CSE 517: Natural Language Processing

New Quals Course!
Instructor: Luke Zettlemoyer
Winter 2013

Slides adapted from Dan Klein
What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!

- End systems that we want to build:
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Unknown: human-level comprehension (is this just NLP?)
Speech Systems

- **Automatic Speech Recognition (ASR)**
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV

```
   -10000  -5000   0    5000  10000
0 1.00 1.20 1.40 1.60 1.80 2.00 2.20
```

- **Text to Speech (TTS)**
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)
Information Extraction

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

<table>
<thead>
<tr>
<th>Person</th>
<th>Company</th>
<th>Post</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times newspaper</td>
<td>president and general manager</td>
<td>start</td>
</tr>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times newspaper</td>
<td>executive vice president</td>
<td>end</td>
</tr>
<tr>
<td>Lance R. Primis</td>
<td>New York Times Co.</td>
<td>president and CEO</td>
<td>start</td>
</tr>
</tbody>
</table>

- **SOTA:** perhaps 80% accuracy for multi-sentence temples, 90%+ for single easy fields
- **But remember:** information is redundant!
QA / NL Interaction

Question Answering:
- More than search
- Can be really easy: “What’s the capital of Wyoming?”
- Can be harder: “How many US states’ capitals are also their largest cities?”
- Can be open ended: “What are the main issues in the global warming debate?”

Natural Language Interaction:
- Understand requests and act on them
- “Make me a reservation for two at Quinn’s tonight”

Example:
- Search query: “How many US states’ capitals are also their largest cities?”
- Google search result:
  - No matches found.
  - Suggestions:
    - Make sure all words are spelled correctly.
    - Try different keywords.
    - Try more general keywords.
    - Try fewer keywords.

External links:
- capital of Wyoming: Information From Answers.com
- Cheyenne: Weather and Much More From Answers.com
Oscar for best actress 1958

Input interpretation:

<table>
<thead>
<tr>
<th>Academy Awards</th>
<th>actress in a leading role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1958 (year of award ceremony)</td>
<td></td>
</tr>
</tbody>
</table>

Result:

Joanne Woodward in The Three Faces of Eve

Other nominees:

- Lana Turner in Peyton Place
- Elizabeth Taylor in Raintree County
- Deborah Kerr in Heaven Knows, Mr. Allison
- Anna Magnani in Wild Is the Wind

Information about Joanne Woodward:

<table>
<thead>
<tr>
<th>full name</th>
<th>Joanne Gignilliat Trimmier Woodward</th>
</tr>
</thead>
<tbody>
<tr>
<td>date of birth</td>
<td>Thursday February 27, 1930 (age: 82 years)</td>
</tr>
<tr>
<td>place of birth</td>
<td>Thomasville, Georgia, United States</td>
</tr>
</tbody>
</table>

Academy Awards and nominations:
Summarization

- Condensing documents
  - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative

- Very context-dependent!

- An example of analysis with generation
Machine Translation

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (next class) vs fidelity (later)
"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959
Business

Latest News

- The exchange of financial stocks fell slightly prominent lower
  12 stocks in Tokyo, ahead of sell orders from the backlash of higher yesterday, with slightly lower values. Nikkei ... ... (11:13) [Full article]

- Negotiation and integration of Japan
  Sompo Japan 興亜 to aggregate in three large camps

Sompo Japan Insurance and it's five to start the negotiations for the merger of NIPPOKOA Insurance Co., Ltd. No. 12, 2007, minutes ... ... (10:33) [Full article]
"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake", because he was:

(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery
US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.
(1) Colorless green ideas sleep furiously.
(2) Furiously sleep ideas green colorless
   It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not." (Chomsky 1957)

70s and 80s: more linguistic focus
   Emphasis on deeper models, syntax and semantics
   Toy domains / manually engineered systems
   Weak empirical evaluation
NLP: machine learning and empiricism

“Whenever I fire a linguist our system performance improves.” –Jelinek, 1988

- **1990s: Empirical Revolution**
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential

- **2000s: Richer linguistic representations used in statistical approaches, scale to more data!**

- **2010s: you decide!**
What is Nearby NLP?

