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

Professor Robert C. Berwick
berwick@csail.mit.edu
The Menu Bar

- The rules of the game:
  http://web.mit.edu/6.863/www/
- Mark Twain was right
- The Big Bang
- Natural language is not WYSIWYG
- We haven’t had a sale in 40 years
- For next time…

6.863J/9.611J SP07
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
Opps, you did it again! However much you love (or hate) her, you just can’t seem to get her name right.

Britney Spears that is – not Brittany or Britny or even Briney!

Click here to find out how Britney got her lucky break

Google, the search engine specialists, have found hundreds of different ways that people spell Britney.

The most popular way is Brittany, followed by Britney.

The misspellings were detected by Google’s spelling correction system.

The clever system counted each time users spelled her name wrong over a three-month period.
Heavy duty:
Natural language computer translation

Source text:
I want my next application to make as much money as Google

Translated text:
Quisiera que mi uso siguiente hiciera tanto dinero como Google
Quisiera que mi uso siguiente hiciera tanto dinero como Google

It wanted that my following use made as much money as Google
It wanted that my following use made as much money as Google.

Translated text:
Deseó que mi uso de siguiente hecho tanto dinero como Google.
Deseó que mi uso de siguiente hecho tanto dinero como Google

It wished that my use of following fact as much money as Google
<table>
<thead>
<tr>
<th>Translation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source text</td>
</tr>
<tr>
<td>Book them, Danno</td>
</tr>
<tr>
<td>Translated text</td>
</tr>
<tr>
<td>Resérvelos, Danno</td>
</tr>
</tbody>
</table>
Ingredient for language:
Apes can label/categorize

But they can’t sing worth spit...

6.863J/9.611J SP07
Fig. 5. Human and nonhuman animals exhibit the capacity to compute numerosities, including small precise number quantification and large approximate number estimation. Humans may be unique, however, in the ability to show open-ended, precise quantificational skills with large numbers, including the integer count list. In parallel with the faculty of language, our capacity for number relies on a recursive computation. [Illustration: John Yanson]
Ingredient for language:
Sophisticated vocal control and vocal learning
Mouse song is ultrasonic

Which bird is this?

Juan ordéñame las cabras
`John milks the goats’

Domingo está enfermo
`Domingo was sick’

Silbo Gomero
Evolution of size and complexity of the human brain

Figure 1: Differences in cerebral cortical size are associated with differences in the cerebral cortex circuit diagram. The top panel shows side views of the brain of a rodent (mouse), a chimpanzee and a human to show relative sizes. The middle panel shows a cross-section of a human and chimpanzee brain, with the cellular composition of the cortex illustrated in the bottom panel (adapted from ref. 3). The cerebral cortex derives from two developmental cell populations: the primordial plexiform layer (PPL) and the cortical plate (CP). The primordial plexiform layer seems to be homologous to simple cortical structures in Amphibia and Reptilia, and appears first temporally during mammalian brain development. The cortical plate develops as a second population that splits the primordial plexiform layer into two layers (layer I at the top and the subplate (SP) at the bottom; numbering follows the scheme of ref. 31). Cortical-plate-derived cortical layers are added developmentally from deeper first (VI, V) to more superficial (III, II) last. Cortical-plate-derived cortical layers are progressively elaborated in mammals with larger brains (for example, insectivores have a single layer II/III/IV that is progressively subdivided into II, III, IV, then IIa, IIb, and so on), so that humans have a larger proportion of these latter-derived neurons, which project locally or elsewhere within the cortex. Images from the top and middle panels are from the Comparative Brain Atlas (http://www.brainmuseum.org).

Room with a View?
Language

Sensory-motor interface (external world)

Semantic (conceptual-intensional) interface (internal world)
What’s the missing ingredient?

Patrick + saw Marvin

I said + Patrick saw Marvin = Cons X,Y

6.863J/9.611J SP07
Do kids/adults use statistics? (to learn language; to use language)

- Two examples: low-level and high-level
- Finding word boundaries in speech
- Figuring out what words mean
What makes natural language challenging and interesting: it is not ‘wysiwyg’

- What you see on the surface is not necessarily what is happening ‘beneath’ or ‘in your head’
- Natural Language is ambiguous
- Not Wysiwyg: ‘fillers’ and ‘gaps’
  - Invisible elements – shared knowledge – window into the mind
  - John is too stubborn to talk to
  - John is too stubborn to talk to Bill
  - John is too stubborn 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 stubborn to talk to
The Problem of Language Understanding

• An example
  • John is too stubborn to talk to

  • \textit{talk to}: predicate
  • two arguments: (talker) (“talkee”)
  • \textit{neither argument is explicitly identified in the sentence}
  • talker = “arbitrary person”
  • talkee = John

