Course logistics in brief

- Instructor: Christopher Manning
- TAs: Paul Baumstarck, Pichuan Chang
- Time: MW 11:00–12:15. (Section: ?? F 11:00–12:15)
- Handouts:
  - Course syllabus, lecture 1, assignment 1
  - Programming language: Java 1.5+
  - Other information: see the webpage.
    - http://cs224n.stanford.edu/

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, 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

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)

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 XML, or menus and drop boxes, or …
- But someone has to work on the hard problems!
  - How can we tell computers about language?
  - Or help them learn it as kids do?
- In this course we seek to identify many of the open research problems in natural language

What/where is NLP?

- Goals can be very far reaching …
  - True text understanding
  - Reasoning about texts
  - Real-time participation in spoken dialogs
- Or very down-to-earth …
  - Finding the price of products on the web
  - Context sensitive spell-checking
  - Analyzing reading level or authorship statistically
  - Extracting facts or relations from documents
- These days, the latter predominate (as NLP becomes increasingly practical, it is increasingly engineering-oriented – also related to changes in approach in AI/NLP)
The hidden structure of language

- We’re going beneath the surface...
  - Not just string processing
  - Not just keyword matching in a search engine
  - Search Google on ‘tennis racquet’ and ‘tennis racquets’ and the results you get are completely different!
  - Search Google on ‘laptop’ and ‘notebook’ and the results you get are completely different!
  - Not just converting a sound stream to a string of words
    - Like Nuance, IBM’s Watson, Philips speech recognition
  - We want to recover and manipulate at least some aspects of 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

- Early NLP (Machine Translation) on machines less powerful than pocket calculators
- Foundational work on automata, formal languages, probabilities, and information theory
- First speech systems (Davis et al., Bell Labs)
- MT heavily funded by military, but basically just word substitution programs
- Little understanding of natural language syntax, semantics, pragmatics
- Problem soon appeared intractable

Why is NLU difficult? The hidden structure of language is hugely ambiguous

- Structures for: Fed raises interest rates 0.5% in effort to control inflation (NYT headline 17 May 2000)

Where are the ambiguities?

Part of speech ambiguities

Syntactic attachment ambiguities

Word sense ambiguities: Fed → “federal agent” interest → a feeling of wanting to know or learn more

Semantic interpretation ambiguities above the word level
Why NLP is difficult: 
Newspaper headlines

- Ban on Nude Dancing on Governor’s Desk
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals Are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks
- Minister Accused Of Having 8 Wives In Jail

LSAT / (former) GRE 
Analytic Section Questions

- Six sculptures - C, D, E, F, G, H - are to be exhibited in rooms 1, 2, and 3 of an art gallery.
  - Sculptures C and E may not be exhibited in the same room.
  - Sculptures D and G must be exhibited in the same room.
  - If sculptures E and F are exhibited in the same room, no other sculpture may be exhibited in that room.
  - At least one sculpture must be exhibited in each room, and no more than three sculptures may be exhibited in any room.
  - If sculpture D is exhibited in room 3 and sculptures E and F are exhibited in room 1, which of the following may be true?
    A. Sculpture C is exhibited in room 1
    B. Sculpture H is exhibited in room 1
    C. Sculpture G is exhibited in room 2
    D. Sculptures C and H are exhibited in the same room
    E. Sculptures G and F are exhibited in the same room

Reference Resolution

U: Where is A Bug's Life playing in Mountain View?
S: A Bug's Life is playing at the Century 16 theater.
U: When is it playing there?
S: It's playing at 2pm, 5pm, and 8pm.
U: I'd like 1 adult and 2 children for the first show. How much would that cost?

- Knowledge sources:
  - Domain knowledge
  - Discourse knowledge
  - World 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
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
    - \( P(\text{maison} \rightarrow \text{house}) \) high
    - \( P(\text{L’avocat général} \rightarrow \text{the general avocado}) \) low
- Some computer scientists think this is a new “A.I.” idea
  - But really it’s an old idea that was stolen 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!)

Machine Translation

The classic acid test for natural language processing.
Requires capabilities in both interpretation and generation.
About $10 billion spent annually on human translation.

Mainly slides from Kevin Knight (at ISI)

Translation (human and machine)

Ref: According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959 billion US dollars of foreign capital, including 40.007 billion US dollars of direct investment from foreign businessmen.

IBM4: the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and

Yamada/Knight: today’s available data of the Ministry of Foreign Trade and Economic Cooperation shows that china’s actual utilization of November this year will include 40.007 billion US dollars for the foreign direct investment among 46.959 billion US dollars in foreign capital

Machine Translation History

- 1950s: Intensive research activity in MT
- 1960s: Direct word-for-word replacement
- 1966 (ALPAC): NRC Report on MT
  - Conclusion: MT no longer worthy of serious scientific investigation.
- 1975-1985: Resurgence (Europe, Japan)
  - Domain specific rule-based systems
- 1985-1995: Gradual Resurgence (US)
- 1995-2005: Statistical MT surges ahead

What happened between ALPAC and Now?

