Lecture 1:
Course Introduction

CMPSCI 585, Fall 2015
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O’Connor
What

- Learn fundamental principles and methods in natural language processing
- Hands-on implementation experience
- Appreciation of basic linguistic issues
- Know about useful open-source NLP tools
- Know when NLP works and when it doesn’t
- Be able to read current research papers in NLP
- “AI systems”
How

• Lots of math!
  • This course: more than traditional CS, but less than traditional physics, engineering
• Lots of data!
• Lots of code that implements math & algorithms
  • Skill: translating from math to code
  • Skill: debugging math/linguistic/algorithm code
• A little bit of linguistics (goes a long way)
Requirements

- (20%) Participation and short exercises
- (30%) Problem sets
  - Written: math and concepts
  - Programming: Python programs
- (20%) Midterm (in-class, Oct 20)
- (30%) Final projects (groups of 1-2)
  - Choose a topic, or select a suggested topic
  - Project Proposal
  - Progress Report
  - In-class presentations (last week)
  - Final Report
Logistics

- Main course website for assignments and links to everything. This is ground truth if there are contradictions
  http://people.cs.umass.edu/~brenocon/inlp2015/

- Piazza for announcements, discussions and restricted access files

- Moodle for homework submissions (sometimes) and to see grades

- Waitlist situation

- To check:
  - SPIRE-registered students should have Piazza invites. Check @umass.edu email if you don’t!
  - Email me if you can’t access Piazza.

- Homework #0 due Thursday!
Readings

• Readings will be provided as PDFs on Piazza
  • Draft chapters of Jurafsky and Martin, *Speech and Language Processing*
  • We will use selections from both
    • 2nd edition (published)
    • 3rd edition (unpublished)
  • Other readings on occasion as well
Related courses at UMass

• Computational Linguistics: Ling 409, 492B (Bhatt, Dillon)
  https://sites.google.com/site/umasslx409/home
  http://www.umass.edu/linguist/courses/detail.php?cid=571

• Speech: Ling 592B (Yu)
  http://courses.umass.edu/linguist592b-kmyu/category/info.html

• Information Retrieval: CS 446, 646 (Allan, Croft)
  http://ciir.cs.umass.edu/cmpsci446/
  http://ciir.cs.umass.edu/~allan/cs646/
NLP is interdisciplinary

- Algorithms
- Statistics + Machine Learning
- Linguistics
- Cognitive Science
- Artificial Intelligence
“Can Machines Think?”

- British mathematician and founding figure in computer science
- Alan Turing (1950)
- How do we know when we have AI?
- “Imitation Game”
NLP imagined
NLP today

• Speech interfaces
• Machine translation
• Sentiment analysis
• Search engines
• ...

• [This course: document text analysis]
NLP today: Speech interfaces

What can I help you with?
“Rao’s coffee in Amherst, Massachusetts”
“Rao’s coffee in Amherst, Massachusetts”
NLP today: Question answering

IBM Watson

Wanted for general evilness, last seen at the Tower of Barad-Dur. It’s a giant eye, folks, kinda hard to miss
At the same time, the research team grew to about 25 full-time researchers and engineers, including several student members from key university partnerships. The team performed and documented more than 8,000 independent experiments by the time Watson went live. Each experiment generated 10 to 20 GB of trace data. Tools were developed to efficiently explore this data and discover failures and their likely causes. On the basis of analysis of this data, the team generated new algorithmic ideas and quantitatively estimated their potential impact on end-to-end performance. This data was used to prioritize, develop, and test new algorithms. Successful algorithmic advances were included in biweekly full-system builds. These were regularly run to produce updated baseline performance. This iterative process was implemented by the core team of researchers working in a single room and supported by more than 200 eight-core servers.

With the DeepQA architecture and the AdaptWatson methodology in place, the team drove the performance of Watson from early baselines delivering roughly 20% Precision@70 to greater than 85% Precision@70 good enough to compete with champions. Many of the papers in this issue describe the result of advancing core algorithms based on using DeepQA as a foundational architecture and the AdaptWatson methodology as a team-oriented process for rapidly creating and advancing a wide diversity of algorithm techniques to meet target performance.

