Drexel UNIVERSITY College of Information Science and Technology

Ph.D. Proposal Defense:
Semantic Annotation and Summarization of Biomedical Literature

Lawrence H. Reeve

Committee Members:
Dr. Hyoil Han (Advisor/Chair)
Dr. Ari D. Brooks
Dr. Xia Lin
Dr. Ani Nenkova
Dr. Il-Yeol Song

Overview

- Research Question / Approach
- Motivation
- Background
  - UMLS, Evaluation Corpus
- Approach
- Evaluation
- Literature Review
- Research Plan

Hypothesis

- The use of domain-specific concepts can be used to identify important areas within a text in order to perform extractive text summarization.

Research Sub-Questions

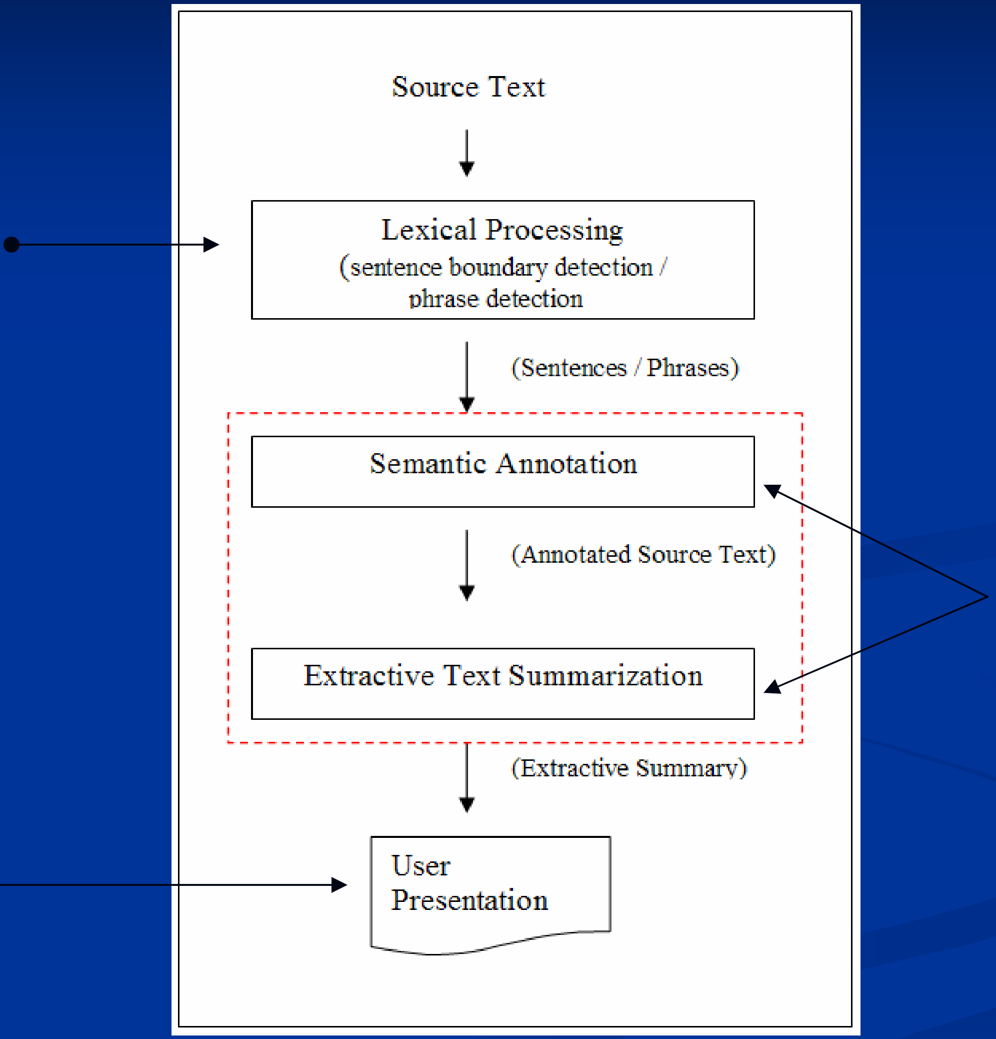
- Can the UMLS Semantic Network and UMLS Metathesaurus be used in the biomedical domain to identify important areas within a biomedical text?
- Can existing text summarization approaches be adapted in new ways to utilize domain-specific concepts?
- Can biomedical semantic annotation time be improved for use in an online environment?

