

CS276B

Web Search and Mining

Lecture 14

Text Mining II

(includes slides borrowed from G. Neumann, M. Venkataramani, R. Altman, L. Hirschman, and D. Radev)

Text Mining

- Previously in Text Mining
 - The General Topic
 - Lexicons
 - Topic Detection and Tracking
 - Question Answering
- Today's Topics
 - Summarization
 - Coreference resolution
 - Biomedical text mining

Summarization

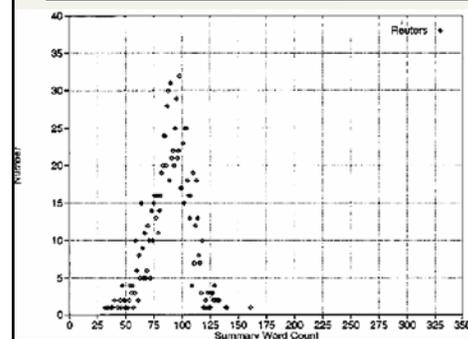
What is a Summary?

- Informative summary
 - Purpose: replace original document
 - Example: executive summary
- Indicative summary
 - Purpose: support decision: do I want to read original document yes/no?
 - Example: Headline, scientific abstract

Why Automatic Summarization?

- Algorithm for reading in many domains is:
 - 1) read summary
 - 2) decide whether relevant or not
 - 3) if relevant: read whole document
- Summary is gate-keeper for large number of documents.
- Information overload
 - Often the summary is all that is read.
- Example from last quarter: summaries of search engine hits
- Human-generated summaries are expensive.

Summary Length (Reuters)



Goldstein
et al.
1999

Characteristics of Summaries People Create for Newswire Stories

- Summary length is approximately constant (Reuters, LA Times)
 - 85-90 words per summary (3-5 sentences)
 - **Note:** Summary length is independent of document length
- 16-17% of the words are proper nouns (named entities)
 - About 3.3 named entities per sentence (20-21 words per sentence)
- 70% of newswire summaries contain the first document sentence
- Summaries usually do not include direct quotes
 - Words such as "said", "adding", "us" and "our" are rare
 - **Note:** Common stopwords might be important
- Summaries are coherent and comprehensible

CMSPSCI 646

Copyright © James Allan

Summarization Algorithms

- Keyword summaries
 - Display most significant keywords
 - Easy to do
 - Hard to read, poor representation of content
- Sentence extraction
 - Extract key sentences
 - Medium hard
 - Summaries often don't read well
 - Good representation of content
- Natural language understanding / generation
 - Build knowledge representation of text
 - Generate sentences summarizing content
 - Hard to do well
- Something between the last two methods?

Sentence Extraction

- Represent each sentence as a feature vector
- Compute score based on features
- Select n highest-ranking sentences
- Present in order in which they occur in text.
- Postprocessing to make summary more readable/concise
 - Eliminate redundant sentences
 - Anaphors/pronouns
 - Delete subordinate clauses, parentheticals
 - Oracle Context

Sentence Extraction: Example

A Trainable Document Summarizer

Julian Knight, Jon Pedersen and Francisco Chen
 School for the Intelligent Agent
 1010 Centre Hill Road, Suite 400, Columbia, SC 29208
 {jkn, pedersen, chen}@lln.cba.cmu.edu

Abstract

• To summarize is to reduce in complexity, and hence in length, while retaining some of the essential qualities of the original.

• This paper focusses on document extracts, a particular kind of computed document summary.

• Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document, which suggests that even shorter extracts may be useful indicative summaries.

• The trends in our results are in agreement with those of Edmundson who used a subjectively weighted combination of features as opposed to training the feature weights using a corpus.

• We have developed a trainable summarization program that is grounded in a sound statistical framework.

Introduction

Summarization is a task that has long interested researchers in the field of natural language processing. The goal is to produce a shorter version of a document that retains the most important information. This is a challenging task because the summary must be both concise and informative. In this paper, we describe a trainable summarization program that is grounded in a sound statistical framework. The program is trained on a corpus of document extracts and is able to generate summaries that are as informative as the full text of the original document.