- **Computational Linguistics**
  - Using computational methods to learn more about how language works
  - We end up doing this and using it

- **Cognitive Science**
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!

- **Speech?**
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP
Problem: Ambiguities

- **Headlines:**
  - Enraged Cow Injures Farmer with Ax
  - Ban on Nude Dancing on Governor’s Desk
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half

- Why are these funny?
Hurricane Emily howled toward Mexico's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun, where frightened tourists squeezed into musty shelters.

- **SOTA:** ~90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples
Semantic Ambiguity

At last, a computer that understands you like your mother.

- **Direct Meanings:**
  - It understands you like your mother (does) [presumably well]
  - It understands (that) you like your mother
  - It understands you like (it understands) your mother

- **But there are other possibilities, e.g. mother could mean:**
  - a woman who has given birth to a child
  - a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

- **Context matters, e.g. what if previous sentence was:**
  - Wow, Amazon predicted that you would need to order a big batch of new vinegar brewing ingredients. 😊

[Example from L. Lee]
Dark Ambiguities

- *Dark ambiguities:* most structurally permitted analyses are so bad that you can’t get your mind to produce them.

This analysis corresponds to the correct parse of

“This will panic buyers!”

- Unknown words and new usages
- **Solution:** We need mechanisms to focus attention on the best ones, probabilistic techniques do this.
Problem: Scale

- People *did* know that language was ambiguous!
  - …but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
  - …they didn’t realize how bad it would be
A corpus is a collection of text
- Often annotated in some way
- Sometimes just lots of text
- Balanced vs. uniform corpora

Examples
- Newswire collections: 500M+ words
- Brown corpus: 1M words of tagged “balanced” text
- Penn Treebank: 1M words of parsed WSJ
- Canadian Hansards: 10M+ words of aligned French / English sentences
- The Web: billions of words of who knows what
Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire
Outline of Topics

- Will be continually updated on website

<table>
<thead>
<tr>
<th>Dates</th>
<th>Topics &amp; Lecture Slides</th>
<th>Notes</th>
<th>Textbook</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 7, 9</td>
<td>Introduction; Language Modeling (LM)</td>
<td>LM Notes</td>
<td>J&amp;M 4; M&amp;S 6</td>
<td>[Smoothing] [Pitman-Yor]</td>
</tr>
<tr>
<td>Jan 14, 16</td>
<td>Hidden Markov Models (HMMs) and Tagging</td>
<td>HMM Notes</td>
<td>J&amp;M 5.1-5.3, 6.1-6.4; M&amp;S 9, 10.1-10.3</td>
<td>[TnT Tagger] [Stanford Tagger]</td>
</tr>
<tr>
<td>Jan 23</td>
<td>PCFGs and Parsing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 28, 30</td>
<td>PCFGs and Parsing (cont'd)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb 4, 6</td>
<td>Machine Translation (MT) Intro.; Word Alignment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb 11, 13</td>
<td>Phrase-based MT; Syntax-based MT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb 20</td>
<td>Log-linear Models; Perceptron</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb 25, 27</td>
<td>Conditional Random Fields; Discriminative Parsing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar 4, 6</td>
<td>Unsupervised Learning and EM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar 11, 13</td>
<td>Compositional Semantics</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Course Details

Books:
- Jurafsky and Martin, Speech and Language Processing, 2nd Edition (not 1st)
- Manning and Schuetze, Foundations of Statistical NLP

Prerequisites:
- CSE 421 (Algorithms) or equivalent
- Some exposure to dynamic programming and probability helpful
- Strong programming
- There will be a lot of math and programming

Work and Grading:
- 60% - Four assignments (individual, submit code + write-ups)
- 40% - Final project (individual or small group)
What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Statistical Modeling Methods
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We’ll focus on what makes the problems hard, and what works in practice…
Class Requirements and Goals

- **Class requirements**
  - Uses a variety of skills / knowledge:
    - Probability and statistics
    - Basic linguistics background
    - Decent coding skills
  - Most people are probably missing one of the above
  - You will often have to work to fill the gaps

- **Class goals**
  - Learn the issues and techniques of modern NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
  - See where the holes in the field still are!