COSI 134 - Statistical Approaches to Natural Language Processing

Ben Wellner
Fall 2010
Course Info

Instructor: Ben Wellner
TA: Chen Lin

Meeting Times
- Lectures: T/Th 5:20-6:30pm
- Office hours: T/Th 4:20pm – 5:20pm

Communication
- Web page: http://www.cs.brandeis.edu/~cs134 (not up-to-date yet)
- My e-mail: wellner@cs.brandeis.edu
- Chen’s e-mail: clin@brandeis.edu
Why NLP?

Computers perform as well as or much better than humans at many tasks that appear to involve ‘intelligence’

- Numeric calculations
- Games (e.g. chess)
- Theorem proving (some theorems)
- Scheduling, planning, etc.

We would like them to process/understand language too:

- Organize, summarize, manage, retrieve information
- Translate from one language to another
- Interface/communicate with humans via human language

Language is too complex, ambiguous, subtle

- Building machines to process language appears to require good linguistics and machine learning/statistical knowledge
**Why Statistical NLP?**

Language contains lots of ambiguity
- Genuine and potential uncertainty to resolve by context

Readily combine lots of pieces of evidence
- Too much for a human-derived heuristics/rules to consider and properly evaluate

Pipelines of statistical systems can minimize cascading errors
- Provide distributions over alternative predictions

Statistical systems can be tuned to (i.e. trained on) different data
- Different domains
- Different genres

Avoid labor intensive knowledge engineering
- *But* replace this with annotation
Information Extraction

Converting unstructured text into database records

- Allow for subsequent knowledge/data mining, inference

In July 1999, Dread Co. purchased 19,335 of Series C Convertible Preferred Shares in foostore.com, an on-line pharmacy, for cash of $9,125, including legal costs.

<table>
<thead>
<tr>
<th>Purchaser</th>
<th>Acquired</th>
<th>Amount</th>
<th>Time/Date</th>
<th>Assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dread Co.</td>
<td>Foostore.com</td>
<td>$9,125</td>
<td>July 1999</td>
<td>19,335 shares</td>
</tr>
</tbody>
</table>

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 Sept. was named president and chief operating officer of the parent.

<table>
<thead>
<tr>
<th>Person</th>
<th>Organization</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>starting</td>
</tr>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times newspaper</td>
<td>Executive vice president</td>
<td>ending</td>
</tr>
<tr>
<td>Lance R. Primis</td>
<td>New York Times Co.</td>
<td>President and COO</td>
<td>starting</td>
</tr>
</tbody>
</table>
Machine Translation

Current performance is now useful in many contexts

- Long way to go, still – but this is a success story
- Lots of statistics
- More and more linguistics integrated into translation models
Question Answering

Keyword search, information retrieval, still dominant

- Often, users are searching for *answers* to a *question*

Can be simple

- “Who is the president of France?”
- “What is the highest mountain in North America?”

More complex, subtle, open-ended

- “How do rockets work?”
- “What issues are important in the healthcare debate?”

Factoid questions can now be answered reasonably, even with textual differences between question and answer
Summarization

Scope of Summarization
- Single-document
- Multi-document

Extractive Summaries
- Extract individual sentences (or fragments) without rewording

Abstractive
- Involves text *generation* or text *re-writing* (i.e. in your own words)
Layers of NLP

**Tokenization/segmentation**
- Identifying what character units constitute words

**Morphology**
- Identifying components of words indicating grammatical function

(Phonetics/Phonology)

**Syntax**
- Grammatical structure; rules for structuring language

**Semantics**
- Lexical or compositional derivation of structures denoting meaning

**Discourse**
- How do sentences, clauses, phrases relate to each other

**Pragmatics**
- What is the intent of a given utterance or set of utterances
**Important NLP Tasks or Components**

**Tokenization – word boundaries**

**Morphological Analysis**
- Lemmatizers – normalize words (e.g. remove clitics)
- Part-of-speech analyzers

**Phrase Identification**
- Named Entity phrases; other task/domain-specific phrases
- Grammatical phrases (NPs, VPs, etc.)

**Co-reference**
- Which phrases refer to the same entity or event

**Word-sense Disambiguation (lexical semantics)**
- To what lexical entry does a word/phrase belong to

**Parsing**
- Constituent, Dependency
NLP Tasks (cont.)

Proposition Extraction (e.g. PropBank)
- Predicate-argument structure

Frame Extraction (e.g. FrameNet)
- Predicate-argument structure with “richer” semantics

Discourse
- Identifying discourse predicates
- Dialog acts, conversation analysis

Generating Logical Forms
- Meaning representation of an utterance, including quantifier scoping

Text Generation
- Mapping meaning representations to text
- Re-writing

Text Classification
State of the Field of NLP

Dominated by statistical, machine learning approaches

Why is this good?

- Better performance on many key NLP tasks
  - Parsing, phrase tagging, word-sense, text classification, etc.
- Improved statistical, machine learning methods and tools
- Some improved insight of contributions of linguistic intuitions
- Better, more rigorous evaluations of systems

Why is this not so good?

- More focus on engineering than science (perhaps)
- Incremental improvements on standard data sets favored over new ideas and new problems/tasks
- Less linguistic understanding of language phenomena
- Linguistic constraints/preferences often hidden in statistics
Course Goals

Broad understanding of statistical underpinnings of NLP
- Appreciate why statistical approaches work
  - And why they don’t always
- Translate linguistic intuitions into
  - Features for statistical models
  - Appropriate model structure
- Understand primary machine learning methods

Ability to apply statistical NLP techniques to real problems
- Use existing software packages and tools
- Ability to implement and understand algorithms for stat NLP

Be able to read and understand research papers in NLP
Identify places for new research
Course Requirements

Pre-requisites

- CS114 or some experience/background in NLP
- OR – statistics/ML background & willingness to pickup some linguistics
- OR – strong linguistics background & willingness to pickup statistics and machine learning
- Programming experience (Python, Java, etc.)