Method

The summarization program is trained on a corpus of document extracts. The program is trained to generate summaries that are as informative as the full text of the original document. The program is trained to generate summaries that are as informative as the full text of the original document.

- Sigr95 paper on summarization by Kupiec, Pedersen, Chen
- **Trainable** sentence extraction
- Proposed algorithm is applied to its own description (the paper)

Sentence Extraction: Example

- To summarize is to reduce in complexity, and hence in length, while retaining some of the essential qualities of the original.
- This paper focusses on document extracts, a particular kind of computed document summary.
- Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document, which suggests that even shorter extracts may be useful indicative summaries.
- The trends in our results are in agreement with those of Edmundson who used a subjectively weighted combination of features as opposed to training the feature weights using a corpus.
- We have developed a trainable summarization program that is grounded in a sound statistical framework.

Feature Representation

- Fixed-phrase feature
 - Certain phrases indicate summary, e.g. "in summary"
- Paragraph feature
 - Paragraph initial/final more likely to be important.
- Thematic word feature
 - Repetition is an indicator of importance
- Uppercase word feature
 - Uppercase often indicates named entities. (Taylor)
- Sentence length cut-off
 - Summary sentence should be > 5 words.

Feature Representation (cont.)

- Sentence length cut-off
 - Summary sentences have a minimum length.
- Fixed-phrase feature
 - True for sentences with indicator phrase
 - "in summary", "in conclusion" etc.
- Paragraph feature
 - Paragraph initial/medial/final
- Thematic word feature
 - Do any of the most frequent content words occur?
- Uppercase word feature
 - Is uppercase thematic word introduced?

Training

- Hand-label sentences in training set (good/bad summary sentences)
- Train classifier to distinguish good/bad summary sentences
- Model used: Naïve Bayes

$$P(s \in \mathcal{S} | F_1, F_2, \dots, F_k) = \frac{\prod_{j=1}^k P(F_j | s \in \mathcal{S}) P(s \in \mathcal{S})}{\prod_{j=1}^k P(F_j)}$$

- Can rank sentences according to score and show top n to user.

Evaluation

- Compare extracted sentences with sentences in abstracts

Direct Sentence Matches	451	79%
Direct Joins	19	3%
Unmatchable Sentences	50	9%
Incomplete Single Sentences	21	4%
Incomplete Joins	27	5%
Total Manual Summary sents	568	

Evaluation of features

- Baseline (choose first n sentences): 24%
- Overall performance (42-44%) not very good.
- However, there is more than one good summary.

Feature	Individual Sents Correct	Cumulative Sents Correct
Paragraph	163 (33%)	163 (33%)
Fixed Phrases	145 (29%)	209 (42%)
Length Cut-off	121 (24%)	217 (44%)
Thematic Word	101 (20%)	209 (42%)
Uppercase Word	100 (20%)	211 (42%)

Multi-Document (MD) Summarization

- Summarize more than one document
- **Why is this harder?**
- But benefit is large (can't scan 100s of docs)
- To do well, need to adopt more specific strategy depending on document set.
- Other components needed for a production system, e.g., manual post-editing.
- DUC: government sponsored bake-off
 - 200 or 400 word summaries
 - Longer → easier

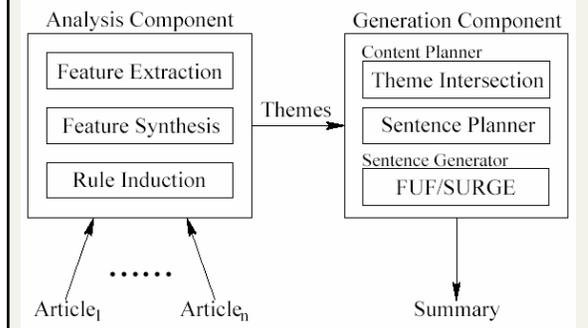
Types of MD Summaries

- Single event/person tracked over a long time period
 - Elizabeth Taylor's bout with pneumonia
 - Give extra weight to character/event
 - May need to include **outcome (dates!)**
- Multiple events of a similar nature
 - Marathon runners and races
 - More broad brush, ignore dates
- An issue with related events
 - Gun control
 - Identify key concepts and select sentences accordingly

