CS276B

Text Information Retrieval, Mining, and Exploitation

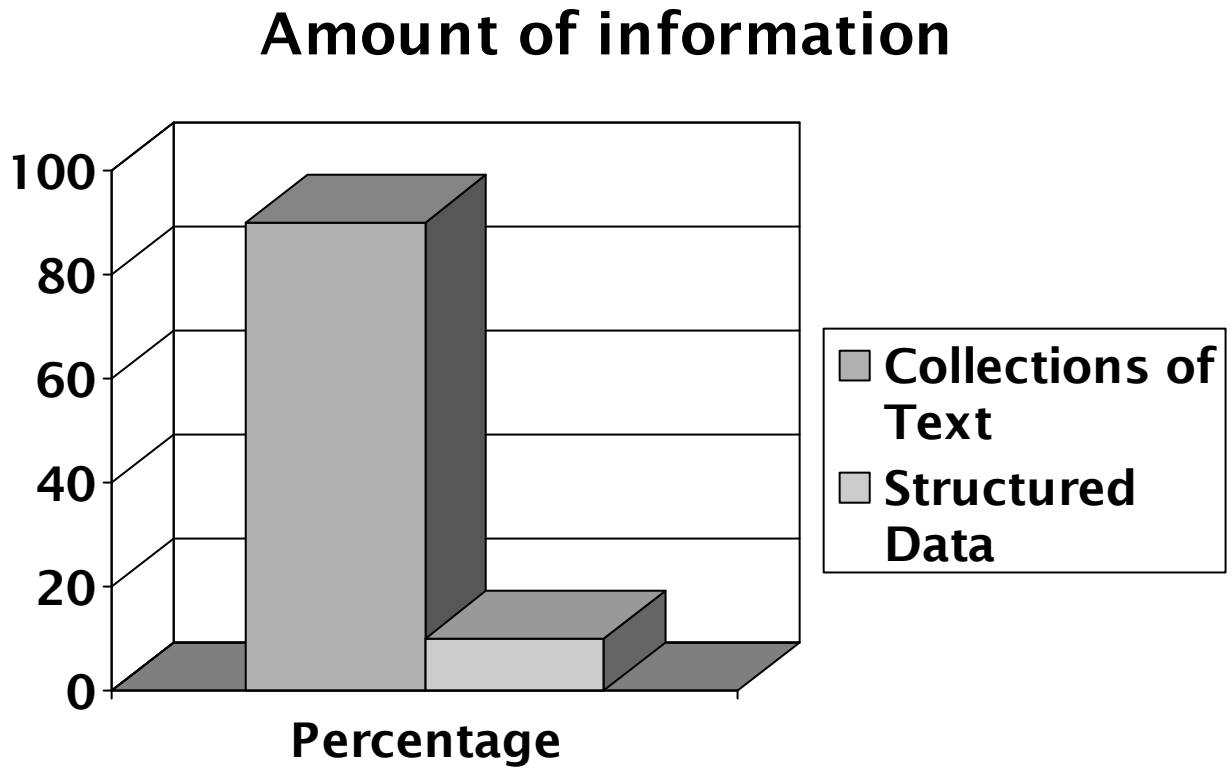
Lecture 12

Text Mining I

Feb 25, 2003

(includes slides borrowed from Marti Hearst,)

The Reason for Text Mining…



Corporate Knowledge “Ore”

- Email
- Insurance claims
- News articles
- Web pages
- Patent portfolios
- IRC
- Scientific articles
- Customer complaint letters
- Contracts
- Transcripts of phone calls with customers
- Technical documents

Text Knowledge Extraction Tasks

- Small Stuff. Useful nuggets of information that a user wants:
  - Question Answering
  - Information Extraction (DB filling)
  - Thesaurus Generation
- Big Stuff. Overviews:
  - Summary Extraction (documents or collections)
  - Categorization (documents)
  - Clustering (collections)
- Text Data Mining: Interesting unknown correlations that one can discover

Text Mining

- The foundation of most commercial "text mining" products is all the stuff we have already covered:
  - Information Retrieval engine
  - Web spider/search
  - Text classification
  - Text clustering
  - Named entity recognition
  - Information extraction (only sometimes)
- Is this text mining? What else is needed?

One tool: Question Answering

- Goal: Use Encyclopedia/other source to answer "Trivial Pursuit-style" factoid questions
- Example: "What famed English site is found on Salisbury Plain?"
- Method:
  - Heuristics about question type: who, when, where
  - Match up noun phrases within and across documents (much use of named entities
    - Coreference is a classic IE problem too!
  - More focused response to user need than standard vector space IR
    - Murax, Kupiec, SIGIR 1993; huge amount of recent work

Another tool: Summarizing

- High-level summary or survey of all main points?
- How to summarize a collection?
- Example: sentence extraction from a single document (Kupiec et al. 1995; much subsequent work)
  - Start with training set, allows evaluation
  - Create heuristics to identify important sentences:
    - position, IR score, particular discourse cues
  - Classification function estimates the probability a given sentence is included in the abstract
  - 42% average precision

IBM Text Miner terminology: Example of Vocabulary found

- Certificate of deposit
- CMOs
- Commercial bank
- Commercial paper
- Commercial Union Assurance
- Commodity Futures Trading Commission
- Consul Restaurant
- Convertible bond
- Credit facility
- Credit line
- Debt security
- Debtor country
- Detroit Edison
- Digital Equipment
- Dollars of debt
- End-March
- Enserch
- Equity warrant
- Eurodollar
- …

What is Text Data Mining?

