CSE 517
Natural Language Processing
Winter 2015

Introduction
Yejin Choi

Slides adapted from Dan Klein, Luke Zettlemoyer
People & Page

- **Course page:**

- **People:**
  - **Instructor:** Yejin Choi
    - Office hours: Mon 3pm at CSE 578
  - **TA:** Eunsol Choi
    - Office hours: TBD
  - **TA:** Andrea M. Kahn
    - Office hours: TBD
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, sentiment analysis, question answering…
  - Unknown: human-level comprehension (is this just NLP?)
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 (second half of this class) vs fidelity (later)
Sélectionnez votre revenu et votre situation pour voir si vous bénéficiez de la pause fiscale.

- Comment le budget pour 2014 est-il réparti ?
- Un budget 2014 soumis aux critiques

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

- How is the budget for 2014 allocated?
- A 2014 budget submitted to criticism

- Budget: these expenses no government can reduce
- Budget 2014: the retail savings

Le chômage baisse pour la première fois depuis avril 2011

Unemployment fell for the first time since April 2011

- Surviving in the Central time looting and anarchy
- Surviving in the Central time looting and anarchy

DE FURSAC

automne-hiver 13/14
"Обиженные люди работают, а иностранцам нам не посидеть"
25.09.2013 19:48

Ректор "Baumanki" Анатолий "Pravda.Ru", какие шаги надо сделать, чтобы открыть ученый в связи с реформой РАН.

Фотосессия

В Индии: 40 жителей эвакуированы
18:55

Наводнение в Индии: 40 жителей эвакуированы
18:55

Президент США президент Путин: "Greenpeace" trying to grab the platform
25.09.2013 14:39

Председатель России, выступая на Арктическом форуме в Салехарде, сказал о северо-восточном плато "Пряжонок".

Expert: It is necessary to encourage participation in the election, rather than returning the column
06/25/2013 13:37

Политический лидер и философ, профессор Олег Матвеев, в свою очередь позитивно отнесся к "Pravda.Ru” Valentina Matviyenko извиниться перед Ратуша "against all".

The British newspaper described the heroes and victims in Nairobi
25/09/2013 10:27

В Кении - поминание жертв террористического нападения в Найроби. "Westgate" wurden überführt.
US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.
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!
Knowledge Graph: “things not strings”
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 (ˌʃeɪ-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
Mobile devices can now answer (some or our) questions and execute commands...

“What's the best movie to see this weekend”

That would probably start an argument. But here’s a list of highly-regarded movies:

“NORTH BY NORTHWEST”
Released July 17, 1959
100%

“THE TREASURE OF THE SIERRA…”
Released January 6, 1948
100%
Will this Be Part of All Our Home Devices?

https://www.youtube.com/watch?v=KkOCeAtKHiC
https://www.youtube.com/watch?v=qGU-SqUTees

- Why are these appearing now?
- What are fundamental limitations in current art?
"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 Naraaia 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
Speech Recognition

- **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)
Today: In 2012 election, automatic sentiment analysis actually being used to complement traditional methods (surveys, focus groups)

Past: “Sentiment Analysis” research started in 2002 or so

Future: Growing research toward computational social science, digital humanities (psychology, communication, literature and more)

Challenge: Need statistical models for deeper semantic understanding --- subtext, intent, nuanced messages
Summarization

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

- Very context-dependent!

- An example of analysis with generation
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.”

Launch 2011, Acquired 2013 for $30M
Can a robot write news?

Despite an expected dip in profit, analysts are generally optimistic about Steelcase as it prepares to report its third-quarter earnings on Monday, December 22, 2014. The consensus earnings per share estimate is 26 cents per share.

The consensus estimate remains unchanged over the past month, but it has decreased from three months ago when it was 27 cents. Analysts are expecting earnings of 85 cents per share for the fiscal year. Revenue is projected to be 5% above the year-earlier total of $784.8 million at $826.1 million for the quarter. For the year, revenue is projected to come in at $3.11 billion.