Determine MD Summary Type

- First, determine which type of summary to generate
- Compute all pairwise similarities
- Very dissimilar articles → multi-event (marathon)
- Mostly similar articles
 - Is most frequent concept named entity?
 - Yes → single event/person (Taylor)
 - No → issue with related events (gun control)

MultiGen Architecture (Columbia)



Generation

- Ordering according to date
- Intersection
 - Find concepts that occur repeatedly in a time chunk
- Sentence generator

Processing

- Selection of good summary sentences
- Elimination of redundant sentences
- Replace anaphors/pronouns with noun phrases they refer to
 - Need coreference resolution
- Delete non-central parts of sentences

Newsblaster (Columbia)

[Powell tells UN Iraq hid arms, deceived weapons inspectors](#) (84 articles)



Britain is likely to introduce a new resolution authorizing the use of force against Iraq after top weapons inspectors return from Baghdad and report to the Security Council on Feb. 14, a British diplomat said Thursday. Chief arms inspectors, in pivotal talks this weekend, expect to gain Iraqi concessions on practical issues, such as reconnaissance flights, but believe Baghdad must finally meet their demand for hard evidence on weapons programs, a senior official said Thursday. A senior Iraqi official said Thursday that an Iraqi weapons expert had submitted to a private interview with arms inspectors, a sign of progress in the deadlock over weapons inspections. He said council members are looking to see a change in attitude from Iraq, which the United States says is concealing illegal weapons programs. Secretary of State Colin Powell "made as strong a case as one could make that Iraq was concealing weapons-related activities" says former weapons inspector and CBS News consultant Steven Black. Despite intense pressure by Blair, French President Jacques Chirac said he remained steadfastly opposed to war against Baghdad without giving weapons inspectors searching for banned weapons as much time they need to do their work.

Query-Specific Summarization

- So far, we've look at **generic summaries**.
- A generic summary makes no assumption about the reader's interests.
- Query-specific summaries are specialized for a single information need, the query.
- Summarization is much easier if we have a description of what the user wants.
- Recall from last quarter:
 - Google-type excerpts – simply show keywords in context

Genre

- Some genres are easy to summarize
 - Newswire stories
 - Inverted pyramid structure
 - The first n sentences are often the best summary of length n
- Some genres are hard to summarize
 - Long documents (novels, the bible)
 - Scientific articles?
- Trainable summarizers are genre-specific.

Discussion

- Correct parsing of document format is critical.
 - Need to know headings, sequence, etc.
- Limits of current technology
 - Some good summaries require natural language understanding
 - Example: President Bush's nominees for ambassadorships
 - Contributors to Bush's campaign
 - Veteran diplomats
 - Others

Coreference Resolution

Coreference

- Two noun phrases referring to the same entity are said to **corefer**.
- Example: Transcription from **RL95-2** is mediated through an ERE element at the 5-flanking region of the **gene**.
- Coreference resolution is important for many text mining tasks:
 - Information extraction
 - Summarization
 - First story detection

Types of Coreference

- Noun phrases: Transcription from **RL95-2** ... the **gene** ...
- Pronouns: **They** induced apoptosis.
- Possessives: ... induces **their** rapid dissociation ...
- Demonstratives: **This gene** is responsible for Alzheimer's

Preferences in pronoun interpretation

- Recency: *John has an Integra. Bill has a legend. Mary likes to drive it.*
- Grammatical role: *John went to the Acura dealership with Bill. He bought an Integra.*
- (?) *John and Bill went to the Acura dealership. He bought an Integra.*
- Repeated mention: *John needed a car to go to his new job. He decided that he wanted something sporty. Bill went to the Acura dealership with him. He bought an Integra.*

Preferences in pronoun interpretation

- Parallelism: *Mary went with Sue to the Acura dealership. Sally went with her to the Mazda dealership.*
- ??? *Mary went with Sue to the Acura dealership. Sally told her not to buy anything.*
- Verb semantics: *John telephoned Bill. He lost his pamphlet on Acuras. John criticized Bill. He lost his pamphlet on Acuras.*

An algorithm for pronoun resolution

- Two steps: discourse model update and pronoun resolution.
- Saliency values are introduced when a noun phrase that evokes a new entity is encountered.
- Saliency factors: set empirically.