- Peoples' first thought:
  - Make it easier to find things on the Web.
  - But this is information retrieval!
- The metaphor of extracting ore from rock:
  - Does make sense for extracting documents of interest from a huge pile.
  - But does not reflect notions of DM in practice. Rather:
    - finding patterns across large collections
    - discovering heretofore unknown information

Real Text DM

- What would finding a pattern across a large text collection really look like?
- Discovering heretofore unknown information is not what we usually do with text.
  - (If it weren't known, it could not have been written by someone!)
- However, there is a field whose goal is to learn about patterns in text for its own sake …
- Research that exploits patterns in text does so mainly in the service of computational linguistics, rather than for learning about and exploring text collections.

Definitions of Text Mining

- Text mining mainly is about somehow extracting the information and knowledge from text;
- 2 definitions:
  - Any operation related to gathering and analyzing text from external sources for business intelligence purposes;
  - Discovery of knowledge previously unknown to the user in text;
- Text mining is the process of compiling, organizing, and analyzing large document collections to support the delivery of targeted types of information to analysts and decision makers and to discover relationships between related facts that span wide domains of inquiry.

TDM using Metadata (instead of Text)

- Data:
  - Reuter's newswire (22,000 articles, late 1980s)
  - Categories: commodities, time, countries, people, and topic
- Goals:
  - distributions of categories across time (trends)
  - distributions of categories between collections
  - category co-occurrence (e.g., topic|country)
- Interactive Interface:
  - lists, pie charts, 2D line plots
    - (Dagan, Feldman, and Hirsh, SDAIR '96)

True Text Data Mining: Don Swanson's Medical Work

- Given
  - medical titles and abstracts
  - a problem (incurable rare disease)
  - some medical expertise
- find causal links among titles
  - symptoms
  - drugs
  - results
- E.g.: Magnesium deficiency related to migraine
  - This was found by extracting features from medical literature on migraines and nutrition

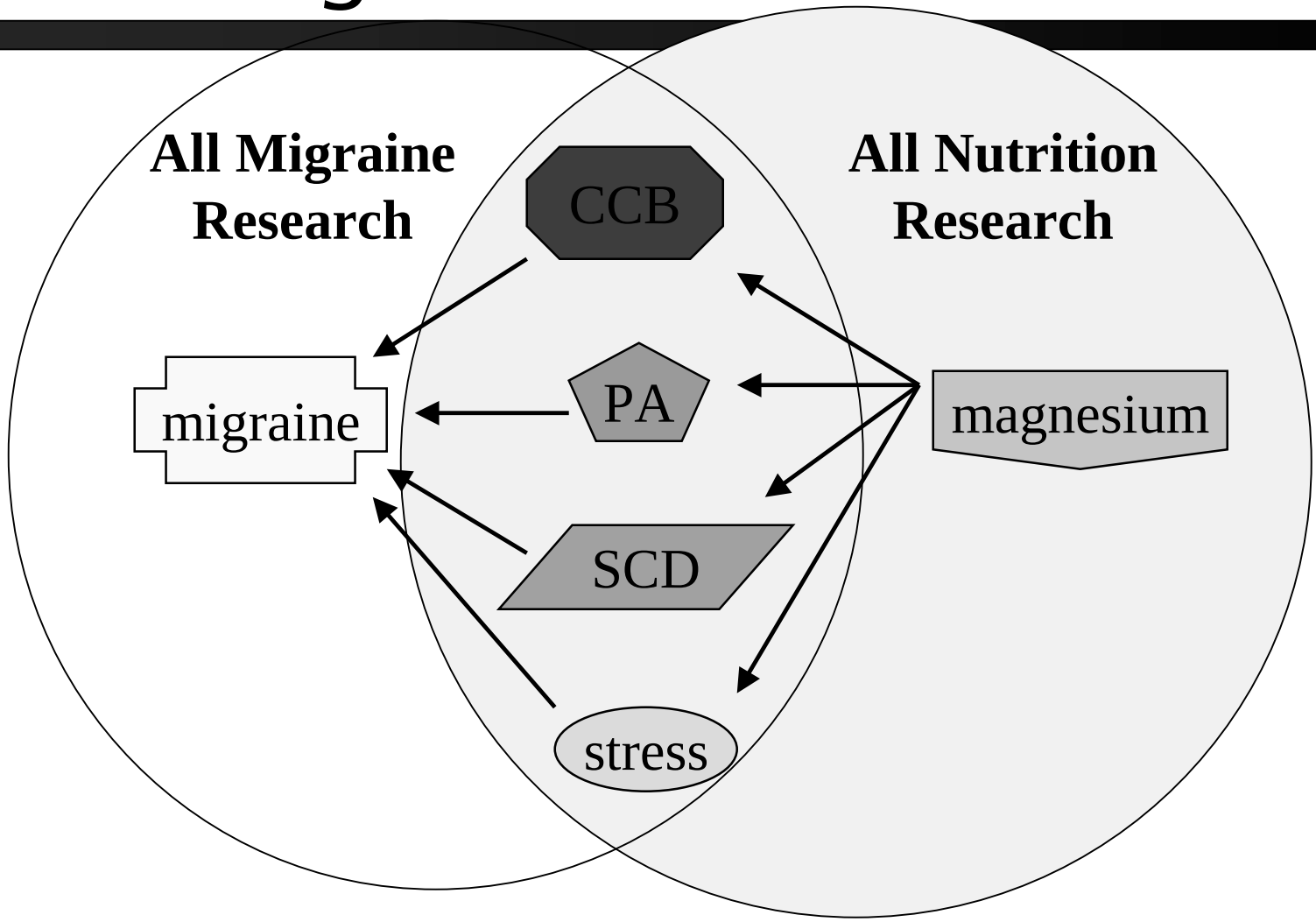
Swanson Example (1991)

