Lecture 1:
Course Introduction

CS 585, Fall 2017
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2017

Brendan O’Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
What

• Learn fundamental principles and methods in natural language processing
• Hands-on implementation experience
• Appreciation of basic linguistic issues
• Know when NLP works and when it doesn’t
• “AI systems”
How

• Math! Algorithms!
• Data!
• Code!
  • Skill: translating from math to code
  • Skill: debugging math/linguistic/algorithm code
• A little bit of linguistics goes a long way
TAs

Satya Narayan Shukla

Abe Handler
Prerequisites

- Comfort with programming, algorithmic thinking
  - Ever debugged a graph algorithm? Know its Big-O time and space requirements?
  - CS 220 or 230
- Comfort with probability and mathematical notation
  - Ever used Bayes Rule?
  - CS 240
- Excitement about language!
- Willingness to learn

- Alternative: Ling 492B for linguistics-track students (imperfect solution; caveat emptor...)

Tuesday, September 5, 17
• “This is a HARD class”

• “The language parts are VERY INTERESTING. The math is next to impossible”

• “The class is moving very slowly, pace can be increased”
Requirements

- (10%) Participation and short exercises
  - Bring pencils/pens/paper to class
  - Laptops?
- (35%) Problem sets
  - Written: math and concepts
  - Programs: in Python
- (15%) Midterm (in-class, first week of Nov.)
- (40%) Final projects (groups of 1-3)
  - Choose a topic, or select a suggested topic
  - Project Proposal
  - Progress Report
  - In-class presentations
  - Final Report
Logistics

- Main course website: http://people.cs.umass.edu/~brenocon/inlp2017/
- Email me and the TAs via: cs585-instructors@googlegroups.com
- Piazza for announcements & discussions
- Gradescope/Moodle for homework submissions

- 585-01 and 585-02 sections are the same

- Due this Friday: HW0, probability review.

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

- Readings will be provided as PDFs on website
- Often draft chapters from Jurafsky and Martin, *Speech and Language Processing*
Related courses at UMass

• http://people.cs.umass.edu/~brenocon/complang_at_umass/
NLP is interdisciplinary

Algorithms

Linguistics

Statistics + Machine Learning

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?
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>
<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. | 20 |
Levels of linguistic structure

Morphology

Characters

Alice talked to Bob.
Levels of linguistic structure

Alice talked to Bob.

Morphology

talked [VerbPast]

Characters

Alice talked to Bob.
Levels of linguistic structure

<table>
<thead>
<tr>
<th>Syntax: Part of Speech</th>
<th>Noun</th>
<th>VerbPast</th>
<th>Prep</th>
<th>Noun</th>
<th>Punct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>Alice</td>
<td>talked</td>
<td>to</td>
<td>Bob</td>
<td>.</td>
</tr>
<tr>
<td>Morphology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Characters</td>
<td></td>
<td>talk -ed</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alice talked to Bob.

Alice talked to Bob.
Levels of linguistic structure

**Syntax:** Constituents

**Syntax:** Part of Speech

**Words**

**Morphology**

**Characters**

The sentence "Alice talked to Bob." has the following parse tree:

```
S
  ↓
 VP
  ↓
 NP  PP
     ↓
Noun  Prep  Noun
```

The words "talked" and "Bob" are in blue, indicating they are part of the past tense. The word "talked" is underlined, indicating it is a past tense verb.

The sentence is also shown as "Alice talked to Bob."
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:

Constituents

Part of Speech

Morphology

Words

Characters

Alice talked to Bob.

talked-ed [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

“diplomacy”
- 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,

“verbal conflict”
- 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,

“material conflict”
- 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

(2) Political dynamics

Israeli–Palestinian Diplomacy

Tuesday, September 5, 17
Earnings for OmniVision Technologies Expected to Fall

By Narrative Science

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/
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.
“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

Google search results for 'umass amherst'

About 8,510 results (0.42 seconds)

UMass Amherst breaks a world record
wwlp.com - 17 hours ago
AMHERST, Mass. (WWLP) – UMass Amherst has broken a world record by serving over 3,000 people a New England clambake in one and a ...

UMass Amherst starts semester with giant clambake
SouthCoastToday.com - 1 minute ago

UMass creates, sets new record for most clambake meals served ...
The Republican - masslive.com - 14 hours ago

UMass gets $37.5 million for environmental projects
The Recorder - 14 hours ago

No cheers here for UMass football program
Opinion - Boston Globe - 10 hours ago

Explore in depth (49 more articles)

UMass football players struggle academically
Boston Globe - Aug 30, 2014
The University of Massachusetts Amherst football team has struggled not only on the field, but in the classroom as well, leaving it barely above ...

Tuesday, September 5, 17
NLP today: Search/summarization

- 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
• Check out HW0 on the website
• See you on Thursday