Saliency weights in Lappin and Leass

Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object and oblique complement emphasis	40
Non-adverbial emphasis	50
Head noun emphasis	80

Lappin and Leass (cont'd)

- Recency: weights are cut in half after each sentence is processed.
- Examples:
 - An Acura Integra is parked in the lot.
 - There is an Acura Integra parked in the lot.
 - John parked an Acura Integra in the lot.
 - John gave Susan an Acura Integra.
 - In his Acura Integra, John showed Susan his new CD player.

Algorithm

1. Collect the potential referents (up to four sentences back).
2. Remove potential referents that do not agree in number or gender with the pronoun.
3. Remove potential referents that do not pass intrasentential syntactic coreference constraints.
4. Compute the total saliency value of the referent by adding any applicable values for role parallelism (+35) or cataphora (-175).
5. Select the referent with the highest saliency value. In case of a tie, select the closest referent in terms of string position.

Observations

- Lappin & Leass - tested on computer manuals - 86% accuracy on unseen data.
- Another well known theory is Centering (Grosz, Joshi, Weinstein), which has an additional concept of a "center". (More of a theoretical model; less empirical confirmation.)

Biological Text Mining

Biological Terminology: A Challenge

- Large number of entities (genes, proteins etc)
- Evolving field, no widely followed standards for terminology → **Rapid Change, Inconsistency**
- Ambiguity: Many (short) terms with multiple meanings (eg, CAN)
- Synonymy: ARA70, ELE1 alpha, RFG
- High complexity → **Complex phrases**

What are the concepts of interest?

- Genes (D4DR)
- Proteins (hexosaminidase)
- Compounds (acetaminophen)
- Function (lipid metabolism)
- Process (apoptosis = cell death)
- Pathway (Urea cycle)
- Disease (Alzheimer's)

Complex Phrases

- Characterization of the repressor function of the nuclear orphan receptor retinoid receptor-related testis-associated receptor/germ nuclear factor

Inconsistency

- No consistency across species

	Protease	Inhibitor	signal
Fruit fly	Tolloid	Sog	dpp
Frog	Xolloid	Chordin	BMP2/BMP4
Zebrafish	Minifin	Chordino	swirl

