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

Professor Robert C. Berwick
berwick@csail.mit.edu
Course home web page:
The Menu Bar

• The rules of the game
• Mark Twain was right
• The Big Bang
• What makes a natural language natural?
• What computers don’t know; what people do know
• Natural language is not WYSIWYG
• We haven’t had a sale in 40 years
• For next time...
The Problem of Language Understanding

- Analyzing the content of sentences
  - beyond scanning for content words or keywords
  - beyond identifying simple phrases (shallow parsing)

human brains are really good at it (largely done without conscious thought)
computer programs are getting better but there’s still a long way to go
Language & computers

• 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?

Original text:
Perhaps this is the right way to translate Spanish.

Automatically translated text:
Quizás esta es la forma correcta de traducir español.

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

Translate Text

Original text:

Book them, Danno

Automatically translated text:

Libreta de ellos, Danno
The problem of Christopher
We want you to learn

• *How* to be as good as Christopher
• *Why* this is a hard problem
• *What* the two key approaches have been to solving it
• *Gain working knowledge* of current methods & their limits
Course organization, I

- Me: Bob Berwick, berwick@csail.mit.edu
- Lab TAs: Yuan Shen (yks@csail.mit.edu); Gabriel Zaccack (gabi@csail.mit.edu)
- 2 Lectures/week; office hours & lab hours if needed
- Lab oriented, w/ Winston-style “reading and responses” (R&R)
- No final exam
- 4-5 Labs, 1 final project lab; final project lab is joint
- Labs typically 2-3 weeks
- 4-5 R&R, out on Weds, due in Monday next class for discussion
- All work can be done jointly, but you must write-up your own reports, identify who you work with
- Email pdf/web URLs for Lab write-ups to: 6.863-graders@mit.edu
Course organization, II

- Grade determined by:
  - 45% Labs
  - 35% Reading & Response - with mandatory class participation
  - 20% final project
- Late days to help with time management (30 days)
(Optional!) 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 – not this course
III. Syntax
IV. Semantics and Pragmatics
V. Applications

For Monday: Read ch. 1 (online)
(see homepage)
So, not so fast, re language being ‘solved’...?

More Problems...

- Christopher’s problem
- 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
- *John is too angry to talk to*
- *John is too angry to talk to Bill*
- *John is too angry to talk to* means: John is too stubborn for someone or other to talk to *John*
- Important to get it right!
- *Bin Laden is too angry to talk to*
- Human language is *ambiguous* – sometimes on purpose
  "We haven’t had a sale in forty years"
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’?
Example: from a complex visible to a simple invisible

- Bin Laden is too angry to talk to

- talk to: predicate
- two arguments: (talker) (“talkee”)
- neither argument is explicitly identified in the sentence
- talker = “arbitrary person”
- talkee = Bin Laden
The Problem of Language Understanding

- An example
  - Bin Laden is too angry to talk to \( \text{talk_to}(X, \text{BL}) \)
  - Bin Laden is too angry to talk to \( \text{Cheney} \)
    - talker = Bin Laden
    - talkee = Cheney
    - \( \text{talk_to}(\text{BL}, \text{Cheney}) \)
The Problem of Language Understanding: Hidden structure in the human mind/brain

• An example

BL is too angry to talk to talk_to(X,BL)

BL is too angry to talk to Cheney talk_to(BL,Cheney)

• A linguistic (i.e., representational) explanation

Missing subject of to+verb (‘to talk’…) should be linked with next higher referring expression
This relation is blocked by intervening element

This explanation can be supported by other data/constructions and as a quick ‘take’: this is analogous to variable-binding frames in programming languages – an example of a (tacit) computational constraint on human language
Problem: Deep Language Analysis

- Sentence/Question Understanding

"Which target did you say was too distant to consider without retargeting?"
And 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
The Problem of Language Understanding

- **Parsing** is the process of
  - breaking down the utterance,
  - filling in the missing information (gaps), and
  - includes figuring out who did what to whom in each part of the sentence

- **encoded using a parse tree**

  (This tree is computed by the PAPPI system)

**gap filling and structural constraints**

part of the **knowledge of grammar (or language)**
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*

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
  
  fox+s → foxiz; cat+s → cats; dog+s → dogz; Bach+s → 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)
What is special about human language?
Ingredient for language:
Apes can label/categorize

But they can’t sing worth spit…
Ingredient for language:
Sophisticated vocal control and vocal learning
Which bird is this?