Understanding questions

The breadth of the Jeopardy! domain is exemplified by the richness of language used, the variety of questions asked, and the huge range of types and topics covered. It is a challenge just to analyze the questions well enough to determine what they might be asking for or how the focus of the clue relates to other key elements in the clue. The more precisely Watson understands the clue, the better chance it has at finding and justifying answers. We refer to the word or phrase that indicates the class of thing the clue is asking for as the lexical answer type, or LAT. The clue in the first example below is asking for a president, which is a useful LAT. However, the LAT in the subsequent clue does not carry much semantic information at all. The third clue below claims to be looking for a star, but, in fact, the answer is a unique synthesis of Tom Cruise and cruise control — no star at all.

**RECENT HISTORY:** President under whom the U.S. gave full recognition to Communist China. (Answer: Jimmy Carter)

**IBM Watson**

25 engineers, 4 years, 200 subsystems, 2,880 CPU cores, 15 TB storage
NLP today: Question answering

From IBM Journal of Research and Development, 2012

Imperfect NLP is still useful

Table 1 DeepQA technology performance on public benchmark sets. (ACE: automatic content extraction; RTE: recognizing textual entailment.)

<table>
<thead>
<tr>
<th>NLP task</th>
<th>Evaluation set</th>
<th>Project start</th>
<th>State of art</th>
<th>Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsing</td>
<td>Wikipedia** accuracy</td>
<td>84.4</td>
<td>81.1 Charniak parser [19]</td>
<td>88.7</td>
</tr>
<tr>
<td>Entity disambiguation</td>
<td>Wikipedia disambiguation $F_1$</td>
<td>72.5</td>
<td>81.9 Hoffart et al. [42]</td>
<td>92.5</td>
</tr>
<tr>
<td>Relation detection</td>
<td>ACE 2004 $F_1$</td>
<td>45.8</td>
<td>72.1 Zhang et al. [43]</td>
<td>73.2</td>
</tr>
<tr>
<td>Textual entailment</td>
<td>RTE-6 2010 $F_1$</td>
<td>34.6</td>
<td>48.0 PKUTM [44]</td>
<td>48.8</td>
</tr>
</tbody>
</table>
Ambiguity: why NLP is hard
Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.
Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.
- Our company is training workers.
Ambiguity: why NLP is hard

- Juvenile Court to Try Shooting Defendant
- Hospitals Are Sued by 7 Foot Doctors
- Alice saw Bob with a telescope.
- Our company is training workers.
- They found that in order to attract settlers -- and make a profit from their holdings -- they had to offer people farms, not just tenancy on manorial estates.
Levels of linguistic structure

Characters

Alice talked to Bob.
Levels of linguistic structure

Morphology

Characters

talk-ed [VerbPast]

Alice talked to Bob.
Levels of linguistic structure

<table>
<thead>
<tr>
<th>Words</th>
<th>Alice talked to Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphology</td>
<td>talk-ed [VerbPast]</td>
</tr>
<tr>
<td>Characters</td>
<td>Alice talked to Bob.</td>
</tr>
</tbody>
</table>
Levels of linguistic structure

Syntax: Part of Speech
Words
Morphology
Characters

[Syntax]

Noun     VerbPast     Prep     Noun     Punct
Alice    talked    to    Bob    .

[Words]

Alice talked to Bob.

[Characters]

Alice talked to Bob.
Levels of linguistic structure

Syntax: Constituents

Syntax: Part of Speech

Words

Morphology

Characters

Alice talked to Bob.