- Problem: Migraine headaches (M)
  - Stress is associated with migraines;
  - Stress can lead to a loss of magnesium;
  - calcium channel blockers prevent some migraines
  - Magnesium is a natural calcium channel blocker;
  - Spreading cortical depression (SCD) is implicated in some migraines;
  - High levels of magnesium inhibit SCD;
  - Migraine patients have high platelet aggregability;
  - Magnesium can suppress platelet aggregability.
- All extracted from medical journal titles

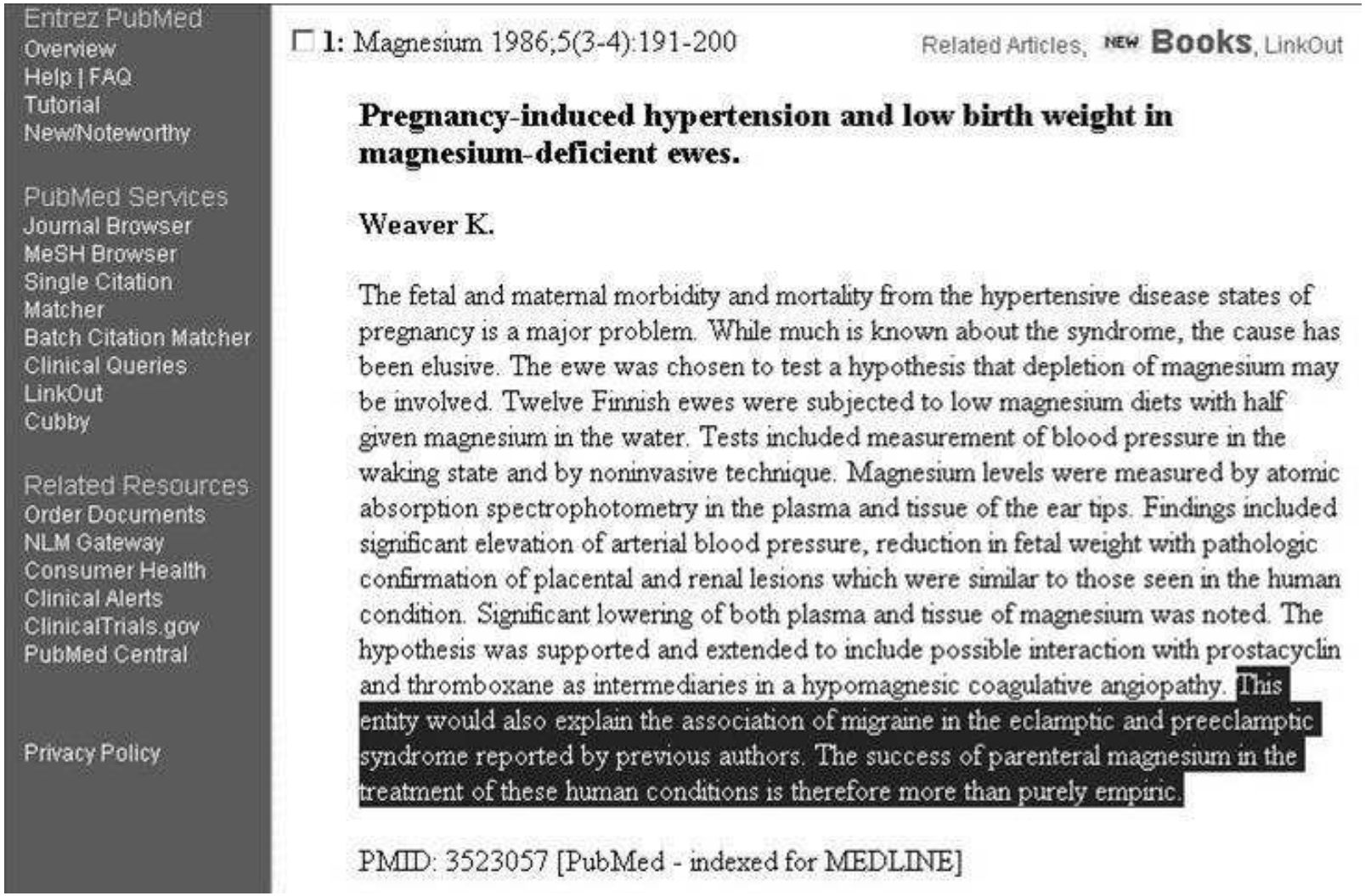
Swanson's TDM

- Two of his hypotheses have received some experimental verification.
- His technique
  - Only partially automated
  - Required medical expertise
- Few people are working on this kind of information aggregation problem.