Juan ordéñame las cabras
`John milks the goats’

Domingo está enfermo
`Domingo was sick’

Silbo Gomero
What’s the missing ingredient?

\[
\begin{align*}
\text{Patrick} & \ + \ \text{saw Marvin} \\
\rightarrow \\
\text{Patrick} & \ \text{saw Marvin}
\end{align*}
\]

\[
\begin{align*}
\text{I said} & \ + \ \text{Patrick saw Marvin} \\
\rightarrow \\
\text{I said} & \ \text{Patrick saw Marvin} \\
= & \ \text{Cons } X, Y
\end{align*}
\]
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 \textit{cons}

Who could ask for anything more?
How is \textit{cons} implemented?

What makes a language a *natural* language?

- Language computes with **hierarchical structure** not by some
  - Example: *Obama likes him; Obama thinks that Paulson 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 case of Christopher

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

What do people do?
What does Christopher do?
The fMRI envelope please...

Better learning structure
Worse learning counting

○ = ‘counting’ rule language
● = ‘structure rule’ language

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

- (Linear) Association: beads on a string (‘bread and...’)
  
- (Recursively) Structural: nested boxes

Mary said Susan likes her
...Correspond to two distinct brain areas...

Ref: Frederici, PNAS, 2006
...Correspond to two distinct brain areas... vs. Hierarchical PSG vs FSG Broca’s Area

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
twasbrilligandtheslithytovesdidgyre

{pabiku, tibudo, daropi, golatu, ...}

pabikutibudodaropipabiku
tibudodaropitibudodaropi
pabikudaropipabikugolatu
tibudogolatutibudogolatu
golatudaropipabikutibudo
daropigolatudaropipabiku
tibudogolatudaropigolatu
daropigolatupabikutibudo
pabikutibudodaropigolatudaropi
...
Challenge: Combining Inference with Cognitive Constraints
(How real people solve real problems can help real computers)

Problem: \textit{twasbrilligandtheslithytovesdidgyreandgimble}

“Standard” solution: \textit{prettybaby} \hspace{1em} \textit{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” 
(\textit{Science}, 1996)

\[
\begin{align*}
Pr(bi|pa) &= 1.0; \quad Pr(ku|bi) = 1.0; \quad Pr(ti|ku) = 0.3, \\
Pr(bu|ti) &= 1; \quad Pr(do|bu) = 1.0; \quad Pr(da|do) = 0.3 \\
Pr(ro|da) &= 1; \quad Pr(pi|ro) = 1.0; \quad Pr(go|pi) = 0.3 \\
Pr(la|go) &= 1.0; \quad Pr(tu|la) = 1.0 \ldots
\end{align*}
\]

\textit{pigola} \rightarrow \textit{pi gola} \hspace{1em} \text{Works great? NO!!!}
Actual results on actual speech to children: works lousy
What’s the answer? Add a single, universal constraint…

What IS this universal constraint????

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

- precision = \[ \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \]
- recall = \[ \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \]
If computers are so smart, why can't they reed?

• Consider, for instance, the four letters *read*; they can be pronounced as either *reed* or *red*.
• How does the computer know in each case which is the correct pronunciation?

(1) The girls will read the paper. (*reed*)
(2) The girls have read the paper. (*red*)
(3) Will the girls read the paper? (*reed*)
(4) Have any men of good will read the paper? (*red*)
(5) Have the executors of the will read the paper? (*red*)
(6) Have the girls who will be on vacation next week read the paper yet? (*red*)
(7) Please have the girls read the paper. (*reed*)
(8) Have the girls read the paper? (*red*)
Modeling sentences: beads on a string?

