28. Applied IR and Natural Language Processing

INFO 202 - 3 December 2008

Bob Glushko

Plan for Today's Class

Overview and Examples of Applied IR and Natural Language Processing

- Machine Translation
- Text Mining
- Text Classification
- Question Answering

Plan for the Home Stretch

Today: Applied information retrieval and natural language processing

Monday 12/8 -- 202 Alumni Day -- IO & IR from the perspective of I-School grads

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

Monday 12/15 -- final exam (9-1 in room 202) & wine tasting (4-6 in downstairs lounge)

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.

Feature	Vendor: Inxight	Autonomy	Clearforest	SAS	Convera	Megaputer	SPSS	IBM
information extraction	X	X	X	X	X	X	X	X
topic tracking		X						
summarization	X	X			X	X		X
categorization	X	X	X	X	X	X	X	X
concept linkage		X	X	X				
clustering		X			X	X		X
information visualization	X						X	
question answering		X				X		

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

SAMPLE PHRASE	BNC	GOOGLE (11/2005)	GOOGLE (11/2006)	GOOGLE (11/2007)	GOOGLE (11/2008)
Medical treatment	**414**	**12,900,000**	**6,430,000**	**4,320,000**	**9,490,000**
Prostate cancer	**39**	**17,100,000**	**3,700,000**	**8.390,000**	**13,000,000**
Deep breath	**732**	**3,200,000**	**1,350,000**	**1,570,000**	**5,950,000**
Acrylic paint	**30**	**1,330,000**	**1,120,000**	**1,060,000**	**1,240,000**
Perfect balance	**38**	**1,700,000**	**1,300,000**	**1,400,000**	**3,350,000**
Electromagnetic radiation	**39**	**1,660,000**	**1,130,000**	**1,230,000**	**1,590,000**
Powerful force	**71**	**1,660.000**	**1,100,000**	**1,120,000**	**1,190,000**
Concrete pipe	**10**	**464,000**	**753,000**	**538,000**	**457,000**
Upholstery fabric	**6**	**781,000**	**1,150,000**	**757,000**	**862,000**
Vital organ	**46**	**169,000**	**250,000**	**264,000**	**177,000**

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

Machine Translation: A Brief History [1]

Great optimism in the 1950s was followed by extreme pessimism

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

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

In the 1970s and 1980s MT systems continued to develop; the dominant technical strategy relied on hand-crafted syntactic parsers, morphological analyzers, and dictionaries - intensely semantic and rule-based approaches.

Machine Translation: A Brief History [2]

The 1990s was a major turning point. IBM research developed the Candide systems 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 bi-directional translations

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 2008):
Microsoft's release of its Xbox 360 video game console begins a new phase in the struggle for the elimination of Sony's PlayStation from the top spot. If it succeeds, the software giant May be tempted to order more incursions into the competitive situation on the market for home entertainment hardware

How Good is Machine Translation? [2]

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 2008):
Microsoft released its Xbox 360 video game console to start a new phase in the battle to remove the Sony PlayStation from the top. If successful, the software giant may be more invasive fierce market competition 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*.

labor cluster	21	labour collective	428
labor grouping	28	work collective	759
labour concern	45	work concern	772
labor concern	77	labor group	3,977
work grouping	124	labour group	10,389
work cluster	279	work group	148,331
labor collective	423		

Topic Categorization in Google News

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:

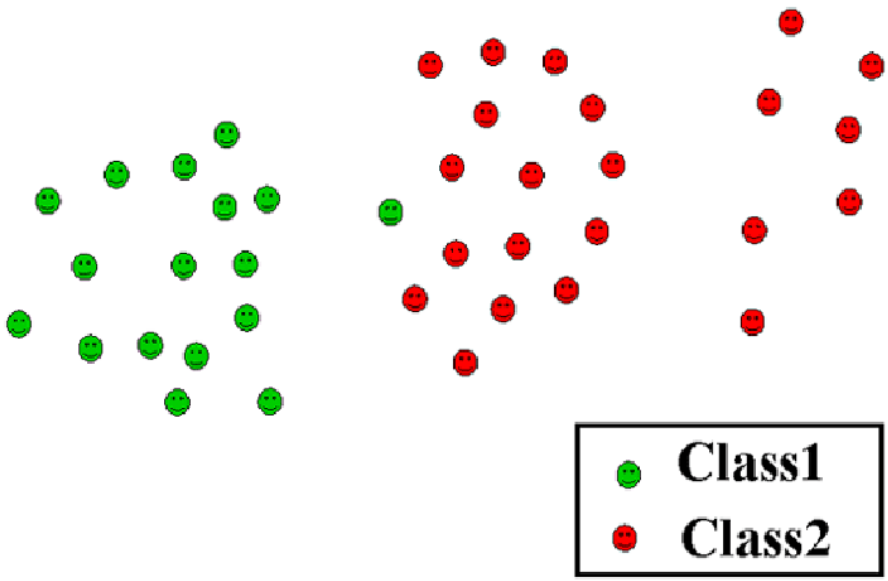
- words - determine part of speech
- words - disambiguate polysemy
- document retrieval - relevant/not relevant?
- author identification - shakespeare or not?
- sentiment classification - positive or negative affect? urgent or not urgent?
- language - English, Spanish, whatever?