Gathering Evidence



Or maybe it was already known?



Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy

PubMed Services
Journal Browser
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Gateway
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Privacy Policy

1: Magnesium 1986;5(3-4):191-200 Related Articles, NEW Books, LinkOut

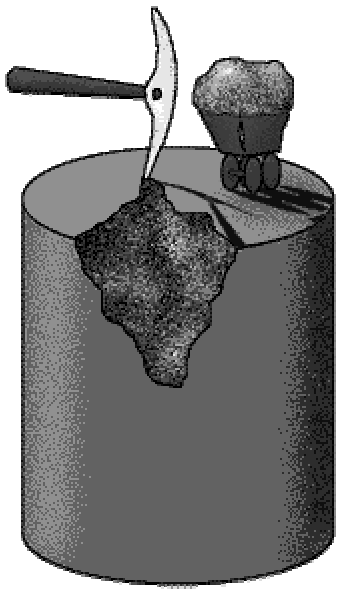
Pregnancy-induced hypertension and low birth weight in magnesium-deficient ewes.

Weaver K.

The fetal and maternal morbidity and mortality from the hypertensive disease states of pregnancy is a major problem. While much is known about the syndrome, the cause has been elusive. The ewe was chosen to test a hypothesis that depletion of magnesium may be involved. Twelve Finnish ewes were subjected to low magnesium diets with half given magnesium in the water. Tests included measurement of blood pressure in the waking state and by noninvasive technique. Magnesium levels were measured by atomic absorption spectrophotometry in the plasma and tissue of the ear tips. Findings included significant elevation of arterial blood pressure, reduction in fetal weight with pathologic confirmation of placental and renal lesions which were similar to those seen in the human condition. Significant lowering of both plasma and tissue of magnesium was noted. The hypothesis was supported and extended to include possible interaction with prostacyclin and thromboxane as intermediaries in a hypomagnesic coagulative angiopathy. This entity would also explain the association of migraine in the eclamptic and preeclamptic syndrome reported by previous authors. The success of parenteral magnesium in the treatment of these human conditions is therefore more than purely empiric.

PMID: 3523057 [PubMed - indexed for MEDLINE]

Extracting Metadata from documents



Why metadata?

- Metadata = “data about data”
- “Normalized” semantics
- Enables easy searches otherwise not possible:
  - Time
  - Author
  - Url / filename
- And gives information on non-text content
  - Images
  - Audio
  - Video

For Effective Metadata We Need:

- Semantics
  - Commonly understood terms to describe information resources
- Syntax
  - Standard grammar for connecting terms into meaningful “sentences”
- Exchange framework
  - So we can recombine and exchange metadata across applications and subjects

Dublin Core Element Set

- Title (e.g., Dublin Core Element Set)
- Creator (e.g., Hinrich Schuetze)
- Subject (e.g, keywords)
- Description (e.g., an abstract)
- Publisher (e.g., Stanford University)
- Contributor (e.g., Chris Manning)
- Date (e.g, 2002.12.03)
- Type (e.g., presentation)
- Format (e.g., ppt)
- Identifier (e.g., http://www.stanford.edu/class/cs276a/syllabus.html)
- Source (e.g. http://dublincore.org/documents/dces/)
- Language (e.g, English)
- Coverage (e.g., San Francisco Bay Area)
- Rights (e.g., Copyright Stanford University)