The company has seen revenue grow for three quarters straight. The less than a percent revenue increase brought the figure up to $786.7 million in the most recent quarter. Looking back further, revenue increased 8% in the first quarter from the year earlier and 8% in the fourth quarter.

The majority of analysts (100%) rate Steelcase as a buy. This compares favorably to the analyst ratings of three similar companies, which average 57% buys. Both analysts rate Steelcase as a buy.

Steelcase is a designer, marketer and manufacturer of office furniture. Other companies in the furniture and fixtures industry with upcoming earnings release dates include: HNI and Knoll.
Some of the formulaic news articles are now written by computers.

- Definitely far from “Op-ed”
- Can we make the generation engine statistically learned rather than engineered?
“Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers”

-- David Stork (HAL’s Legacy, 2001) on A. Rosenfeld’s vision
What begins to work (e.g., Kuznetsova et al. 2014)

The flower was so vivid and attractive.

Blue flowers are running rampant in my garden.

Spring in a white dress.

Blue flowers have no scent. Small white flowers have no idea what they are.

Scenes around the lake on my bike ride.

This horse walking along the road as we drove by.

We sometimes do well: 1 out of 4 times, machine captions were preferred over the original Flickr captions:
But many challenges remain (better examples of when things go awry)

- The couch is definitely bigger than it looks in this photo.
- My cat laying in my duffel bag.
- Yellow ball suspended in water.
- Incorrect Object Recognition
- Incorrect Scene Matching
- Incorrect Composition

- A high chair in the trees.
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 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.

- **Unknown words and new usages**

- **Solution**: We need mechanisms to focus attention on the best ones, probabilistic techniques do this.

This analysis corresponds to the correct parse of

“This will panic buyers!”

```
S   
  /  
NP  VP
   /  
DT VBZ NP
  /  /  
This is panic buying
```
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