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)

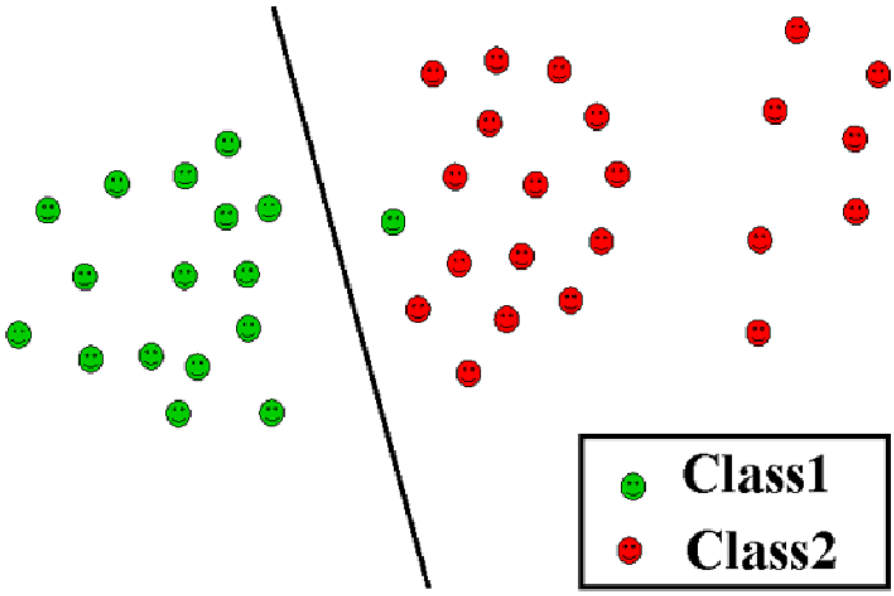
Feature Selection

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



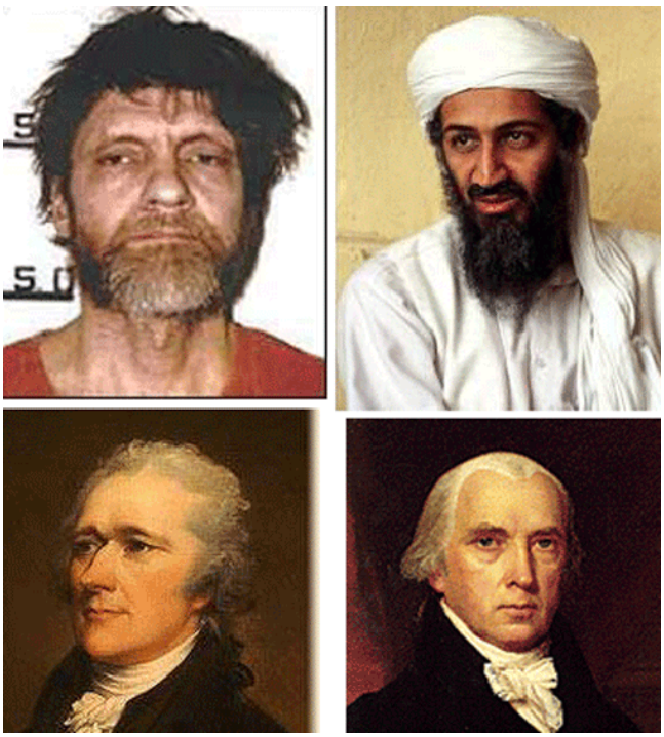
Identifying Authorship

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

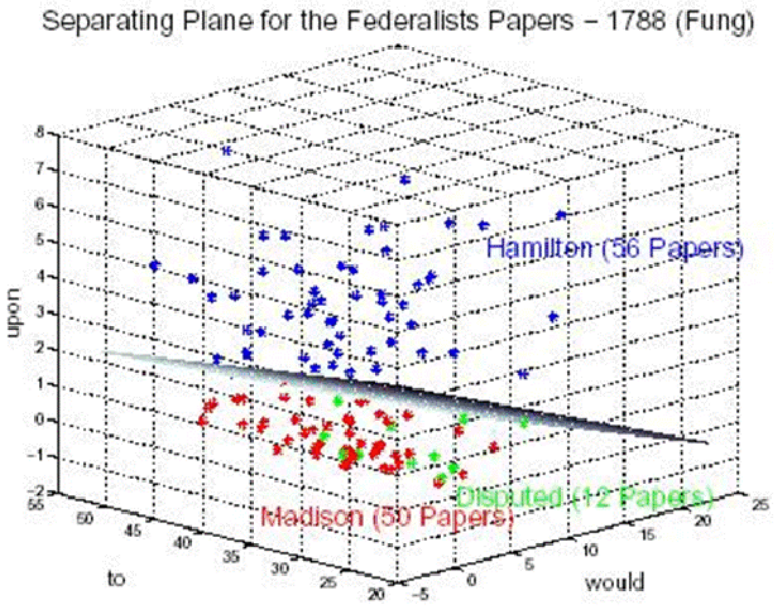


Figure 1: Obtained Hyperplane in 3 dimensions

A Plan For Spam [1]

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

A Plan For Spam [2]

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 [1]

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

Hypothesis Testing [2]

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

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.

Type I and Type II Errors

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
- http://www.intuitor.com/statistics/CurveApplet.html shows how differences in power and confidence levels affect the proportions of Type I and Type II errors

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

Important IE Application Areas

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*...

From IE to Text Data Mining

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

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 8

- Readings to be assigned by our returning alums

The Returning Alums

Carolyn Cracraft (Primitive Logic)

Zach Gillen (SF General Hospital)

Ben Hill (ex-eBay)

Mano Marks (Google)

Patrick Schmidt(UC Berkeley)