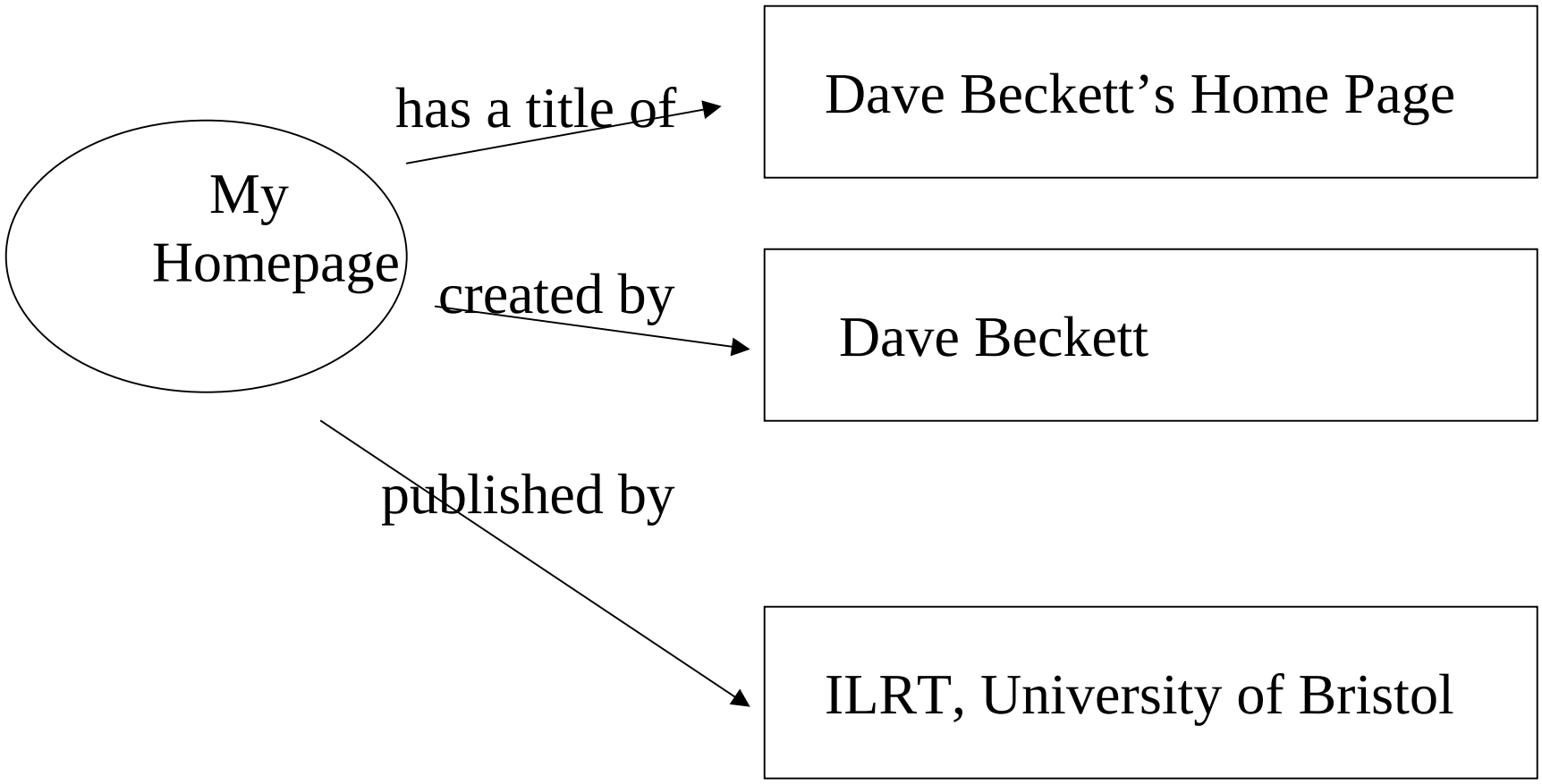
RDF = Resource Description Framework

- Emerging standard for metadata
- W3C standard
  - Part of W3C's metadata framework
- Specialized for WWW
- Desiderata
  - Combine different metadata modules (e.g., different subject areas)
  - Syndication, aggregation, threading

RDF example in XML

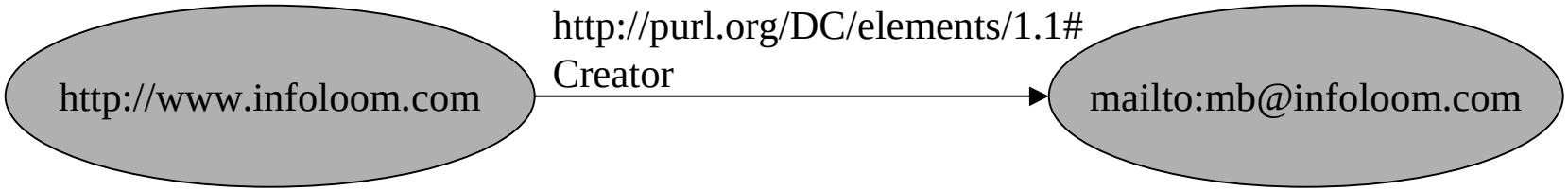
```
<?xml version="1.0"?> <rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#" xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description
rdf:about="http://www.ilrt.org/people/cmdjb/">
<dc:title>Dave Beckett's Home Page</dc:title>
<dc:creator>Dave Beckett</dc:creator>
<dc:publisher>ILRT, University of
Bristol</dc:publisher> </rdf:Description> </rdf:RDF>
```

RDF example


has a title of
Dave Beckett's Home Page
My Homepage
created by
Dave Beckett
published by
ILRT, University of Bristol

Resource Description Framework (RDF)

- RDF was conceived as a way to wrap metadata assertions (eg Dublin Core information) around a web resource.
- The central concept of the RDF data model is the *triple*, represented as a labeled edge between two nodes.
- The subject, the object, and the predicate are all resources, represented by URIs



- Properties can be multivalued for a resource, and values can be literals instead of resources
- Graph pieces can be chained and *nested*
- RDF Schema gives frame-based language for ontologies and reasoning over RDF.