Rapid Change

Mouse Genome Nomenclature Events 8/25

Ch. Symbol	Gene Name	IR	Protein	Protein Symbol
1	Acad10	Academy-10-like domain containing 10	Acad10	Acad10
1	Acad11	Academy-11-like domain containing 11	Acad11	Acad11
1	Acad12	Academy-12-like domain containing 12	Acad12	Acad12
1	Acad13	Academy-13-like domain containing 13	Acad13	Acad13
1	Acad14	Academy-14-like domain containing 14	Acad14	Acad14
1	Acad15	Academy-15-like domain containing 15	Acad15	Acad15
1	Acad16	Academy-16-like domain containing 16	Acad16	Acad16
1	Acad17	Academy-17-like domain containing 17	Acad17	Acad17
1	Acad18	Academy-18-like domain containing 18	Acad18	Acad18
1	Acad19	Academy-19-like domain containing 19	Acad19	Acad19
1	Acad20	Academy-20-like domain containing 20	Acad20	Acad20
1	Acad21	Academy-21-like domain containing 21	Acad21	Acad21
1	Acad22	Academy-22-like domain containing 22	Acad22	Acad22
1	Acad23	Academy-23-like domain containing 23	Acad23	Acad23
1	Acad24	Academy-24-like domain containing 24	Acad24	Acad24
1	Acad25	Academy-25-like domain containing 25	Acad25	Acad25
1	Acad26	Academy-26-like domain containing 26	Acad26	Acad26
1	Acad27	Academy-27-like domain containing 27	Acad27	Acad27
1	Acad28	Academy-28-like domain containing 28	Acad28	Acad28
1	Acad29	Academy-29-like domain containing 29	Acad29	Acad29
1	Acad30	Academy-30-like domain containing 30	Acad30	Acad30
1	Acad31	Academy-31-like domain containing 31	Acad31	Acad31
1	Acad32	Academy-32-like domain containing 32	Acad32	Acad32
1	Acad33	Academy-33-like domain containing 33	Acad33	Acad33
1	Acad34	Academy-34-like domain containing 34	Acad34	Acad34
1	Acad35	Academy-35-like domain containing 35	Acad35	Acad35
1	Acad36	Academy-36-like domain containing 36	Acad36	Acad36
1	Acad37	Academy-37-like domain containing 37	Acad37	Acad37
1	Acad38	Academy-38-like domain containing 38	Acad38	Acad38
1	Acad39	Academy-39-like domain containing 39	Acad39	Acad39
1	Acad40	Academy-40-like domain containing 40	Acad40	Acad40
1	Acad41	Academy-41-like domain containing 41	Acad41	Acad41
1	Acad42	Academy-42-like domain containing 42	Acad42	Acad42
1	Acad43	Academy-43-like domain containing 43	Acad43	Acad43
1	Acad44	Academy-44-like domain containing 44	Acad44	Acad44
1	Acad45	Academy-45-like domain containing 45	Acad45	Acad45
1	Acad46	Academy-46-like domain containing 46	Acad46	Acad46
1	Acad47	Academy-47-like domain containing 47	Acad47	Acad47
1	Acad48	Academy-48-like domain containing 48	Acad48	Acad48
1	Acad49	Academy-49-like domain containing 49	Acad49	Acad49
1	Acad50	Academy-50-like domain containing 50	Acad50	Acad50
1	Acad51	Academy-51-like domain containing 51	Acad51	Acad51
1	Acad52	Academy-52-like domain containing 52	Acad52	Acad52
1	Acad53	Academy-53-like domain containing 53	Acad53	Acad53
1	Acad54	Academy-54-like domain containing 54	Acad54	Acad54
1	Acad55	Academy-55-like domain containing 55	Acad55	Acad55
1	Acad56	Academy-56-like domain containing 56	Acad56	Acad56
1	Acad57	Academy-57-like domain containing 57	Acad57	Acad57
1	Acad58	Academy-58-like domain containing 58	Acad58	Acad58
1	Acad59	Academy-59-like domain containing 59	Acad59	Acad59
1	Acad60	Academy-60-like domain containing 60	Acad60	Acad60
1	Acad61	Academy-61-like domain containing 61	Acad61	Acad61
1	Acad62	Academy-62-like domain containing 62	Acad62	Acad62
1	Acad63	Academy-63-like domain containing 63	Acad63	Acad63
1	Acad64	Academy-64-like domain containing 64	Acad64	Acad64
1	Acad65	Academy-65-like domain containing 65	Acad65	Acad65
1	Acad66	Academy-66-like domain containing 66	Acad66	Acad66
1	Acad67	Academy-67-like domain containing 67	Acad67	Acad67
1	Acad68	Academy-68-like domain containing 68	Acad68	Acad68
1	Acad69	Academy-69-like domain containing 69	Acad69	Acad69
1	Acad70	Academy-70-like domain containing 70	Acad70	Acad70
1	Acad71	Academy-71-like domain containing 71	Acad71	Acad71
1	Acad72	Academy-72-like domain containing 72	Acad72	Acad72
1	Acad73	Academy-73-like domain containing 73	Acad73	Acad73
1	Acad74	Academy-74-like domain containing 74	Acad74	Acad74
1	Acad75	Academy-75-like domain containing 75	Acad75	Acad75
1	Acad76	Academy-76-like domain containing 76	Acad76	Acad76
1	Acad77	Academy-77-like domain containing 77	Acad77	Acad77
1	Acad78	Academy-78-like domain containing 78	Acad78	Acad78
1	Acad79	Academy-79-like domain containing 79	Acad79	Acad79
1	Acad80	Academy-80-like domain containing 80	Acad80	Acad80
1	Acad81	Academy-81-like domain containing 81	Acad81	Acad81
1	Acad82	Academy-82-like domain containing 82	Acad82	Acad82
1	Acad83	Academy-83-like domain containing 83	Acad83	Acad83
1	Acad84	Academy-84-like domain containing 84	Acad84	Acad84
1	Acad85	Academy-85-like domain containing 85	Acad85	Acad85
1	Acad86	Academy-86-like domain containing 86	Acad86	Acad86
1	Acad87	Academy-87-like domain containing 87	Acad87	Acad87
1	Acad88	Academy-88-like domain containing 88	Acad88	Acad88
1	Acad89	Academy-89-like domain containing 89	Acad89	Acad89
1	Acad90	Academy-90-like domain containing 90	Acad90	Acad90
1	Acad91	Academy-91-like domain containing 91	Acad91	Acad91
1	Acad92	Academy-92-like domain containing 92	Acad92	Acad92
1	Acad93	Academy-93-like domain containing 93	Acad93	Acad93
1	Acad94	Academy-94-like domain containing 94	Acad94	Acad94
1	Acad95	Academy-95-like domain containing 95	Acad95	Acad95
1	Acad96	Academy-96-like domain containing 96	Acad96	Acad96
1	Acad97	Academy-97-like domain containing 97	Acad97	Acad97
1	Acad98	Academy-98-like domain containing 98	Acad98	Acad98
1	Acad99	Academy-99-like domain containing 99	Acad99	Acad99
1	Acad100	Academy-100-like domain containing 100	Acad100	Acad100

