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?)
Why NLP

- To access information & knowledge
US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.
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?”
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)
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 is allocated? INTERACTIVE VISUAL
- A 2014 budget submitted to criticism
- Budget: these expenses no government can reduce
- Budget 2014: the retail savings INTERACTIVE VISUAL
Crash d’un Boeing en Iran : le Canada affirme que « l’avion a été abattu par un missile iranien »

Le premier ministre Justin Trudeau a affirmé disposer d’informations « de sources multiples », « Ce n’était peut-être pas intentionnel », a-t-il ajouté.

- Le Canada réclame « une enquête approfondie » après le crash du Boeing à Téhéran
- Le crash d’un Boeing ukrainien peu après son décollage à Téhéran fait 176 morts

Le premier ministre Justin Trudeau a affirmé disposer d’informations « de sources multiples », « Ce n’était peut-être pas intentionnel », a-t-il ajouté.

- Canada calls for "a thorough investigation" after the Boeing crash in Tehran
- Ukrainian Boeing crash shortly after takeoff in Tehran kills 176
Why NLP

- To access information & knowledge
- To communicate
Human-Machine Interactions

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

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

25 MOVIES

- NORTH BY NORTHWEST
  Released July 17, 1959
  100%

- THE TREASURE OF THE SIERRA...
  Released January 6, 1948
  100%

What can I help you with?

“You need to start understanding me Siri”

I’ll make a note of that.

“Yeah you better make a note of that”

Noted:

Of that
Will this Be Part of All Our Home Devices?

- Will it rain tomorrow?
- Set an alarm for eight a.m.
- Play music by Bruno Mars
- How many teaspoons are in a tablespoon?
- Add gelato to my shopping list
- Wikipedia: Abraham Lincoln
- When is Thanksgiving?
- Play my "dinner party" playlist
- What's the weather in Los Angeles this weekend?
- Add "make hotel reservations" to my to-do list
University of Washington

Sounding Board

Sounding Board

Location: Seattle, WA, USA
Faculty Advisor: Mari Ostendorf
Why NLP

- To access information & knowledge
- To communicate
- To understand our society
Analyzing public opinion, making political forecasts

- **Today**: In 2012 election, automatic sentiment analysis actually being used to complement traditional methods (surveys, focus groups)
- **Past**: “Sentiment Analysis” research started in 2002
- **Future**: computational social science and NLP for digital humanities (psychology, communication, literature and more)
- **Challenge**: Need statistical models for deeper semantic understanding --- subtext, intent, nuanced messages
Why NLP

- To access information & knowledge
- To communicate
- To understand our society
- And to make our lives easier
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."

Launched 2011, Acquired 2013 for $30M
OpenAI has published the text-generating AI it said was too dangerous to share

The lab says it’s seen ‘no strong evidence of misuse so far’

By James Vincent | Nov 7, 2019, 7:24am EST
Why NLP

- To access information & knowledge
- To communicate
- To understand our society
- To make our lives easier
- NLP and AI
"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
“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:
Table of Content

- Definition of NLP
- Historical account of NLP
(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!**
NLP: deep learning / neural networks

“The idea of what an internal representation would look like was it would be some kind of symbolic structure. That has completely changed with these big neural nets.”

–Hinton, 2016

- ~2014-now: Neural networks
  - Big models, more data, less and less linguistic bias
  - Can be brittle to adversarial inputs
  - Can be difficult to interpret

- 2020s: What comes next?
  - Hybrid models? Just deeper networks?
  - You decide!!!
2019, the year of BERT….

- Train a big NN as a masked language model on *lots* of unlabeled data

  **Input:** The man went to the [MASK]$_1$. He bought a [MASK]$_2$ of milk.

  **Labels:** [MASK]$_1$ = store, [MASK]$_2$ = gallon

- Fine tune for end task with labeled data
- Over 3,000 citations in first year alone…
BERT is in Google Search!

2019 brazil traveler to usa need a visa

BEFORE

9:00

google.com

U.S. citizens can travel to Brazil without the red tape of a visa ...

Mar 21, 2019 · Starting on June 17, you can go to Brazil without a visa and ... Australia, Japan and Canada will no longer need a visa to ... washingtonpost.com; © 1996-2019 The Washington Post ...

AFTER

9:00

google.com

Tourism & Visitor | U.S. Embassy & Consulates in Brazil

In general, tourists traveling to the United States require valid B-2 visas. That is unless they are eligible to travel visa ...
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
<table>
<thead>
<tr>
<th>Table of Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition of NLP</td>
</tr>
<tr>
<td>Historical account of NLP</td>
</tr>
<tr>
<td>Unique challenges of NLP</td>
</tr>
</tbody>
</table>
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**: ~95% 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.
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)
<table>
<thead>
<tr>
<th>Table of Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Definition of NLP</td>
</tr>
<tr>
<td>▪ Historical account of NLP</td>
</tr>
<tr>
<td>▪ Unique challenges of NLP</td>
</tr>
<tr>
<td>▪ Class administrivia / discussion</td>
</tr>
</tbody>
</table>
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...
Comparisons with Other Classes

- Compared to ML
  - Typically multivariate, dynamic programming everywhere
  - Structural Learning & Inference
  - Insights into language matters (a lot!)
  - DL: RNNs, LSTMs, Seq-to-seq, Attention, ...

- Compared to undergrad NLP
  - Faster paced
  - Stronger engineering skills & higher degree of independence assumed

- Compared to CompLing classes
  - More focus on core algorithm design, technically more demanding in terms of math, algorithms, and programming
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!