*It was a bright cold day in April...*
N-grams: Unigrams

- Given a corpus of text, the n-grams are the sequences of n consecutive words that are in the corpus
- example (12 word sentence)
  - the cat that sat on the sofa also sat on the mat
- \( n=1 \) (8 unigrams)
  - the 3
  - sat 2
  - on 2
  - cat 1
  - that 1
  - sofa 1
  - also 1
  - mat 1
N-grams: Bigrams

- **example** (12 word sentence)
  - the cat that **sat on** the sofa also sat on the mat
- **N=2** (8 bigrams)
  - sat on 2
  - on the 2
  - the cat 1
  - cat that 1
  - that sat 1
  - the sofa 1
  - sofa also 1
  - also sat 1
  - the mat 1
N-grams: Trigrams

- example (12 word sentence)
  - the cat that sat on the sofa also sat on the mat
- N=3 (9 trigrams)
  - most language models stop here, some stop at quadrigrams
    - too many n-grams
    - low frequencies
  - sat on the 2
  - the cat that 1
  - cat that sat 1
  - that sat on 1
  - on the sofa 1
  - the sofa also 1
  - sofa also sat 1
  - also sat on 1
  - on the mat 1
N-grams: Quadrigrams

- Example: (12 word sentence)
  - the cat that sat on the sofa also sat on the mat
- N=4 (8 quadrigrams)
  - the cat that sat 1
  - cat that sat on 1
  - that sat on the 1
  - sat on the sofa 1
  - on the sofa also 1
  - the sofa also sat 1
  - sofa also sat on 1
  - also sat on the 1
  - sat on the mat 1
Language Models and $N$-grams

Bigram model of language

- Approximate the probability of $n^{th}$ word given $n-1$ previous words:
  - $P(w_n|w_1 \cdot w_2 \cdot ... \cdot w_{n-1})$ as
  - $P(w_n|w_{n-1})$

  i.e., the Markov assumption: conditioned on just the single preceding word

  (what would be the exact conditional probability? Why do we need to use this approximation?)
3 Steps: step 1, get estimate

Let $S = w_1, w_2, \ldots, w_n$ be a word sequence.
Given a training corpus, an intuitive estimate of the probability of the sequence $P(S)$, is the relative frequency of the string $w_1, w_2, \ldots, w_n$ in the corpus, the maximum likelihood estimate (MLE):

$$P_{MLE} = \frac{C(w_1, \ldots, w_n)}{N}$$

where $C(w_1, \ldots, w_n)$ is the frequency or count of the string $w_1, \ldots, w_n$ in the corpus and $N$ is the total number of strings of length $n$. 
Step 2, use chain rule

\[
P(S) = P(w_1, \ldots, w_n) \\
= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \ldots P(w_n|w_1, \ldots, w_{n-1}) \\
= \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1}) \\
P(S) = P(it) \times P(was|it) \times P(a|it, was) \times P(bright|it, was, a) \times \ldots \times \\
P(April|it, was, a, bright, cold, day, in)
\]
Step 3: use Markov assumption to limit ‘history’ length (context)

Make the Markov assumption, either order 1 (1 word history), or order 2 (2 word history):

\[
P(w_i | w_1, \ldots, w_{i-1}) \approx P(w_i | w_{i-1})
\]
\[
P(w_i | w_1, \ldots, w_{i-1}) \approx P(w_i | w_{i-2}, w_{i-1})
\]

trigram model:

\[
P(S) \approx P(it) \times P(was|it) \times P(a|it, was) \times P(bright|was, a) \times \ldots \times P(April|day, in)
\]
For next time (Monday)

- Do R&R #1 on GoogleTalk (3 short readings), and 2-page write up, email this to me at least 3 hrs before class; bring hardcopy to class
- Install NLTK (see instructions in R&R)
- Do background reading on NLTK, on words (nltk docs); in Jurafsky & Martin chapter 4
- For history and background, read ch. 1 of Jurafsky & Martin, 2nd edition, also linked on home page:

  http://www.cs.colorado.edu/%7Emartin/SLP/Updates/1.pdf
- See webpage for all lecture notes:

  After all, there might be some ambiguity...
Ambiguity

- 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
Ambiguity

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

- We haven’t had a sale in forty years