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

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 arguments
  - Real-time participation in spoken dialogs
- 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 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)
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.

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

<table>
<thead>
<tr>
<th>VB</th>
<th>VBP</th>
<th>VBJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>NNS</td>
<td>NN</td>
</tr>
<tr>
<td>Fed raises interest rates 0.5% in effort to control inflation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

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
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." 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!)

Daily Question!

- What is the ambiguity in this (authentic!) newspaper headline?

Minister Accused Of Having 8 Wives In Jail

Translation (human and machine)

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 businesses.

IBM:
- 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 this year 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

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an email 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 $10 billion spent annually on human translation.

Mainly slides from Kevin Knight (at ISI)

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.
- 1966-1975: ’Recovery period’
- 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

Statistical Solution

- Parallel Texts
  - Rosetta Stone

  Hieroglyphs
  Demotic
  Greek

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 (1965-18, quoting a letter he wrote in 1947)

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp
### Centauri/Arcturan [Knight, 1997]

<table>
<thead>
<tr>
<th>1a. ok-voom oorreum sprlok</th>
<th>7a. latok farok oorreum latok sprlok izok emenemok</th>
<th>6a. lalok sprok izok jok stok</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b. at-voom bechot dat</td>
<td>7b. wat ju bechot wat dat vat emenok</td>
<td>5b. totat jjat quat cat</td>
</tr>
<tr>
<td>2a. ok-drubel at-voom amok pleok sprlok</td>
<td>8a. latok brok amok pleok nok</td>
<td>5a. wiwok farok izok stok</td>
</tr>
<tr>
<td>2b. at-drubel at-voom pippad rat dat</td>
<td>8b. lat lat pippad rat mat</td>
<td>4a. ok-voon anok drok brok jok</td>
</tr>
<tr>
<td>3a. erok sprlok izok bibok grihok</td>
<td>9a. mitok nok izok kantok ok-yaap</td>
<td>4b. at-drubel at-voon pippad rat dat</td>
</tr>
<tr>
<td>3b. wat dat arrat vat bila</td>
<td>9b. sent muul gaat elat ok-yaap</td>
<td>3a. erok sprok izok bibok emenemok</td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok</td>
<td>10a. latok mok nok yorok ghirok clok</td>
<td>3b. totat dat arrat vat hilat</td>
</tr>
<tr>
<td>4b. at-voon kent pippad rat lat</td>
<td>10b. wat muat gat mat bat bila</td>
<td>2a. ok-drubel ok-voon bichat wat dat vat eneat</td>
</tr>
<tr>
<td>5a. wiwok farok izok kantok ok-yaap</td>
<td>11a. latok mok nok yorok ghirok clok</td>
<td>2b. at-drubel at-voon pippad rat dat</td>
</tr>
<tr>
<td>5b. totat jjat quat cat</td>
<td>11b. sent muul gaat elat ok-yaap</td>
<td>1a. ok-voon oorreum sprlok</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok</td>
<td>12a. latok mok nok yorok ghirok clok</td>
<td>1b. at-voon bichat dat</td>
</tr>
<tr>
<td>6b. wat dat kent quart cat</td>
<td>12b. wat muat front arrat vat gat</td>
<td>1a. ok-voon ororok sprok</td>
</tr>
</tbody>
</table>

---

### Centauri/Arcturan [Knight, 1997]

<table>
<thead>
<tr>
<th>1a. ok-voom oorreum sprlok</th>
<th>7a. latok farok oorreum latok sprlok izok emenemok</th>
<th>6a. lalok sprok izok jok stok</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b. at-voom bechot dat</td>
<td>7b. wat ju bechot wat dat vat emenok</td>
<td>5b. totat jjat quat cat</td>
</tr>
<tr>
<td>2a. ok-drubel at-voom amok pleok sprlok</td>
<td>8a. latok brok amok pleok nok</td>
<td>5a. wiwok farok izok stok</td>
</tr>
<tr>
<td>2b. at-drubel at-voom pippad rat dat</td>
<td>8b. lat lat pippad rat mat</td>
<td>4a. ok-voon anok drok brok jok</td>
</tr>
<tr>
<td>3a. erok sprlok izok bibok grihok</td>
<td>9a. mitok nok izok kantok ok-yaap</td>
<td>4b. at-drubel at-voon pippad rat dat</td>
</tr>
<tr>
<td>3b. wat dat arrat vat bila</td>
<td>9b. sent muul gaat elat ok-yaap</td>
<td>3a. erok sprok izok bibok emenemok</td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok</td>
<td>10a. latok mok nok yorok ghirok clok</td>
<td>3b. totat dat arrat vat hilat</td>
</tr>
<tr>
<td>4b. at-voon kent pippad rat lat</td>
<td>10b. wat muat gat mat bat bila</td>
<td>2a. ok-drubel ok-voon bichat wat dat vat eneat</td>
</tr>
<tr>
<td>5a. wiwok farok izok kantok ok-yaap</td>
<td>11a. latok mok nok yorok ghirok clok</td>
<td>2b. at-drubel at-voon pippad rat dat</td>
</tr>
<tr>
<td>5b. totat jjat quat cat</td>
<td>11b. sent muul gaat elat ok-yaap</td>
<td>1a. ok-voon oorreum sprlok</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok</td>
<td>12a. latok mok nok yorok ghirok clok</td>
<td>1b. at-voon bichat dat</td>
</tr>
<tr>
<td>6b. wat dat kent quart cat</td>
<td>12b. wat muat front arrat vat gat</td>
<td>1a. ok-voon ororok sprok</td>
</tr>
</tbody>
</table>

