Plan for Today's Class

Overview and Examples of Applied IR and Natural Language Processing

- Machine Translation
- Text Classification
- Text Mining
Plan for the Home Stretch

Today: Applied information retrieval and natural language processing

Wednesday: 202 Alumni Day -- IO & IR from the perspective of recent I-School grads

Monday 12/10 - last day of class -- course review

Wednesday 12/12 -- final exam (9-12 in room 202; due by 3pm)
Natural Language Processing

NLP has the goal of creating computers and machines that can use natural language -- i.e., the language used by people -- as their inputs and outputs.

The field is broad, and involves computer science, linguistics, cognitive psychology, statistics.

We're including this "taste" of NLP in this course because it illustrates many IR techniques and in some cases illustrates the tradeoffs between IO and IR.
Real World Applications

- Machine Translation
- Spelling Suggestions/Corrections
- Grammar Checking
- Speech Processing
- Text Categorization and Clustering
# "Text Mining" Applications

Table 1. Text-mining technologies offered by commercial vendors.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Inxight</th>
<th>Autonomy</th>
<th>ClearForest</th>
<th>SAS</th>
<th>Convera</th>
<th>Megaputer</th>
<th>SPSS</th>
<th>IBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>information extraction</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>topic tracking</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>summarization</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>categorization</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>concept linkage</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clustering</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>information visualization</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>question answering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Combining Linguistic and Statistical Approaches

Both the linguistic and data-driven or statistical approaches are seen as integral and complementary parts of an NLP application.

Systems employ sophisticated techniques for dictionaries and grammars to identify parts of speech and do morphological analysis.

But the statistics of co-occurrence / conditional probability yield many practical techniques for estimating the substitutability or semantic equivalence of words in larger text segments that make no use of their "languageness".

In particular, the web is such a huge corpus that statistical approaches can be surprisingly informative and robust.
Text Corpora

Computational linguists, computer scientists, experimental psychologists and others rely on text corpora for their research.

Prominent pre-web examples include the Brown corpus (Kucera and Francis, 1967) that includes a million words of contemporary American English...

... and the British National Corpus (http://www.natcorp.ox.ac.uk/) that contains 100 million words of contemporary British English.

But as large as the BNC is, because of Zipf’s Law most words occur fewer than 50 times in 100,000,000 words -- not frequent enough to draw statistical conclusions.
The Web as Corpus

The web is not representative of anything other than itself, but then neither are other text corpora

But the web dwarfs any other (possible?) corpus -- Google probably indexes a few trillion words, making it orders of magnitude larger than any other text collection

And most of it is freely available
# Phrases in BNC and Google

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical treatment</td>
<td>414</td>
<td>12,900,000</td>
<td>6,430,000</td>
<td>4,320,000</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>39</td>
<td>17,100,000</td>
<td>3,700,000</td>
<td>8,390,000</td>
</tr>
<tr>
<td>Deep breath</td>
<td>732</td>
<td>3,200,000</td>
<td>1,350,000</td>
<td>1,570,000</td>
</tr>
<tr>
<td>Acrylic paint</td>
<td>30</td>
<td>1,330,000</td>
<td>1,120,000</td>
<td>1,060,000</td>
</tr>
<tr>
<td>Perfect balance</td>
<td>38</td>
<td>1,700,000</td>
<td>1,300,000</td>
<td>1,400,000</td>
</tr>
<tr>
<td>Electromagnetic radiation</td>
<td>39</td>
<td>1,660,000</td>
<td>1,130,000</td>
<td>1,230,000</td>
</tr>
<tr>
<td>Powerful force</td>
<td>71</td>
<td>1,660,000</td>
<td>1,100,000</td>
<td>1,120,000</td>
</tr>
<tr>
<td>Concrete pipe</td>
<td>10</td>
<td>464,000</td>
<td>753,000</td>
<td>538,000</td>
</tr>
<tr>
<td>Upholstery fabric</td>
<td>6</td>
<td>781,000</td>
<td>1,150,000</td>
<td>757,000</td>
</tr>
<tr>
<td>Vital organ</td>
<td>46</td>
<td>169,000</td>
<td>250,000</td>
<td>264,000</td>
</tr>
</tbody>
</table>
Machine Translation: An Apocryphal but Important Example

A story often told about the early days of machine translation research (1950s) is that the English sentence:

*The spirit is willing, but the flesh is weak*

when translated into Russian, and then back to English became:

*The vodka is strong but the meat is rotten*
Great optimism in the 1950s was followed by extreme pessimism.