Syntax:

Part of Speech

Noun

VerbPast

Prep

Noun

Punct

S

NP

VP

PP

talk\_ed

[VerbPast]
Levels of linguistic structure

Discourse

Semantics

Syntax: Constituents

Syntax: Part of Speech

Words

Morphology

Characters

CommunicationEvent(e)  SpeakerContext(s)
Agent(e, Alice)  TemporalBefore(e, s)
Recipient(e, Bob)

Syntax:

Part of Speech

Noun

VerbPast

Prep

Noun

Punct

Alice talked to Bob.

talked [VerbPast]

Alice talked to Bob.
NLP today: Machine translation

Xinhua Beijing September 2 (Reporter Liu) Politburo Standing Committee of the CPC Central Committee Liu Yunshan the 2nd met with the delegation led by the main 席斯塔尼舍夫 European Socialists in Beijing.

Liu Yunshan said China attaches great importance to China-EU relations and is willing to work together with the EU President Xi Jinping and implement the important consensus reached by leaders of the EU, around to create peace, growth, reform and civilization of the four partnerships, implementation of good "2020 China-EU cooperation in strategic planning," , deepen pragmatic cooperation in various fields. Communist Party of China is willing to develop good inter-party relations with the European Socialist Party, through various forms of
NLP today: Machine translation

Xinhua Beijing September 2 (Reporter Liu) Politburo Standing Committee of the CPC Central Committee Liu Yunshan the 2nd met with the delegation led by the main European Socialists in Beijing.

Liu Yunshan said China attaches great importance to China-EU relations and is willing to work together with the EU President Xi Jinping and implement the important consensus reached by leaders of the EU, around to create peace, growth, reform and civilization of the four partnerships, implementation of good "2020 China-EU cooperation in strategic planning," deepen pragmatic cooperation in various fields. Communist Party of China is willing to develop good inter-party relations with the European Socialist Party, through various forms of
NLP today: Trend analysis

Data: news articles

Dependency parsing to identify events

Machine learning from text:

(1) Event class dictionaries

```
arrive in, visit, meet with, travel to, leave,
hold with, meet, meet in, fly to, be in, arrive
for talk with, say in, arrive with, head to,
hold in, due in, leave for, make to, arrive to,
```

(2) Political dynamics

```
accuse, blame, say, break with, sever with,
blame on, warn, call, attack, rule with,
charge, say ← ccomp come from, say ← ccomp,
suspect, slam, accuse government ← poss,
```

```
kill in, have troops in, die in, be in, wound
in, have soldier in, hold in, kill in attack in,
remain in, detain in, have in, capture in, stay
in, about ← pobj troops in, kill, have troops
```

Israeli–Palestinian Diplomacy
Earnings for OmniVision Technologies Expected to Fall

Wall Street is expecting lower profit for OmniVision Technologies when the company reports its first quarter results on Thursday, August 28, 2014. Analysts are expecting earnings per share of 39 cents after the company booked a profit of 42 cents a share a year earlier.

The consensus estimate has risen from 16 cents over the past three months. Analysts are expecting earnings of 99 cents per share for the fiscal year. Revenue is projected to eclipse the year-earlier total of $373.7 million by 2%, finishing at $381.5 million for the quarter. For the year, revenue is projected to come in at $1.39 billion.

http://www.forbes.com/sites/narrativescience/
NLP today: Search/summarization

“We had the **crystal shrimp dumplings** that is seen wowing all yelpers, and they were superb.” in 5 reviews

“In addition to having the best pork buns in the area, they also have the best **scallion pancakes** and xiaolongbao.” in 8 reviews

“Usually I eat at the large dim sum restaurants in Boston with the rolling **carts**, but this experience was as good as any.” in 5 reviews
NLP today: Search/summarization
NLP today: Search/summarization

Entity Extraction

- Have technology (thanks to R6) – for English, Arabic and Chinese
- Allow queries like:
  - Show me all the word documents with references to IAEA
  - Show me all documents that reference Osama Bin Laden
  - Will allow a ‘show me more like this’ capability

NSA slides from Snowden leaks

http://www.theguardian.com/world/interactive/2013/jul/31/nsa-xkeyscore-program-full-presentation
• HW0
• See you on Thursday