---

### Centauri/Arcturan [Knight, 1997]

<table>
<thead>
<tr>
<th>1a. ok-voom oorreum sprlok</th>
<th>7a. latok farok oorreum latok sprlok izok emenemok</th>
<th>6a. lalok sprok izok jok stok</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b. at-voom bechot dat</td>
<td>7b. wat ju bechot wat dat vat emenok</td>
<td>5b. totat jjat quat cat</td>
</tr>
<tr>
<td>2a. ok-drubel at-voom amok pleok sprlok</td>
<td>8a. latok brok amok pleok nok</td>
<td>5a. wiwok farok izok stok</td>
</tr>
<tr>
<td>2b. at-drubel at-voom pippad rat dat</td>
<td>8b. lat lat pippad rat mat</td>
<td>4a. ok-voon anok drok brok jok</td>
</tr>
<tr>
<td>3a. erok sprlok izok bibok grihok</td>
<td>9a. mitok nok izok kantok ok-yaap</td>
<td>4b. at-drubel at-voon pippad rat dat</td>
</tr>
<tr>
<td>3b. wat dat arrat vat bila</td>
<td>9b. sent muul gaat elat ok-yaap</td>
<td>3a. erok sprok izok bibok emenemok</td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok</td>
<td>10a. latok mok nok yorok ghirok clok</td>
<td>3b. totat dat arrat vat hilat</td>
</tr>
<tr>
<td>4b. at-voon kent pippad rat lat</td>
<td>10b. wat muat gat mat bat bila</td>
<td>2a. ok-drubel ok-voon bichat wat dat vat eneat</td>
</tr>
<tr>
<td>5a. wiwok farok izok kantok ok-yaap</td>
<td>11a. latok mok nok yorok ghirok clok</td>
<td>2b. at-drubel at-voon pippad rat dat</td>
</tr>
<tr>
<td>5b. totat jjat quat cat</td>
<td>11b. sent muul gaat elat ok-yaap</td>
<td>1a. ok-voon oorreum sprlok</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok</td>
<td>12a. latok mok nok yorok ghirok clok</td>
<td>1b. at-voon bichat dat</td>
</tr>
<tr>
<td>6b. wat dat kent quart cat</td>
<td>12b. wat muat front arrat vat gat</td>
<td>1a. ok-voon ororok sprok</td>
</tr>
</tbody>
</table>
Your assignment, translate this to Arcturan:

6b. wat dat krat quat cat.

6a. lalok sprok izok jok stok.

5b. totat jjat quat cat.

5a. wiwok farok izok stok.

4b. at-voon krat pippat sat lat.

4a. ok-voon anok drok brok jok.

3b. totat dat arrat vat hilat.

3a. erok sprok izok hihok ghirok.

2b. at-drubel at-voon pippat rrat dat.

2a. ok-drubel ok-voon anok plok sprok.

1b. at-voon ororok sprok.

1a. ok-voon ororok sprok.

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan:

12b. wat nnat forat arrat vat gat.

11b. wat nnat arrat mat zanzanat.

11a. lalok nok crrrok hihok yorok zanzanok.

10a. lalok mok nok yorok ghirok clok.

9b. totat nnat quat oloat at-yurp.

9a. wiwok nok izok kantok ok-yurp.

8b. iat lat pippat rrat nnat.

8a. lalok brok anok plok nok.

7b. wat jjat bichat wat dat vat eneat.

7a. lalok farok ororok lalok sprok izok enemok.

6b. wat nnat forat arrat vat gat.

6a. lalok rarok nok izok hihok mok.

process of elimination
It's Really Spanish/English

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clients do not sell pharmaceuticals in Europe</td>
<td>Clientes no venden medicinas en Europa</td>
</tr>
<tr>
<td>Garcia and associates</td>
<td>García y asociados</td>
</tr>
<tr>
<td>Carlos Garcia has three associates</td>
<td>Carlos García tiene tres asociados</td>
</tr>
<tr>
<td>Los asociados no son fuertes</td>
<td>Los grupos modernos venden medicinas fuertes</td>
</tr>
<tr>
<td>Los asociados son enemigos</td>
<td>Los asociados son enemigos</td>
</tr>
<tr>
<td>Los grupos no venden zenzanine</td>
<td>Los grupos no venden zenzanine</td>
</tr>
<tr>
<td>Los clientes están enfadados</td>
<td>Los asociados también están enfadados</td>
</tr>
</tbody>
</table>

Speech Recognition: Acoustic Waves

- Human speech generates a wave – like a loudspeaker moving
- A wave for the words “speech lab” looks like:

Spectral Analysis

- Frequency gives pitch; amplitude gives volume
- Fourier transform of wave displayed as a spectrogram – darkness indicates energy at each frequency
- Hundreds to thousands of frequency samples

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

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
    \[
    \text{ArgMax}_w P(w|a) = \text{ArgMax}_w P(a|w) P(w) / P(a) \\
    = \text{ArgMax}_w P(a|w) P(w)
    \]
- Acoustic Model: \( P(a|w) \)
- Language Model: \( P(w) \)

Probabilistic Language Models

- Assign probability \( P(w) \) to word sequence \( w = w_1, w_2, \ldots, w_k \)
- Can’t directly compute probability of long sequence – one needs to decompose it
- Chain rule provides a history-based model:
  \[
  P(w_1, w_2, \ldots, w_k) = P(w_1) P(w_2|w_1) P(w_3|w_1, w_2) \cdots P(w_k|w_1, \ldots, w_{k-1})
  \]
- Cluster histories to reduce number of parameters
  - E.g., just based on the last word (1st order Markov model):
  \[
  P(w_1, w_2, \ldots, w_k) = P(w_1) P(w_2|w_1) P(w_3|w_1, w_2) \cdots P(w_k|w_{k-1})
  \]
- How do we estimate these probabilities?
  - We count word sequences in corpora
  - We “smooth” probabilities so as to allow unseen sequences