In 1966 a report by the Automatic Language Processing Advisory Committee (ALPAC) concluded that "there is no immediate or predictable prospect of useful machine translation" and instead recommended the development of computer aids for human translators.

Fortunately, ALPAC also recommended the continued support of basic research in computational linguistics.

In the 1970s and 1980s MT systems continued to develop, strongly driven by the emergence of microcomputers and text processing. The dominant technical strategy relied on hand-crafted syntactic parsers, morphological analyzers, and dictionaries - intensely semantic and rule-based approaches.
The beginning of the 1990s was a major turning point. An IBM research group developed a system called Candide that relied purely on statistical analysis and "example-based" methods for phrase matching and translation.

Candide used a very large corpus of English and French documents that had extremely reliable translations (in both directions).

This approach has really taken off with the emergence of the Web for obvious reasons...
How Good is Machine Translation? [1]

Microsoft's release of its Xbox 360 video-games console begins a new phase in the battle to remove Sony's PlayStation from the top spot. If it succeeds, the software giant may be tempted to make more incursions into the competitive market for home-entertainment hardware.

Roundtrip through German (Nov 2005): Release Microsofts of its video game console Xbox 360 begins a new phase in the battle for removing from PlayStation Sonys from the upper point. If it follows, the software giant can be provoked, in order to form more ideas into the free market for house maintenance small articles.

Roundtrip through German (Nov 2007): Microsoft's release of its Xbox 360 video game console begins a new phase in the battle to remove Sony's PlayStation from the top. If it succeeds, the software giant may be tempted to find out more incursions into the competitive market for home-entertainment hardware.
Microsoft's release of its Xbox 360 video-games console begins a new phase in the battle to remove Sony's PlayStation from the top spot. If it succeeds, the software giant may be tempted to make more incursions into the competitive market for home-entertainment hardware.

*Roundtrip through Chinese (Nov 2005):*
Its Xbox 360 video-games control bench Microsoft. The s release starts one new stage removes Sony in this battle; s PlayStation from this top spot. If it succeeds, perhaps the software giant does invades into the competitive market for the family entertainment hardware

*Roundtrip through Chinese (Nov 2007):*
Microsoft released its xbox 360 video game consoles began a new stage in the battle to remove Sony playstation from the throne of the world. To be successful, the software giant may be tempted to make more into the competitive market for home-entertainment hardware
Using Web Corpus to Improve Translation

Word selection in translation:

- French phrase *groupe de travail*
- *groupe* translates to cluster, group, grouping, concern, collective
- *travail* translates to work, labor, labour

**Table 4**
AltaVista frequencies for candidate translations of *groupe de travail*.

<table>
<thead>
<tr>
<th>labor cluster</th>
<th>21</th>
<th>labour collective</th>
<th>428</th>
</tr>
</thead>
<tbody>
<tr>
<td>labor grouping</td>
<td>28</td>
<td>work collective</td>
<td>759</td>
</tr>
<tr>
<td>labour concern</td>
<td>45</td>
<td>work concern</td>
<td>772</td>
</tr>
<tr>
<td>labor concern</td>
<td>77</td>
<td>labor group</td>
<td>3,977</td>
</tr>
<tr>
<td>work grouping</td>
<td>124</td>
<td>labour group</td>
<td>10,389</td>
</tr>
<tr>
<td>work cluster</td>
<td>279</td>
<td>work group</td>
<td>148,331</td>
</tr>
<tr>
<td>labor collective</td>
<td>423</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Google News gathers stories from more than 4,500 English-language news sources worldwide, and automatically arranges them to present the most relevant news first.

