Lecture 1:
Course Introduction

CMPSCI 585, Fall 2014
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2014/

Brendan O’Connor (http://brenocon.com)

Some material borrowed from
Chris Manning and Jacob Eisenstein
Who, where, when

<table>
<thead>
<tr>
<th>Instructor:</th>
<th>Brendan O'Connor</th>
<th>Office hours: Thursday after lecture (4pm–?), room 222</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA:</td>
<td>David Belanger</td>
<td>Office hours TBD (starts week of 9/8)</td>
</tr>
<tr>
<td>Lectures:</td>
<td>CS Building</td>
<td>room 142, Tues and Thurs, 2:30–3:45pm</td>
</tr>
</tbody>
</table>

- My email: brenocon@cs.umass.edu
- Course website: http://people.cs.umass.edu/~brenocon/inlp2014/ (Reachable from my website)
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”
Requirements

• Submit all homeworks electronically via moodle.umass.edu

• (~15%) Short exercises (~weekly): basic linguistic analysis to sharpen intuitions, <1 hour of work. Due before class; bring paper copy to class.

• (~25%) Problem sets (several): conceptual and programming problems.
  • Python preferred. We provide starter code.
  • IPython Notebook: http://ipython.org/notebook.html

• (~20%) Midterm (in-class, mid-October)

• (~20%) Final exam

• (~20%) Group-based final projects
First exercise - due Thursday

- Find several ambiguous sentences from news articles, and explain.
- Exercise will be posted on website later today.
- Due this Thursday, electronic submission before class.
  Bring paper copy to class!
Readings

• Readings will be provided -- see website.
• Suggested textbook:
  • Jurafsky and Martin, *Speech and Language Processing*
• Other great references:
  • Murphy, *Machine Learning: a Probabilistic Perspective*
  • Bender, *Linguistic Fundamentals for NLP*
  • Smith, *Linguistic Structure Prediction* (free access from UMass campus)
  • Manning and Schütze, *Foundations of Stat. NLP*
Related courses at UMass

- Computational Linguistics: Ling 409, 492B (Bhatt, Dillon)
  Intended to prepare non-CS students for 585
  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
- ...
NLP today: Speech interfaces

What can I help you with?
“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
NLP today: Question answering

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: 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

Google Translate

新华网北京9月2日电（记者刘华）中共中央政治局常委、中央书记处书记刘云山2日在北京会见了由主席斯塔尼舍夫率领的欧洲社会党代表团。

刘云山说，中方高度重视中欧关系，愿与欧方一道落实习近平主席与欧方领导人达成的重要共识，围绕打造和平、增长、改革、文明四大伙伴关系，执行好《中欧合作2020战略规划》，深化各领域务实合作。中国共产党愿与欧洲社会党发展良好党际关系，通过多种形式深化治国理政经验交流，促进中欧全面战略伙伴关系持续健康发展。刘云山应询介绍了中共党建有关情况。

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 等。李云山表示中欧关系是中国和中国的欧盟联盟关系的重要组成部分。 未来中国将致力于推动中欧关系的可持续发展。
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

“diplomacy”

A: Israel-Jordan Peace Treaty
B: Hebron Protocol
C: U.S. Calls for West Bank Withdrawal
D: Deadlines for West Bank Peace Accord
E: Negotiations in Mecca
F: Annapolis Conference

“verbal conflict”

“material conflict”

0.0 0.4 0.8

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/
“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

- Have technology (thanks to R6) – for English, Arabic and Chinese
- Allow queries like:
  - Show me all the word documents with references to IAEO
  - 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
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.
But we need **domain knowledge**, **discourse knowledge**, **world knowledge**, **linguistic knowledge**.
Demo
• Many constraints for meaning
  We use probability to combine
• Full language understanding is a long ways off
• But it is surprising how far simple methods can get you