- Need for MT and other NLP applications confirmed
- Change in expectations
- Computers have become faster, more powerful
- WWW
- Political state of the world
- Maturation of Linguistics
- Availability of data
- Development of statistical and hybrid statistical/symbolic approaches

**Three MT Approaches: Direct, Transfer, Interlingual** (Vauquois triangle)

- **Interlingua**
- **Semantic Structure**
  - Translation
- **Syntactic Structure**
  - Word Structure
- **Morphological Analysis**
  - Source Text
  - Target Text

**Statistical Solution**

- **Parallel Texts**
  - Rosetta Stone
    - Hieroglyphs
    - Demotic
    - Greek

- **Statistical Solution**
  - Instruction Manuals
  - Hong Kong Legislation
  - Macao Legislation
  - Canadian Parliament Hansards
  - United Nations Reports
  - Official Journal of the European Communities

- **Warren Weaver**
  - "Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'"
  - Warren Weaver (1955:18, quoting a letter he wrote in 1947)

**Centauri/Arcturan [Knight, 1997]**

- **Your assignment, translate this to Arcturan**
  - fonsk emrok bihok yorok clok kamok ok-ysyp

- **Centauri/Arcturan [Knight, 1997]**

  | 1a. ok-oom emerok sprok | 7. lalok fonsk emerok halok sprok izok enemok . |
  | 1b. ok-oom bichat dat | 7b. uot jot bichat uot dat uat amers . |
  | 2a. ok-oom emerok plek sprok | 8. lalok trook amok plek nok . |
  | 2b. ok-oom bichat dat uat amers | 8b. uat jot sprok rook uat amers . |
  | 3a. fonsk emerok izok hihok ghirok | 9a. hmof nok izok kamok ok-ysyp . |
  | 3b. totat dat uat vut hikat | 9b. totat amers uat violet at-ysyp . |
  | 4a. ok-oom oomok oomok jok | 10a. lalok trook rook yorok clok . |
  | 4b. at-oomok kimok pai kip | 10b. totat amers uat vat hikat . |
  | 5a. fonsk emerok izok hihok | 11a. lalok meck meck yorok clok . |
  | 5b. totat jot uat amers | 11b. totat amers uat vat hikat . |
  | 6a. lalok emerok izok hihok ghirok | 12a. lalok trook meck izok hihok clok . |
  | 6b. totat amers uat vat hikat . | 12b. totat amers uat vat hikat . |
### Centauri/Arcturan [Knight, 1997]

#### Your assignment, translate this to Arcturan:

<table>
<thead>
<tr>
<th>English</th>
<th>Arcturan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-voon eremok sprok.</td>
<td>7a. lakul farok eremok lakul sprok izok enmekok.</td>
</tr>
<tr>
<td>1b. at-voon bichat dat.</td>
<td>7b. wat just bichat wat dat tar enemok.</td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plek sprok.</td>
<td>8a. lakul brok anok plek nok.</td>
</tr>
<tr>
<td>2b. at-drubel at-voon pippat mat dat.</td>
<td>8b. lat pippat mat tar nont.</td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghiruk.</td>
<td>9a. sutolok nek izok karmaok ok-eryap.</td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat.</td>
<td>9b. totat mat tar enemok ok-eryap.</td>
</tr>
<tr>
<td>4a. ok-voon anok brok jok.</td>
<td>10a. lakul nek izok yolkok ghiruk cl i klok.</td>
</tr>
<tr>
<td>4b. at-voon leen pippat sat lat.</td>
<td>10b. wat mat gat mat tar bat hilat.</td>
</tr>
<tr>
<td>5a. wiwok farok toxk sprok.</td>
<td>11a. lakul nek eremok hihok yolkok zanenmekok.</td>
</tr>
<tr>
<td>5b. totat jot quartz mat.</td>
<td>11b. wat mat arrat mat zanenmekok.</td>
</tr>
<tr>
<td>6a. lakul sprok toxk jot toxk.</td>
<td>12a. lakul relok toxk izok hihok molok.</td>
</tr>
<tr>
<td>6b. wat dat jot quartz mat.</td>
<td>12b. wat mat foret arrat vat gat.</td>
</tr>
</tbody>
</table>