"Google News has no human editors selecting stories or deciding which ones deserve top placement. Our headlines are selected by computer algorithms, based on factors including how often and on what sites a story appears online."

"Our grouping technology examines numerous data points for each article published by the Google News sources, including the titles, text and publication time. We then use clustering algorithms to identify closely related articles."
Text Classification

Classification assigns objects in some domain to two or more classes or categories:

- words - determine part of speech
- words - disambiguate polysemy
- document retrieval - relevant/not relevant?
- author identification - shakespeare or not?
- sentiment classification - positive or negative affect? urgent or not urgent?
- language - English, Spanish, whatever?
Text Classification

Text Classification assumes a system of categories and some labeled instances so we can train a system to assign new instances to the appropriate categories.

The system's learning is Supervised learning.
Classification Problem
The Text Classification Process

- Specify classes
- Label text
- Extract features
- Choose a classifier algorithm
- Train and test
- Classify new examples
Features for Text Classification

Linguistic Features
- Words (stems?)
- Phrases
- Word and character level "N-grams"
- Punctuation
- Part of speech

Non-linguistic features (especially formatting)
Feature Selection

Not all features are equally good

So we need to eliminate, weight, and normalize features

Feature selection can be done in a task- and domain-independent or dependent manner
Classification Solution
Identifying Authorship

Given:

- A text with unknown author
- A list of possible authors
- A sample of their writing

Can we automatically determine which person wrote the text?
Motivation and Applications
The Disputed Federalist Papers

The Federalist papers were 77 short essays written in 1787-1788 by Hamilton, Jay and Madison to persuade NY to ratify the US Constitution; published under a pseudonym.

Historians disputed the authorship of 12 of the papers.

Two statisticians (Mosteller and Wallace, 1964) solved the problem by identifying 70 words whose usage patterns distinguished the papers with known authors.

Their statistical classifier concluded that the author was Madison.
Author Identification for the Federalist Papers

Separating Plane for the Federalists Papers – 1788 (Fung)
A Plan For Spam

Classifying email as "spam" or "not spam" using the simple and obvious approach of classifying messages as "spam" when they contain words most often contained in spam messages yields many false positives.

But if you are conservative in classifying messages as "spam" you have too many misses.

Bayesian approaches assign a "spam probability" to each word, then combines them into a single probability for the email. This combined score considers the good and bad words in an email.

This approach evolves with spam as it learns new words and considers their probabilities.

Trying to trick a Bayesian filter with misspelled words like "V1AG RA" just trains it to be more reliable.
Probability 101: Hypothesis Testing

We assume that there is some "true" state or value - called the "null hypothesis" - and we conduct some tests or make some observations to determine whether to believe it or to instead reject it and accept an "alternative hypothesis"

Example null hypotheses - the patient doesn't have the disease, the defendant is innocent, this message isn't spam, the graduation rate for starting football players is 90%

Alternative hypotheses - the patient has the disease, the defendant is guilty, this message is spam, the graduation rate isn't 90%

We conduct experiments / make observations to determine if we should reject the null hypothesis

The number of observations we make and their variability gives us more or less confidence about the hypotheses
Type I and Type II Errors

Our experiments or observations may suggest that the null hypothesis is false - that is, a "positive" test that the patient has the disease, the defendant is guilty, the message is spam, the graduation rate for starting football players isn't 90%

Or the results might be "negative" and not provide enough evidence for the disease, conviction, etc.

These outcomes or conclusions might be wrong in two ways:

- A **Type I error or false positive**
  is the error of rejecting a null hypothesis when it is in fact true; the supposedly positive evidence was observed due to chance

- A **Type II error or false negative**
  is the the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature; the test or observations made weren't powerful enough to detect the evidence that was there
Thinking About Probabilities

Most people think of probability using a *frequentist* approach, which focuses on identifying the "true" probability of some event, defined as the limit of its relative frequency in a large number of trials or samples.

In contrast, the *Bayesian* approach is a more subjective interpretation of probability, defined as a person's degree of belief about some event.