Metadata Pros and Cons

- CONS
  - Most authors are unwilling to spend time and energy on
    - learning a metadata standard
    - annotating documents they author
  - Authors are unable to foresee all reasons why a document may be interesting.
  - Authors may be motivated to sabotage metadata (patents).
- PROS
  - Information retrieval often does not work.
  - Words poorly approximate meaning.
  - For truly valuable content, it pays to add metadata.
- Synthesis
  - In reality, most documents have some valuable metadata
  - If metadata is available, it improves relevance and user experience
  - But most interesting content will always have inconsistent and spotty metadata coverage

Metadata and TextCat/IE

- The claim of metadata proponents is that metadata has to be explicitly annotated, because we can't hope to get, say, a book price from varied documents like:

<H1>
<The Rhyme of the Ancient Mariner>
</H1>
The Rhyme of the Ancient Mariner, by Samuel Coleridge, is available for the low price of $9.99. This Dover reprint is beautifully illustrated by Gustave Dore.
<p>
Julian Schnabel recently directed a movie, Pandemonium, about the relationship between Coleridge and Wordsworth.

Metadata and TextCat/IE

- … but with IE/TextCat, these are exactly the kind of things we *can* do
- Of course, we can do it more accurately with human authored metadata
  - But, of course, the metadata might not match the text (*metadata spamming*)
- Opens up an interesting world where agents use metadata if it's there, but can synthesize it if it isn't (by text cat/IE), and can verify metadata for correctness against text
  - Seems a promising area; not much explored!

Lexicon Construction

What is a Lexicon?

- A database of the vocabulary of a particular domain (or a language)
- More than a list of words/phrases
- Usually some linguistic information
  - Morphology (manag- e/es/ing/ed -> manage)
  - Syntactic patterns (transitivity etc)
- Often some semantic information
  - Is-a hierarchy
  - Synonymy

Lexica in Text Mining

- Many text mining tasks require named entity recognition.
- Named entity recognition requires a lexicon in most cases.
- Example 1: Question answering
  - Where is Mount Everest?
  - A list of geographic locations increases accuracy
- Example 2: Information extraction
  - Consider scraping book data from amazon.com
  - Template contains field “publisher”
  - A list of publishers increases accuracy
- Manual construction is expensive: 1000s of person hours!
- Sometimes an unstructured inventory is sufficient

Lexicon Construction (Riloff)

- Attempt 1: Iterative expansion of phrase list
- Start with:
  - Large text corpus
  - List of seed words
- Identify “good” seed word contexts
- Collect close nouns in contexts
- Compute confidence scores for nouns
- Iteratively add high-confidence nouns to seed word list. Go to 2.
- Output: Ranked list of candidates

Lexicon Construction: Example

- Category: weapon
- Seed words: bomb, dynamite, explosives
- Context: <new-phrase> and <seed-phrase>
- Iterate:
  - Context: They use TNT and other explosives.
  - Add word: TNT
- Other words added by algorithm: rockets, bombs, missile, arms, bullets

Lexicon Construction: Attempt 2

- Multilevel bootstrapping (Riloff and Jones 99)
- Generate two data structures in parallel
  - The lexicon
  - A list of extraction patterns
- Input as before
  - Corpus (not annotated)
  - List of seed words

Multilevel Bootstrapping

- Initial lexicon: seed words
- Level 1: Mutual bootstrapping
  - Extraction patterns are learned from lexicon entries.
  - New lexicon entries are learned from extraction patterns
  - Iterate
- Level 2: Filter lexicon
  - Retain only most reliable lexicon entries
  - Go back to level 1
- 2-level performs better than just level 1.

Scoring of Patterns

- Example
  - Concept: company
  - Pattern: owned by <x>
- Patterns are scored as follows
  - score(pattern) = F/N log(F)
  - F = number of unique lexicon entries produced by the pattern
  - N = total number of unique phrases produced by the pattern
  - Selects for patterns that are
    - Selective (F/N part)
    - Have a high yield (log(F) part)

Scoring of Noun Phrases

- Noun phrases are scored as follows
  - score(NP) = sum_k (1 + 0.01 * score(pattern_k))
  - where we sum over all patterns that fire for NP
  - Main criterion is number of independent patterns that fire for this NP.
  - Give higher score for NPs found by high-confidence patterns.
- Example:
  - New candidate phrase: boeing
  - Occurs in: owned by <x>, sold to <x>,

Shallow Parsing

- Shallow parsing needed
  - For identifying noun phrases and their heads
  - For generating extraction patterns
- For scoring, when are two noun phrases the same?
  - Head phrase matching
  - X matches Y if X is the rightmost substring of Y
  - “New Zealand” matches “Eastern New Zealand”
  - “New Zealand cheese” does not match “New Zealand”

Seed Words

Web Company:	*co. company corp. corporation inc. incorporated limited ltd. plc*
Web Location:	*australia canada china england france germany japan mexico switzerland united_states*
Web Title:	*ceo cfo president vice-president vp*
Terr. Location:	*bolivia city colombia district guatemala honduras neighborhood nicaragua region town*
Terr. Weapon:	*bomb bombs dynamite explosive explosives gun guns rifle rifles tnt*

Mutual Bootstrapping

Generate all candidate extraction patterns from the training corpus using AutoSlog.