#### Centauri/Arcturan [Knight, 1997]

<table>
<thead>
<tr>
<th>English</th>
<th>Arcturan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-voon eremok sprok.</td>
<td>7a. lakul farok eremok lakul sprok izok enmekok.</td>
</tr>
<tr>
<td>1b. at-voon bichat dat.</td>
<td>7b. wat just bichat wat dat tar enemok.</td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plek sprok.</td>
<td>8a. lakul brok anok plek nok.</td>
</tr>
<tr>
<td>2b. at-drubel at-voon pippat mat dat.</td>
<td>8b. lat pippat mat tar nont.</td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghiruk.</td>
<td>9a. sutolok nek izok karmaok ok-eryap.</td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat.</td>
<td>9b. totat mat tar enemok ok-eryap.</td>
</tr>
<tr>
<td>4a. ok-voon anok brok jok.</td>
<td>10a. lakul nek izok yolkok ghiruk cl i klok.</td>
</tr>
<tr>
<td>4b. at-voon leen pippat sat lat.</td>
<td>10b. wat mat gat mat tar bat hilat.</td>
</tr>
<tr>
<td>5a. wiwok farok toxk sprok.</td>
<td>11a. lakul nek eremok hihok yolkok zanenmekok.</td>
</tr>
<tr>
<td>5b. totat jot quartz mat.</td>
<td>11b. wat mat arrat mat zanenmekok.</td>
</tr>
<tr>
<td>6a. lakul sprok toxk jot toxk.</td>
<td>12a. lakul relok toxk izok hihok molok.</td>
</tr>
<tr>
<td>6b. wat dat jot quartz mat.</td>
<td>12b. wat mat foret arrat vat gat.</td>
</tr>
</tbody>
</table>
Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia y asociados.
1b. Garcia is a company.
2a. Garcia tiene tres asociados.
2b. Garcia has three associates.
3a. Los grupos son fuertes.
3b. The groups are strong.
4a. Los grupos venden medicinas.
4b. The groups sell strong pharmaceuticals.
5a. Los grupos son fuertes.
5b. The groups are strong.
6a. Los grupos venden medicinas.
6b. The groups sell strong pharmaceuticals.
7a. Los asociados son fuertes.
7b. The associates are strong.
8a. Los asociados son fuertes.
8b. The associates are strong.
9a. Los grupos venden medicinas.
9b. The groups sell strong pharmaceuticals.
10a. Los grupos son fuertes.
10b. The groups are strong.
11a. Los asociados son fuertes.
11b. The associates are strong.
12a. Los grupos venden medicinas.
12b. The groups sell strong pharmaceuticals.
Speech Recognition: Acoustic Waves

- Human speech generates a wave
  - like a loudspeaker moving

- A wave for the words "speech lab" looks like:

  ![Waveform Image]

  "I" to "a" transition:

  ![Transition Image]

Acoustic Sampling

- 10 ms frame (ms = millisecond = 1/1000 second)
- ~25 ms window around frame [wide band] to allow/smooth signal processing – it let’s you see formants

  ![Sampling Image]

Result:

Acoustic Feature Vectors

(after transformation, numbers in roughly $R^{10}$)

Spectral Analysis

- Frequency gives pitch; amplitude gives volume
  - sampling at ~8 kHz phone, ~16 kHz mic (kHz=1000 cycles/sec)

- Fourier transform of wave displayed as a spectrogram
  - darkness indicates energy at each frequency
  - hundreds to thousands of frequency samples

The Speech Recognition Problem

- The Recognition Problem: Noisy channel model
  - We started out with English words, they were encoded as an audio signal, and we now wish to decode.
  - Find most likely sequence $w$ of “words” given the sequence of acoustic observation vectors $a$
  - Use Bayes’ rule to create a generative model and then decode
    
    \[ \text{ArgMax}_w \ P(w | a) = \text{ArgMax}_w \ P(a | w) \ P(w) / P(a) \]

- Acoustic Model: $P(a | w)$
- Language Model: $P(w)$

A probabilistic theory of a language

Probabilistic Language Models

- Assign probability $P(w)$ to word sequence $w = w_1, w_2, ..., w_k$
- Can’t directly compute probability of long sequence – one needs to decompose it
- Chain rule provides a history-based model:
  
  \[ P(w_1, w_2, ..., w_k) = P(w_1) \ P(w_2 | w_1) \ P(w_3 | w_1, w_2) \cdots P(w_k | w_1, ..., w_{k-1}) \]

- Cluster histories to reduce number of parameters
- E.g., just based on the last word (1st order Markov model):
  
  \[ P(w_1, w_2, ..., w_k) = P(w_1 | <s>) \ P(w_2 | w_1) \ P(w_3 | w_2) \cdots P(w_k | w_{k-1}) \]

- How do we estimate these probabilities?
  - We count word sequences in corpora
  - We “smooth” probabilities so as to allow unseen sequences

N-gram Language Modeling

- n-gram assumption clusters based on last n-1 words
  - unigrams = $P(w_i)$
  - bigrams = $P(w_i | w_{i-1})$
  - trigrams = $P(w_i | w_{i-2}, w_{i-1})$

- Trigrams often interpolated with bigram and unigram:
  
  \[ \hat{P}(w_i | w_{i-2}, w_{i-1}) = \lambda \ P(w_i | w_{i-2}, w_{i-1}) + \lambda \ P(w_i | w_{i-1}) + \lambda \ P(w_i) \]

  - the $\lambda$ typically estimated by maximum likelihood estimation on held out data ($P(\cdot | \cdot)$ are relative frequencies)
  - many other interpolations exist (another standard is a non-linear backoff)