<table>
<thead>
<tr>
<th>Week</th>
<th>Dates</th>
<th>Topics &amp; Lecture Slides</th>
<th>Notes (Required)</th>
<th>Textbook</th>
<th>Supplementary Readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jan 5, 7</td>
<td>Introduction; Language Models (LM)</td>
<td><strong>LM Notes</strong></td>
<td>J&amp;M 4.1-4; M&amp;S 6</td>
<td>[Large LMs] [Berkeley LM]</td>
</tr>
<tr>
<td>2</td>
<td>Jan 12, 14</td>
<td><strong>Sequences:</strong> Language Models and Smoothing; Hidden Markov Models (HMMs)</td>
<td><strong>HMM Notes</strong></td>
<td>J&amp;M 4.5-7; M&amp;S 6</td>
<td>[Smoothing]</td>
</tr>
<tr>
<td>3</td>
<td>Jan 21</td>
<td>Hidden Markov Models (HMMs) &amp; Part-Of-Speech Tagging</td>
<td></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] [SOTA POS]</td>
</tr>
<tr>
<td>4</td>
<td>Jan 26, 28</td>
<td><strong>Trees:</strong> Probabilistic Context Free Grammars (PCFG) and Parsing</td>
<td><strong>PCFG Notes, Lexicalized PCFGs</strong></td>
<td>J&amp;M 13-14; M&amp;S 11-12</td>
<td>[Syntax Intro] [Incremental] [Best First] [A* Parsing] [Lexicalized] [Unlexicalized] [Split Merge]</td>
</tr>
<tr>
<td>5</td>
<td>Feb 2, 4</td>
<td>More Parsing; Expectation Maximization (EM)</td>
<td><strong>EM Notes, Forward-backward, Inside-outside</strong></td>
<td>J&amp;M 6.5; M&amp;S 9.3-4; 11.3-4</td>
<td>[Semi-supervised Naive Bayes] [EM Tutorial] [EM for Feature-Rich]</td>
</tr>
<tr>
<td>6</td>
<td>Feb 9, 11</td>
<td><strong>Semantics:</strong> Frame Semantics; Distributional Semantics</td>
<td></td>
<td>J&amp;M 19.4; 20.7; 20.9; M&amp;S 8</td>
<td>[Fillmore-Tribute Workshop] [Frame-Semantic Parsing] [Composition in Distributional Models]</td>
</tr>
<tr>
<td>7</td>
<td>Feb 18</td>
<td><strong>Machine Translation (MT):</strong> Noisy Channel &amp; Word Alignment</td>
<td><strong>IBM Models 1 and 2</strong></td>
<td>J&amp;M 25.1-6; M&amp;S 13</td>
<td>[IBM Models] [HMM Model] [MERT Training]</td>
</tr>
<tr>
<td>8</td>
<td>Feb 23, 25</td>
<td>Phrase-based MT &amp; Syntax-based MT</td>
<td><strong>Phrase-based Notes</strong></td>
<td>J&amp;M 25.6-10; M&amp;S 13</td>
<td>[SCFG Tutorial] [Hiero] [Tree-to-String] [Tree-to-Tree]</td>
</tr>
<tr>
<td>9</td>
<td>Mar 2, 4</td>
<td><strong>Log-Linear / Feature-Rich Models:</strong> Conditional Random Fields (CRFs); Revisiting Sequences and Trees</td>
<td><strong>Log-linear models CRF Notes</strong></td>
<td>J&amp;M 6.6-6.8; M&amp;S 16.2-16.3</td>
<td>[MaxExt] [CRF Tutorial] [CRF LM] [CRF Parsing]</td>
</tr>
<tr>
<td>10</td>
<td>Mar 9, 11</td>
<td><strong>Knowledge &amp; Semantic Relations:</strong> Information Extraction; Entailment; Paraphrasing</td>
<td></td>
<td>J&amp;M 22</td>
<td>[Entailment Graphs] [Paraphrasing w/ MT] [Paraphrasing and Entailment]</td>
</tr>
</tbody>
</table>
Course Details

- **Books (recommended but required):**
  - Jurafsky and Martin, Speech and Language Processing, 2\textsuperscript{nd} Edition (not 1\textsuperscript{st})
  - 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:**
  - 3~4 programming-based homeworks (60%), a final mini-project (20%), non-programming assignments (15%) and course/discussion board participation (5%). No midterm or final exam.

- **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)
Assignments

Programming-based Assignments (code + write-up)
- Assignment 1: Language Models (Due Tue Jan 20, 5pm)
- Assignment 2: HMMs (Due Wed Feb 4, 5pm)
- Assignment 3: Parsing (Due Fri Feb 20, 5pm)
- Assignment 4: MT (Due Fri Mar 6, 5pm)

You may choose 3~4 assignments above depending on your plan for the final mini-project

Non-programming-based Assignments (15% grade)
- Lightweight 2~3 written homework assignments
Final Mini-project

- A final mini-project (20% of grade) will be completed during the last weeks of the term.
  - Do 3 out of 4 programming-based assignments + programming-based final project
  - Or do all 4 programming-based assignments + survey-based final project

- Students are encouraged to design projects that overlap with their research interests.

- Proposal (Due Feb 13, 5pm) <=== plan early!
- Final report (Due Mar 13, 5pm)
Useful Pointers

- ACL Anthology (http://www.aclweb.org/anthology/)
- All NLP conference and journal papers, free of charge
Useful Pointers

- Upcoming NLP Conference Deadlines & Venues
  - [http://www.cs.rochester.edu/~tetreaul/conferences.html](http://www.cs.rochester.edu/~tetreaul/conferences.html)

- Upcoming ACL-sponsored events
  [http://www.aclweb.org/portal/acl_sponsored_events](http://www.aclweb.org/portal/acl_sponsored_events)
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!