Apply the candidate extraction patterns to the training corpus and save the patterns with their extractions to *EPdata*

SemLex = {seed_words}

Cat_EPlist = {}

MUTUAL BOOTSTRAPPING LOOP

1. Score all extraction patterns in *EPdata*.
2. *best_EP* = the highest scoring extraction pattern not already in *Cat_EPlist*
3. Add *best_EP* to *Cat_EPlist*
4. Add *best_EP*'s extractions to *SemLex*.
5. Go to step 1

Extraction Patterns

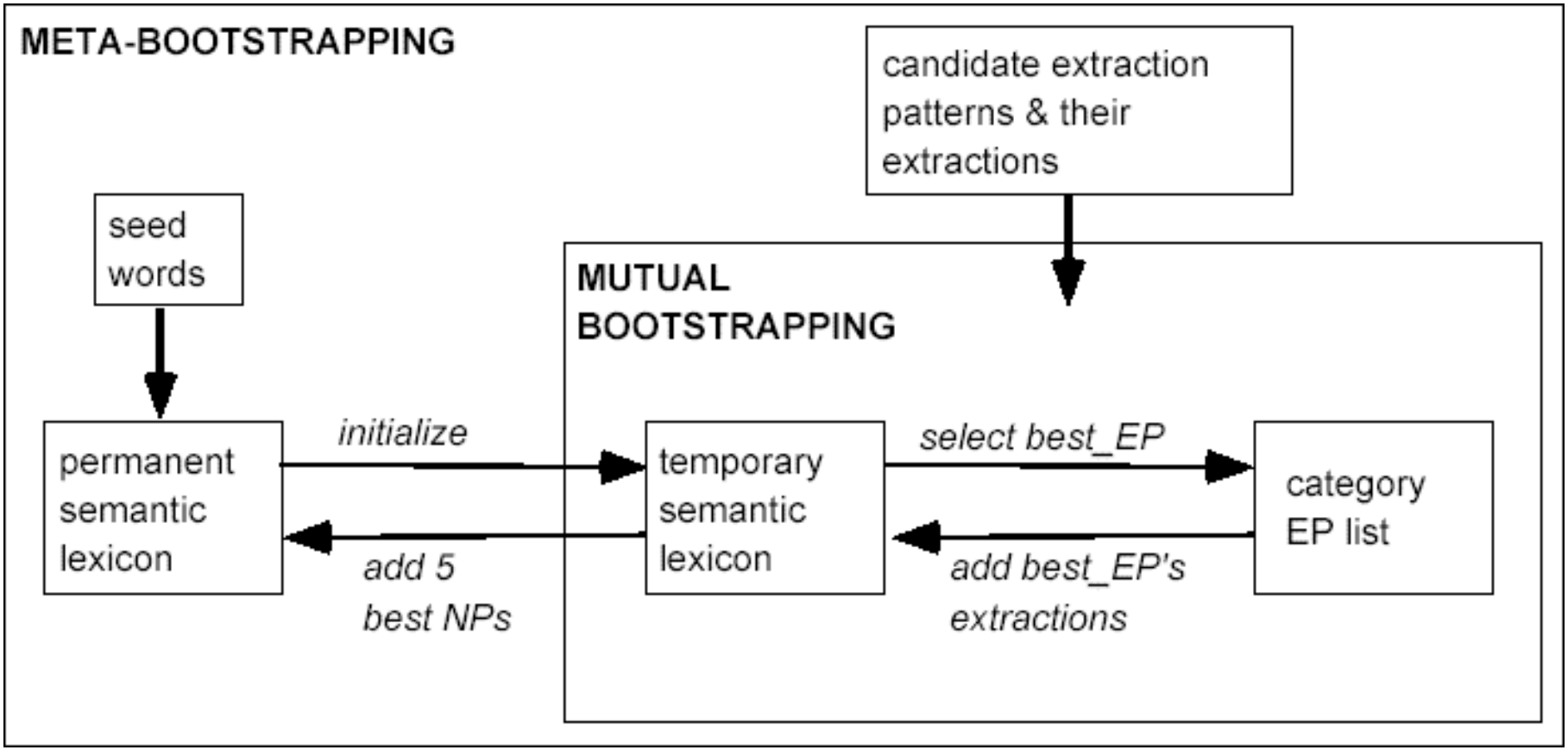
Web Company Patterns
owned by <x>
both as <x>
<x> employed
<x> is distributor
<x> positioning
marks of <x>
motivated <x>
<x> trust company
sold to <x>
devoted to <x>
<x> consolidated stmts.
<x> thrive
message to <x>
<x> is obligations
<x> request information
<x> is foundation
<x> has positions
incorporated as <x>
offices of <x>
<x> required to meet

