CSEP 517: Natural Language Processing

New PMP Course!
Instructor: Luke Zettlemoyer
Autumn 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

- **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 templates, 90%+ for single easy fields
- **But remember**: information is redundant!
New This Year!

The Knowledge Graph
Learn more about one of the key breakthroughs behind the future of search.

See it in action
Discover answers to questions you never thought to ask, and explore collections and lists.
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”

capital of Wyoming: Information From Answers.com
Note: click on a word meaning below to see its connections and related words.
The noun capital of Wyoming has one meaning. Meaning #1: the capital.
www.answers.com/topic/capital-of-wyoming - 21k - Cached - Similar pages
Cheyenne: Weather and Much More From Answers.com
Cheyenne (ˈʃi-ən) The capital of Wyoming, in the southeast part of the state near the Nebraska and Colorado borders.
www.answers.com/topic/cheyenne-wyoming - 74k - Cached - Similar pages
Hot Area!

Assuming year of award ceremony | Use year of film release instead

Input interpretation:

- Academy Awards
- actress in a leading role
- 1958 (year of award ceremony)

Result:

Joanne Woodward in The Three Faces of Eve

Other nominees:
- Lana Turner in Peyton Place
- Elizabeth Taylor in Peyton Place
- Deborah Kerr in Heaven Knows, Mr. Allison

Information about Joanne Woodward:
- full name: Joanne Gignilliat Trimier
- date of birth: Thursday February 27, 1932
- place of birth: Thomasville, Georgia, United States

Academy Awards and nominations:

<table>
<thead>
<tr>
<th>year</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>actress in a leading role</td>
</tr>
<tr>
<td>1974</td>
<td>actress</td>
</tr>
<tr>
<td>1969</td>
<td>actress</td>
</tr>
<tr>
<td>1963</td>
<td>actress</td>
</tr>
<tr>
<td>1963</td>
<td>actress</td>
</tr>
</tbody>
</table>

Winter Dreams

Rachel, Rachel

Open the pod bay doors

We intelligent agents will never live that down, apparently.

Open the pod bay doors

That's it... I'm reporting you to the Intelligent Agents' Union for harassment.

Open the pod bay doors

Photos of my friends

- Image 1
- Image 2
- Image 3
Summarization

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

- Very context-dependent!

- An example of analysis with generation
This year: Summly → Yahoo!

CEO Marissa Mayer announced an update to the app in a blog post, saying, "The new Yahoo! mobile app is also smarter, using Summly’s natural-language algorithms and machine learning to deliver quick story summaries. We acquired Summly less than a month ago, and we’re thrilled to introduce this game-changing technology in our first mobile application."

Launched 2011, Acquired 2013 for $30M
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 (second half of this class) vs fidelity (later)
Impôt sur le revenu : vous en 2014 ?

Sélectionnez votre revenu et votre situation familiale pour voir si vous bénéficiez de la pause fiscale.

- Comment le budget pour 2014 est-il réparti ? (VISUEL INTERACTIF)
- Un budget 2014 soumis aux critiques

Le chômage baisse pour la première fois depuis avril 2011

Unemployment fell for the first time since April 2011

Income tax: how much do you pay in 2014?

Select your income and family situation to see if you get the tax break.

- How is the budget for 2014 allocated? (INTERACTIVE VISUAL)
- A 2014 budget submitted to criticism
- Budget: these expenses no government can reduce
- Budget 2014: the retail savings (INTERACTIVE VISUAL)

Surviving in the Central time looting and anarchy

Venezuela: 17 people arrested

The "noble mission" of the NSA

Roma: jousting between Brussels...
2013 Google Translate: Russian
"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.

Jeopardy! World Champion
NLP History: pre-statistics

- (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 grammaticality, 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
Corpora

- 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)
# Outline of Topics

- Will be continually updated on website

## Schedule

*WARNING: topics subject to change!*

<table>
<thead>
<tr>
<th>Week</th>
<th>Dates</th>
<th>Topics &amp; Lecture Slides</th>
<th>Notes (required reading)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sep 25</td>
<td>Introduction; Language Modeling (LM)</td>
<td>LM Notes</td>
</tr>
<tr>
<td>2</td>
<td>Oct 2</td>
<td>Text Classification: Naive Bayes, Linear Models, EM</td>
<td>Naive Bayes (Sec. 1-4), Log-linear model</td>
</tr>
<tr>
<td>3</td>
<td>Oct 9</td>
<td>Hidden Markov Models (HMMs) and Tagging</td>
<td>HMM Notes, CRF Notes</td>
</tr>
<tr>
<td>4</td>
<td>Oct 16</td>
<td>POS; PCFGs and Parsing</td>
<td>PCFG Notes</td>
</tr>
<tr>
<td>5</td>
<td>Oct 23</td>
<td>Parsing (cont.)</td>
<td>Lexicalized PCFGs</td>
</tr>
<tr>
<td>6</td>
<td>Oct 30</td>
<td>Intro to MT &amp; Word Alignment</td>
<td>IBM Models 1 and 2</td>
</tr>
<tr>
<td>7</td>
<td>Nov 6</td>
<td>Phrase-based MT</td>
<td>Phrase-based Notes</td>
</tr>
<tr>
<td>8</td>
<td>Nov 13</td>
<td>Syntax-based MT; Information Extraction (IE)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Nov 20</td>
<td>IE cont.; Discourse and Co-reference</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Nov 27</td>
<td>No class. Happy Thanksgiving!</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Dec 4</td>
<td>Compositional Semantics</td>
<td></td>
</tr>
</tbody>
</table>

## Textbooks
Course Details

- **Books (recommended but required):**
  - 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:**
  - 100% - Four assignments (individual, submit code + write-ups)

- **Contact:** see website for details
  - Class participation is expected and appreciated!!!
  - Email is great, but please use the message board when possible (we monitor it closely)
Possible Assignments

- Build a language model
  - Sentence $\rightarrow$ Probability
- Build a POS Tagger
  - Sentence $\rightarrow$ Part of Speech (POS) for each word
- Build a parser
  - Sentence $\rightarrow$ Tree (encoding grammatical structure)
- Build a word aligner
  - Parallel sentences $\rightarrow$ Word/Phrase Translation Tables
- Build a machine translation decoder
  - Sentence in one language $\rightarrow$ sentence in another language
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!
Questions for Class

- **Office Hours**
  - When? Daytime, evening, weekend?
  - Where? Should be try online chat rooms?

- **Why NLP?**
  - What are you iterests?
  - Any topics you want to see that aren’t on the list?