In 1 week, 166 events involving change of nomenclature

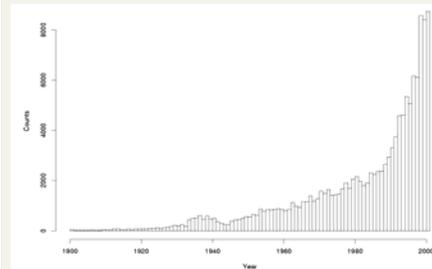
L. Hirschmann

Where's the Information?

- Information about function and behavior is mainly in text form (scientific articles)
- Medical Literature on line.
- Online database of published literature since 1966 = Medline = PubMed resource
- 4,000 journals
- 10,000,000+ articles (most with abstracts)
- www.ncbi.nlm.nih.gov/PubMed/

Curators Cannot Keep Up with the Literature!

FlyBase References By Year

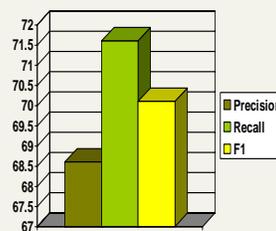


Biomedical Named Entity Recognition

- The list of biomedical entities is growing.
 - New genes and proteins are constantly being discovered, so explicitly enumerating and searching against a list of known entities is not scalable.
 - Part of the difficulty lies in identifying previously unseen entities based on contextual, orthographic, and other clues.
- Biomedical entities don't adhere to strict naming conventions.
 - Common English words such as *period*, *curved*, and *for* are used for gene names.
 - The entity names can be ambiguous. For example, in FlyBase, "clk" is the gene symbol for the "Clock" gene but it also is used as a synonym of the "period" gene.
- Biomedical entity names are ambiguous
 - Experts only agree on whether a word is even a gene or protein 69% of the time. (Krauthammer *et al.*, 2000)

Results of Finkel et al. (2004) MEMM-based BioNER system

- BioNLP task – Identify genes, proteins, DNA, RNA, and cell types



Precision	Recall	F1
68.6%	71.6%	70.1%

$$\text{precision} = \text{tp} / (\text{tp} + \text{fp})$$

$$\text{recall} = \text{tp} / (\text{tp} + \text{fn})$$

$$\text{F1} = \frac{2(\text{precision})(\text{recall})}{(\text{precision} + \text{recall})}$$

Abbreviations in Biology

- Two problems
 - "Coreference"/Synonymy
 - What is PCA an abbreviation for?
 - Ambiguity
 - If PCA has >1 expansions, which is right here?
- Only important concepts are abbreviated.
- Effective way of jump starting terminology acquisition.