Level 1: Mutual Bootstrapping

Best pattern	"headquartered in <x>" (F=3,N=4)
Known locations	*nicaragua*
New locations	*san miguel, chapare region, san miguel city*
Best pattern	"gripped <x>" (F=2,N=2)
Known locations	*colombia, guatemala*
New locations	none
Best pattern	"downed in <x>" (F=3,N=6)
Known locations	*nicaragua, san miguel*, city*
New locations	*area, usulutan region, soyapango*
Best pattern	"to occupy <x>" (F=4,N=6)
Known locations	*nicaragua, town*
New locations	*small country, this northern area, san sebastian neighborhood, private property*
Best pattern	"shot in <x>" (F=5,N=12)
Known locations	*city, soyapango**
New locations	*jauja, central square, head, clash, back, central mountain region, air, villa el_salvador district, northwestern guatemala, left side*

- Drift can occur.
- It only takes one bad apple to spoil the barrel.
- Example: head
- Introduce level 2 bootstrapping to prevent drift.

Level 2: Meta-Bootstrapping



Evaluation

Recall/Precision (%)	*Baseline*	*Lexicon*	*Union*
Web Company	10/32	18/47	18/45
Web Location	11/98	51/77	54/74
Web Title	6/100	46/66	47/62

Collins&Singer: CoTraining

- Similar back and forth between
  - an extraction algorithm and
  - a lexicon
- New: They use word-internal features
  - Is the word all caps? (IBM)
  - Is the word all caps with at least one period? (N.Y.)
  - Non-alphabetic character? (AT&T)
  - The constituent words of the phrase ("Bill" is a feature of the phrase "Bill Clinton")
- Classification formalism: Decision Lists

Collins&Singer: Seed Words

full-string=New_York	→	Location
full-string=California	→	Location
full-string=U.S.	→	Location
contains(Mr.)	→	Person
contains(Incorporated)	→	Organization
full-string=Microsoft	→	Organization
full-string=I.B.M.	→	Organization

Note that categories are more generic than in the case of Riloff/Jones.

Collins&Singer: Algorithm

- Train decision rules on current lexicon (initially: seed words).
  - Result: new set of decision rules.
- Apply decision rules to training set
  - Result: new lexicon
- Repeat

Collins&Singer: Results

Learning Algorithm	Accuracy (Clean)	Accuracy (Noise)
Baseline	45.8%	41.8%
EM	83.1%	75.8%
(Yarowsky 95)	81.3%	74.1%
Yarowsky-cautious	91.2%	83.2%
DL-CoTrain	91.3%	83.3%
CoBoost	91.1%	83.1%

Per-token evaluation?

Lexica: Limitations

- Named entity recognition is more than lookup in a list.
- Linguistic variation
  - Manage, manages, managed, managing
- Non-linguistic variation
  - Human gene MYH6 in lexicon, MYH7 in text
- Ambiguity
  - What if a phrase has two different semantic classes?
  - Bioinformatics example: gene/protein metonymy

Lexica: Limitations - Ambiguity

- Metonymy is a widespread source of ambiguity.
- Metonymy: A figure of speech in which one word or phrase is substituted for another with which it is closely associated. (king – crown)
- Gene/protein metonymy
  - The gene name is often used for its protein product.
  - TIMP1 inhibits the HIV protease.
  - TIMP1 could be a gene or protein.
  - Important difference if you are searching for TIMP1 protein/protein interactions.
- Some form of disambiguation necessary to identify correct sense.

Discussion

Partial resources often available.

- E.g., you have a gazetteer, you want to extend it to a new geographic area.

- Some manual post-editing necessary for high-quality.
- Semi-automated approaches offer good coverage with much reduced human effort.
- Drift not a problem in practice if there is a human in the loop anyway.
- Approach that can deal with diverse evidence preferable.
- Hand-crafted features (period for “N.Y.”) help a lot.

Terminology Acquisition

- Goal: find heretofore unknown noun phrases in a text corpus (similar to lexicon construction)
- Lexicon construction
  - Emphasis on finding noun phrases in a specific semantic class (companies)
  - Application: Information extraction
- Terminology Acquisition
  - Emphasis on term normalization (e.g., viral and bacterial infections -> viral_infection)
  - Applications: translation dictionaries, information retrieval

Lexica For Research Index

- Lexica of which classes would be useful?

References

- Julian **Kupiec**, Jan Pedersen, and Francine Chen. A trainable document summarizer. http://citeseer.nj.nec.com/kupiec95trainable.html
- Julian Kupiec. *Murax: A robust linguistic approach for question answering using an on-line encyclopedia*. In the Proceedings of 16th SIGIR Conference, Pittsburgh, PA, 2001.
- Don R. Swanson: Analysis of Unintended Connections Between Disjoint Science Literatures. SIGIR 1991: 280-289
- Tim Berners Lee on semantic web: http://www.sciam.com/2001/0501issue/0501berners-lee.html
- http://www.xml.com/pub/a/2001/01/24/rdf.html
- Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping (1999) Ellen Riloff, Rosie Jones. Proceedings of the Sixteenth National Conference on Artificial Intelligence
- Unsupervised Models for Named Entity Classification (1999) Michael Collins, Yoram Singer