This degree of belief, called the *prior distribution*, is then changed by any data or observations -- i.e., your opinion can change if you get new information.

You updated degree of belief, the *posterior distribution*, is computed using Bayes' Rule.
Bayes' Rule

Bayes' Rule defines the "maximum amount of knowledge" you can get out of some piece of evidence

\[ p(A|B) = \frac{p(A)p(B|A)}{p(B)} \]

\[ = \frac{p(A) p(B|A)}{p(B|A) p(A) + p(B|\sim A) p(\sim A)} \]
The Need to Do Better than "Just Document Retrieval"

- Retrieve only the most relevant documents (better classification and ranking)
- Summarize the relevant documents
- Extract the important information to support question answering
- Answer questions directly
Motivation for Information Extraction

You're a baker and want to change jobs. Search for "baker job opening" on Google

That's hopeless. Much better to search for "baker" at monster.com
Information Extraction Application
But aggregating jobs from all over the web isn't the only IE application...

- Sales intelligence and lead generation
- Market intelligence
- Business intelligence
- "Central Intelligence" and Homeland Security

So IE is often a second step in topical categorization; after a text is categorized, the "information nuggets" in it can be extracted using topic-dependent rules
"Named Entity" Recognition

People, organizations, locations etc. can be identified with high accuracy in most kinds of documents using a combination of dictionaries, directories, gazetteers and rules.

Domain-specific knowledge and rules can be used for "named entities" like chemicals, species, proteins, etc.

Important entities are likely to be mentioned many times in a text, but are often described by different noun phrases each time, requiring co-reference resolution.

- Microsoft's release of its Xbox 360 video-games console begins a new phase in the battle... If it succeeds, the software giant may be tempted ...
  Gates and his army...
IE systems are successful in populating templates when matching rules can encode lots of information about the domain of the retrieved documents.

This means it works best when there is an implicit or explicit schema that describes the structure of the text to be extracted.

This means that IE is suitable for answering highly-structured questions where the answer can be assumed to exist somewhere in the "mixed content" of unstructured or semi-structured text.

But once information has been extracted from many documents of a particular type or topic, the aggregated collection of "information nuggets" can be "mined" to discover new facts or patterns -- put another way, we can now answer "harder" questions.
Text Data Mining: Examples

Positive Examples:

- hypothesis that magnesium deficiency can contribute to migraine headaches "mined" from a collection of scientific literature too broad for any one scientist to have read
- mined information about research funding, patents, and publications revealed a greater impact of government funding than suspected

Negative example: Your purchasing patterns reveal your values and vices
From Data Mining to Question Answering

People have questions, not queries -- but most web search engines aren't designed to handle natural language questions.

Question answering systems are designed to give the user a short answer to their question, not a long list of URLs.

More precisely, answering systems actually answer questions, while search engines give you a list of sites that mention the questions.

Example QA system is Brainboost.
How QA Systems Work

QA systems have been built on both ends of the language vs statistical learning dimension.

Very sophisticated systems using lots of NLP have traditionally done best, but new approaches that exploit the massive Web corpus are catching up fast.

These statistical systems rewrite the question into multiple queries in which the keywords occur in different orders.

This increases the probability of finding the answers, but is very inefficient, so use Bayes Rule to learn which query rewrites are best and stop doing useless ones.

QA systems also use IE techniques for "named entity recognition" to identify the parts of the retrieved documents where the questions are most likely to be answered.
Automated Customer Service

All of these NLP techniques come together in applications for automated customer support or "self-service"

Classify incoming messages / emails

Extract information to identify customer / product / problem

Use learning techniques to learn which words or phrases in message best classify the intent, urgency, sentiment of customer

Generate messages with information retrieved from enterprise applications to personalize the reply

If incoming message can't be handled automatically, route it to the human service agent whose knowledge is most appropriate to the customer concern
For December 5

No new readings

Survey on topics and readings
The Returning Alums

Peter Charles (Recommind)
Carolyn Cracraft (Primitive Logic)
Ben Hill (eBay)
Mano Marks (Google)
Patrick Schmidt
Kelly Snow (Cisco)