Ambiguity Example PCA has >60 expansions

"p-chloroamphetamine" "p-chloroaniline" "p-coumaric acid"
 "p.tochromycin c. complex activity" "para-chloroamphetamine" "parietal cell antibodies" "parietal cell autoantibodies" "paroxysmal cerebellar ataxia" "passive cutaneous anaphylactic" "patient care appraisal"
 "patient controlled analgesia" "patient controlled anesthesia" "pca"
 "pentachloroisole" "percent cortical area" "perchloric acid"
 "perchloric acid" "percutaneous coronary angioplasty" "percutaneous coronary atherectomy" "pericardial artery" "peritoneal carcinomatosis"
 "peritoneal carcinoma" "personal care attendant"
 "phenazine-1-carboxylic acid" "phenylcyclohexylacetic acid"
 "phenylethylhexylamine" "physical capacity assessment" "pig coronary artery" "plate count agar" "pneumococcal capsular antigen" "pole climbing avoidance" "polypeptidase" "polyclonal antibody"
 "polyclonal antisera" "polycyclic aromatic content" "porous coated anatomic" "porous coated total hip arthroplasty" "porous-coated hip arthroplasties" "porous-coated patellar component" "porta-caval anastomosis" "portable clinical analyzer" "portocaval anastomosis"
 "portocaval shunt" "post chipper attachment" "postconceptional age" "posterior cerebral arteries" "posterior communicating artery"
 "posterior cortical atrophy" "posterior crico-arytenoid" "potassium channel activators" "presence of parietal cell" "primary cardiac arrest" "primary congenital aphakia" "principal component analyses"
 "procoagulant" "procoagulant activities" "procoagulant cellular activity" "procoagulatory activity" "prostatic carcinoma" "protein c activator" "prothrombin complex activity" "protocatechuate"
 "protocatechuic acid" "protocaval anastomosis" "pulmonary corpora amygdala" "pyroglutamic acid" "pyrrolidone-5-carboxylic acid"
 "pyrrolidone-2-carboxylic acid" "pyrrolidone-5-carboxylic acid"
 "pyrrolidone-5-carboxylate"

Problem 1: Ambiguity

- “Senses” of an abbreviation are usually not related.
- Long form often occurs at least once in a document.
- Disambiguating abbreviations is easy.

Problem 2: “Coreference”

- Goal: Establish that abbreviation and long form are coreferring.
- Strategy:
 - Treat each pattern $w^*(c^*)$ as a hypothesis.
 - Reject hypothesis if well-formedness conditions are not met.
 - Accept otherwise.

Approach

- Generate a **set of good candidate alignments**
- Build feature representation
- Classify feature representation using logistic regression classifier (or SVM would be equally good) to choose best one.

Features for Classifier

- Describes the abbreviation.
 - Lower Abbrev
- Describes the alignment.
 - Aligned
 - Unused Words
 - AlignsPerWord
- Describes the characters aligned.
 - WordBegin
 - WordEnd
 - SyllableBoundary
 - HasNeighbor