NLP very much inter-disciplinary

- Most people will have some gaps
- Some additional effort to fill these will be required
Course Work

Quizzes (10%)
- 2 quizzes – first half of the course

Mid-term Exam (15%)

Paper Summaries (15%)
- Read and discuss 10-12 research papers
- Summary and questions submitted for each paper

3 Homework Assignments (30%)
- Written work, programming and running experiments

Course Project (25%)
- 1) Programming and/or experimentation
  - Written report
- 2) Literature review paper
  - Both options: Class presentation
Flexibility on Assignments

- Students have different backgrounds and interests
- Homework assignments will have options that emphasize:
  - Algorithm implementation
  - Experimentation and analysis
- Java and Python preferred

Project

- Original work – OR re-implement existing algorithm
- Aim for a conference short paper in terms of work, presentation
- Abstracts will be due late October
- Individual effort; possible to pair-up
Materials

Main Text
- Manning and Schütze – Statistical Approaches to NLP
- Available online

Additional Texts
- Russell and Norvig - Artificial Intelligence: A modern approach
- Koller and Friedman – Probabilistic Graphical Models: Principles and Techniques

Software
- MALLET (mallet.cs.umass.edu)
- Natural Language Tool Kit (NLTK)
- Carafe Toolkit
Syllabus at a Glance

Technical Methods
- Probability, math essentials
- Supervised classification
  - Naïve Bayes, Maximum Entropy
- Sequence models
  - HMMs, MEMMs, CRFs
- Margin-based learning
  - SVMs, perceptron
- Graphical models
  - Bayesian networks
  - Markov random fields

Application/Task Areas
- Language modeling
- Part-of-speech tagging
- Phrase tagging
  - Named Entities, Chunking
- Text classification
- Topics, opinions/sentiment
- Co-reference
- Machine Translation
- Summarization
- Parsing
  - Constituent and Dependency
- Semantic Role labeling
- Discourse
A Look at Ambiguity

News Headlines

- Iraqi Head Seeks Arms
- Ban on Nude Dancing on Governor’s Desk
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- Hospitals Are Sued by 7 Foot Doctors
Syntactic and Semantic Ambiguity

Syntactic Ambiguity

- Bear left at the zoo
- I’m going to sleep
- Flying planes can be dangerous
- Time flies like an arrow

Attachment ambiguity

- Drag the file next to the item

  NP attachment: Drag \([_{NP} \text{the file} \ [_{PP} \text{next to the item}]])
  VP attachment: Drag \([_{NP} \text{the file} \ [_{PP} \text{next to the item}]])

Semantic (scope) Ambiguity/Underspecification

- Someone ate every tomato
Ambiguity, Vagueness, Noise, etc.

Statistical-based systems help deal with these problems

- Rely on human-annotated data

Ambiguity

- Some genuine – or, result of inadequate context/scope

Vagueness

- Occurs frequently for some tasks
- Will result in human disagreements without proper care

Noise

- Human annotators make mistakes
- Guidelines are never perfect; difficult corner-cases arise frequently
- Statistical-based systems can handle (some) noise
Corpus-based Methodology

A corpus is a collection of text

- Usually annotated by humans (linguists) for some specific linguistic phenomena (or task)

Large corpora provide:

- Broad coverage – lots of different examples and contexts
- Given, realistic data (not in the minds of linguists)
- Statistical information
  - How often is a named entity a person vs. a location phrase?
  - How often do NPs dominate PPs?
  - How often does a certain preposition attach low/high?
- A means to accurately evaluate our systems on real data
  - Compare system output (on unseen data) with human annotations
# Initial Statistical View of Corpora

<table>
<thead>
<tr>
<th>Token Freq.</th>
<th>Type Freq.</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>3</td>
</tr>
<tr>
<td>to</td>
<td>1725</td>
<td>4</td>
</tr>
<tr>
<td>of</td>
<td>1440</td>
<td>5</td>
</tr>
<tr>
<td>was</td>
<td>1161</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>&gt;100</td>
<td>102</td>
</tr>
<tr>
<td>TOTAL</td>
<td>71370</td>
<td></td>
</tr>
</tbody>
</table>

**Tom Sawyer Word Distributions**
- **Token vs. Type**
- 8,018 word types
- Nearly half occur just once
- Most common 100 words account for over half of text

**Zipf’s Law**
- Frequency is inversely proportional to frequency rank
  \[ F = \frac{1}{r} \]
- Small number of very frequent words
- Many, many very rare words – problem for Statistical Methods!
- This will tend to generalize beyond words
The Annotate-Train-Test Cycle

1) Identify an NLP Task
   - Note – this is where a lot of good linguistic insight is required

2) Get a lot of annotated (i.e. labeled) data created by humans

3) Build a simple system (and train it if appropriate)

4) Evaluate the system

5) Repeat:
   - Identify errors
   - Add additional resources, customize features based on what evidence humans bring to bear
   - Modify machine learning methods, models and representations to fit the problem

We will see evidence of this cycle in the papers we read
Most Class Projects will follow this methodology
Reading

Read Manning & Schütze Chapter 1, 2, 3

- Brandeis is a member of Cognet and the book is available for free
- E-mail me if you have problems accessing the book