams)
  - "h" is *really* common after "(space) t" etc. …
How to define this sort of language ‘model’?

- Simplest idea:
  A probability function that assigns probabilities to word sequences, \( w, p(w) \), e.g.,
  \( w = (i, love, the, Red, Sox) \)
- Turn on radio at some arbitrary moment & you hear these next 4 words (perhaps in the middle of some longer sentence)
- Note that the space of ‘possible events’ \( E \) here is actually infinite (since an infinite # of sequences starting with “...
For next time (Monday)

- Do R&R #1 on GoogleTalk (3 short readings), and 2 page write up
  - email this to me, berwick@csail.mit.edu by 6pm SUNDAY, EST
  - bring hardcopy to class for in-class discussion & revision, then turn in this final hardcopy Monday

- Install NLTK (see instructions in R&R #1)
- Do background reading for RR as per website
- See webpage for all lecture notes, labs, etc., etc.: http://www.mit.edu/~6.863/spring2011/

After all, there might be some ambiguity…
Ambiguity

• British Left Waffles on Falkland Islands
• Red Tape Holds Up New Bridges
• Man Struck by Lightning Faces Battery Charge
• Clinton Wins on Budget, but More Lies Ahead
• Hospitals Are Sued by 7 Foot Doctors

• We haven’t had a sale in forty years