Text-Enhanced Sequence Homology Detection

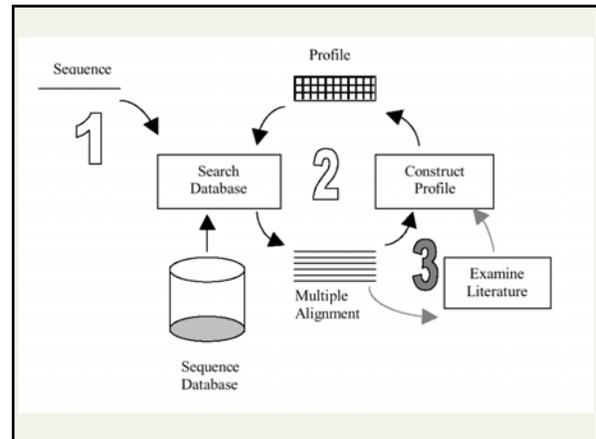
- Obtaining sequence information is easy; characterizing sequences is hard.
- Organisms share a common basis of genes and pathways.
- Information can be predicted for a novel sequence based on sequence similarity:
 - Function
 - Cellular role
 - Structure
- Nearly all information about functions is in textual literature

PSI-BLAST

- Used to detect protein sequence homology. (Iterated version of universally used BLAST program.)
- Searches a database for sequences with high sequence similarity to a query sequence.
- Creates a profile from similar sequences and iterates the search to improve sensitivity.

Text-Enhanced Homology Search (Chang, Raychaudhuri, Altman)

- PSI-BLAST Problem: Profile Drift
 - At each iteration, could find non-homologous (false positive) proteins.
 - False positives create a poor profile, leading to more false positives.
- OBSERVATION: Sequence similarity is only one indicator of homology.
 - More clues, e.g. protein functional role, exist in the literature.
- SOLUTION: incorporate MEDLINE text into PSI-BLAST matching process.

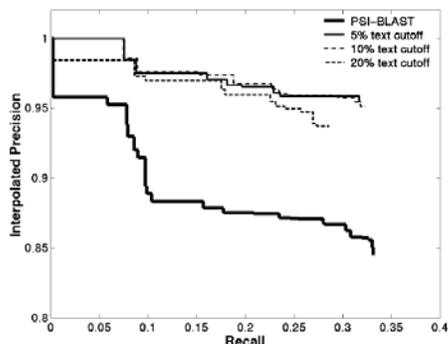


Modification to PSI-BLAST

- Before including a sequence, measure similarity of literature. Throw away sequences with least similar literatures to avoid drift.
- Literature is obtained from SWISS-PROT gene annotations to MEDLINE (text, keywords).
- Define domain-specific “stop” words (< 3 sequences or >85,000 sequences) = 80,479 out of 147,639.
- Use similarity metric between literatures (for genes) based on word vector cosine.

Evaluation

- Created families of homologous proteins based on SCOP (gold standard site for homologous proteins-- <http://scop.berkeley.edu/>)
- Select one sequence per protein family:
 - Families must have \geq five members
 - Associated with at least four references
 - Select sequence with worst performance on a non-iterated BLAST search
- Compared homology search results from original and modified PSI-BLAST.



Resources

- A Trainable Document Summarizer (1995) Julian Kupiec, Jan Pedersen, Francine Chen *Research and Development in Information Retrieval*
- [The Columbia Multi-Document Summarizer for DUC 2002](http://www.ldc.upenn.edu/Projects/ACE/PHASE2/Annotation/guidelines/EDT/coreference.shtml) K. McKeown, D. Evans, A. Nenkova, R. Barzilay, V. Hatzivassiloglou, B. Schiffman, S. Blair-Goldensohn, J. Klavans, S. Sigelman, Columbia University
- Coreference: detailed discussion of the term: <http://www.ldc.upenn.edu/Projects/ACE/PHASE2/Annotation/guidelines/EDT/coreference.shtml>
- <http://www.smi.stanford.edu/projects/helix/psb01/chang.pdf> Pac Symp Biocomput. 2001;:374-83. PMID: 11262956
- <http://www.smi.stanford.edu/projects/helix/psb03> Genome Res 2002 Oct;12(10):1582-90 Using text analysis to identify functionally coherent gene groups. Raychaudhuri S, Schutze H, Altman RB
- Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair. 2004. Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web. Joint Workshop on Natural Language Processing in Biomedicine and its Applications at Coling 2004.