Research Approach

- Construct an integrated system to perform semantic annotation and summarization of biomedical text which has performance competitive with existing systems.
  - Semantic Annotation: Research issues for improving performance in an online environment
  - Text Summarization: Research if use of concepts can be used for effective text summarization

Proposed Integrated System

MetaMap,
LingPipe,
Sliding Window,
other methods



Source Text

Lexical Processing
(sentence boundary detection / phrase detection

(Sentences / Phrases)

Semantic Annotation

(Annotated Source Text)

Extractive Text Summarization

(Extractive Summary)

User Presentation

Focus of research

Text File

Expected Contributions

- Biomedical text annotator:
  - New design approach: concept filtering
  - Use two new methods for:
    - Coverage metric: IDF-based scoring method
    - Coherence metric: skip-bigram method to allow inexact matches

- Single-document biomedical text summarizer
  - use domain concepts to identify relevant sentences for extraction
  - develop two novel algorithms:
    - Concept chaining
    - Frequency Distribution

- Explore the characteristics of biomedical text (e.g., summary size, sections)

- Generate summary model corpus of biomedical text & use for evaluation

Expected Significance

- Biomedical Text Summarizer:
  - New algorithms designed for use with domain-specific concepts can be used to improve system-generated summaries over existing summarization approaches
  - Concepts, rather than terms, can be used to present a UI to personalize biomedical summaries
- Semantic Annotation:
  - Improved time to perform biomedical text annotation will allow use in an online environment