\texttt{talk\_to(X,John)}
The Problem of Language Understanding

- An example
  - John is too stubborn to talk to $\text{talk_to}(X,\text{John})$
  - John is too stubborn to talk to $\text{Bill}$
    - talker = John
    - talkee = Bill
      - $\text{talk_to}(\text{John},\text{Bill})$
The Problem of Language Understanding

• An example
  John is too stubborn to talk to talk_to(X,John)

  John is too stubborn to talk to Bill talk_to(John,Bill)

• A linguistic explanation
  missing subject of an infinitival clause
  should be co-referential with next higher referring expression
  relation is blocked by intervening operator element
  *this explanation can be supported by other data/constructions*

  “linguistic principle”
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 - which we shall demo)

- gap filling and structural constraints part of the **knowledge of grammar (or language)**

6.863J/9.611J SP07
The Problem of Language Understanding

• applications of **deep parsing** are many...

• **Machine Translation**

• **Text Understanding**
  • or anything that might require sentence or question comprehension
    • Dialogue Understanding
    • Intelligent Interactive Systems
    • Question-Answering (QA)
    • etc.
Stochastic Methods

- **Resources**
  - treebanks
    - pre-parsed corpora
  - large amounts of labeled text
    - part-of-speech labeling

- **Typical Properties**
  - shallower analyses
  - quick, cheap and adaptable
    - use machine learning methods
  - win competitions *(bake-offs)*

- **Statistical Parsing**
- **Statistical Machine Translation**
  - exploit bilingual corpora
  - can be “syntax-based”
  - **Language Weaver** *(2004, 2005)*
  - **Google** *(forthcoming?)*
    - winner of the 2005 NIST Machine Translation Evaluation
Not so fast…?
Challenges…

- Challenge with ‘deep’ understanding (‘who did what to whom’)
- Challenge with shallow syntax
- Challenge in moving to other languages
- Challenge with translation
- Challenge of how kids learn so fast and so easily (and cognitive questions generally)
- Challenge of word meaning

...
Challenge: The Problem of Deep Language Analysis

- Sentence/Question Understanding

“Which proposal did you say was too illogical to consider without redrafting?”
An Example

- consider the question:
  - Which report did you file without reading? =
  - Which report did you file [that report] without [you] reading [that report]
- 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?
The Problem of Language Analysis

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

<table>
<thead>
<tr>
<th>which</th>
<th>proposal</th>
<th>did</th>
<th>you</th>
<th>say</th>
<th>is</th>
<th>too</th>
<th>illogical</th>
<th>to</th>
<th>consider</th>
<th>without</th>
<th>redrafting</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>utterance</td>
<td></td>
<td></td>
<td>sayer</td>
<td></td>
<td></td>
<td>topic</td>
<td></td>
<td>action</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R-A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>of</td>
<td></td>
<td>comment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R-A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

perhaps **state-of-the-art** performance but performance on this task is poor
The Problem of Language Analysis

(This tree is computed by the PAPPI system)

you=[2]=redrafter
proposal=[1]=item to be redrafted
you=[2]=thinker/considerer
proposal=[1]=item to be considered
proposal=[1]=too illogical

you=[2]=“sayer”
which proposal=[1]
you=[2]
which report did you file without reading

One parse found
The **semantic** interface seems to take functional priority in language design.

**did Patrick see what**

**what did Patrick see what**

**Sensory-motor interface**

**what did Patrick see …**

(what you **speak** or **hear**)

**Semantic interface**

**For what x, did Patrick see x**

(what you **think** inside your head)
Fast forward 40 years
Prolog:

[3] canan okuldan eve geldi


Loaded comments for
Completed loading files for language Tur
To interleave operations: licenseClausalArguments, licenseAdjuncts, sBarDeletion, coindexSubjAndINFL, whInSyntax
licenseClausalArguments interleaved for categories: pp, v1, vp, ap, np, i2
licenseAdjuncts interleaved for categories: ap, np, vp
sBarDeletion interleaved for categories: v1, vp
coin dexSubjAndINFL interleaved for categories: i2
whInSyntax interleaved for categories: c2
sub jacency merged wrt: chainFormation
Completed loading parser file
/Applications/pappi2.app/Contents/MacOS/pappi2/j5parser.pl (11/05/04 00:20:06)
Parsing: [3] canan okul dan eve geldi
LF (1):

```
  C2
 /___\  
|     |
I2    C
|     |
NP[1]  I1
```

Filters

- Theta Criterion
- D-structure Theta Condition
- Wh-movement in Syntax
- Case Filter
- Case Condition on ECs
- Coindex Subject
- Condition A
- Condition B
- Condition C
- ECP
- Control
- License Clitics
- License Object pro
- ECP at LF
- Fl: License operator/variable
- Fl: Quantifier Scoping
- Fl: Reanalyze Bound Pro
- License Clausal Arguments
- License Syntactic Adjoin