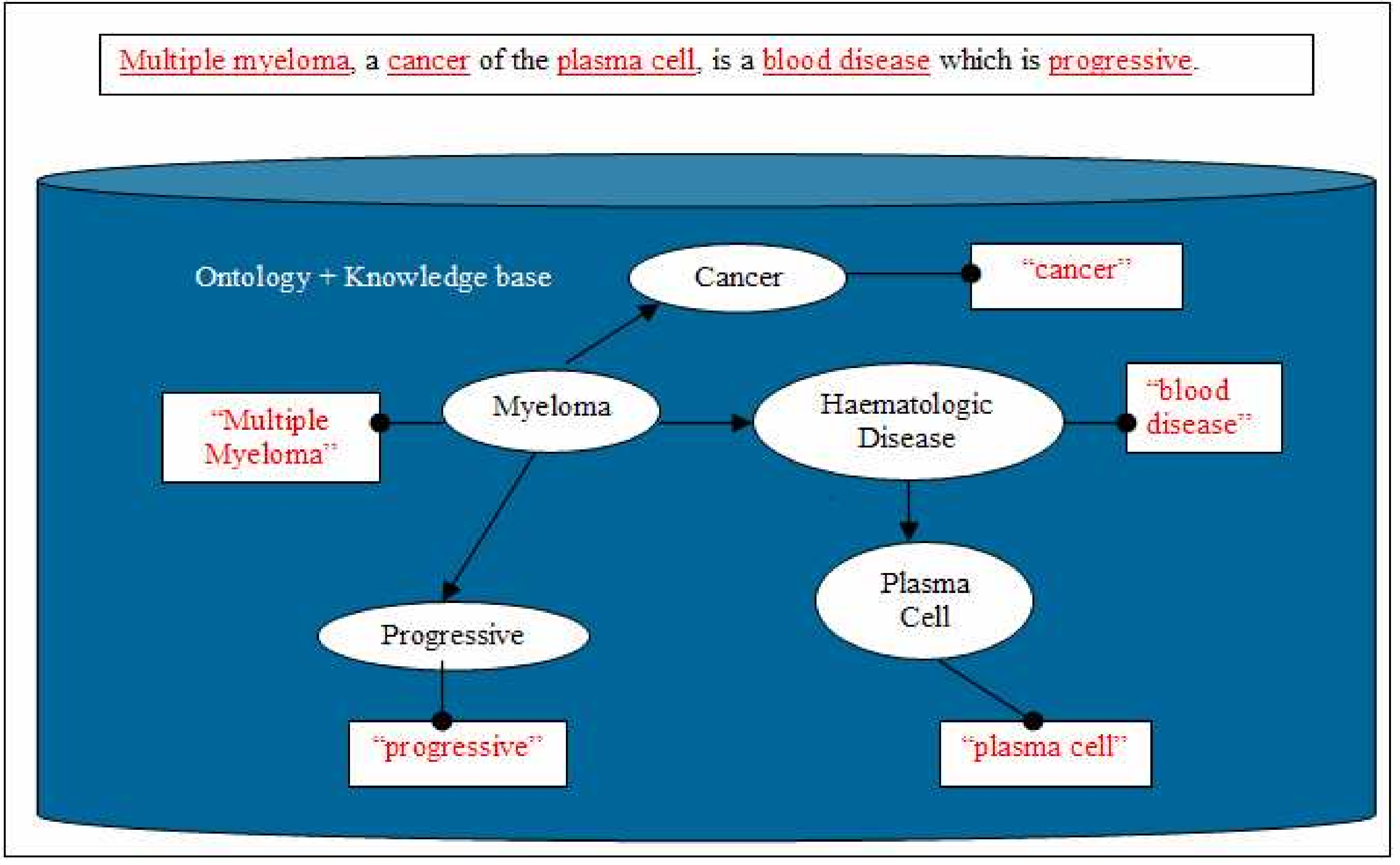
Motivation

- Biomedical Semantic Annotation
  - *Purpose*: add machine-understandable meaning by finding domain-specific concepts within free-form texts
  - *Uses*: (a) synonym merging, and (b) semantic filtering
  - *How*: Domain-specific resources: ontology, thesaurus
- Text Summarization
  - *Purpose*: Reduce amount of data required to understand main points of a text
  - *Uses*: Quick overviews of research, focus on certain details
  - *How*: Extract subset of sentences from source text

Background

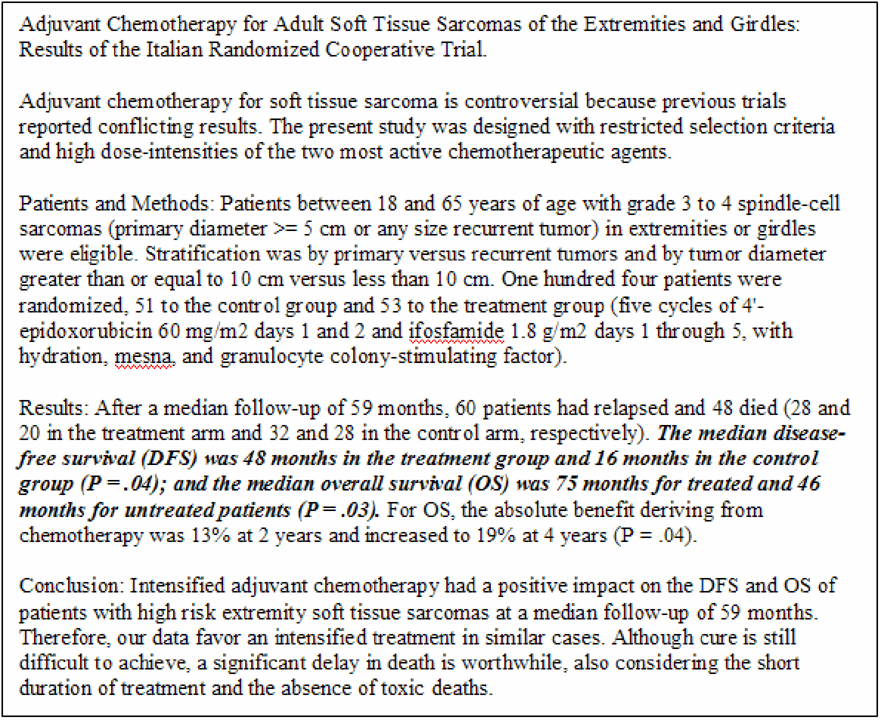
(Semantic Annotation, Text Summarization, UMLS, MetaMap, Evaluation Corpus)

Semantic Annotation



Ovals: biomedical concepts; *Rectangles*: instances of the biomedical concepts from the sentence; *Directed lines* - relationships between the concepts; *Undirected lines* – concept instances linked to concepts.

Text Summarization



Adjuvant Chemotherapy for Adult Soft Tissue Sarcomas of the Extremities and Girdles: Results of the Italian Randomized Cooperative Trial.

Adjuvant chemotherapy for soft tissue sarcoma is controversial because previous trials reported conflicting results. The present study was designed with restricted selection criteria and high dose-intensities of the two most active chemotherapeutic agents.

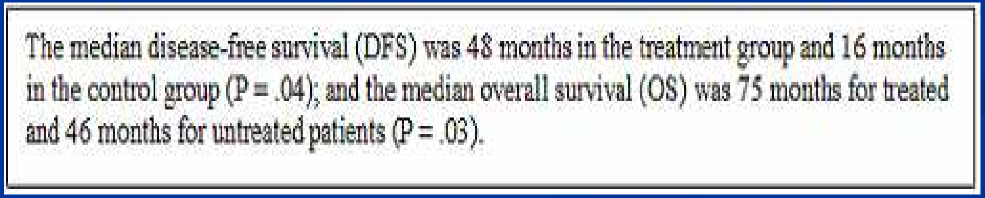
Patients and Methods: Patients between 18 and 65 years of age with grade 3 to 4 spindle-cell sarcomas (primary diameter >= 5 cm or any size recurrent tumor) in extremities or girdles were eligible. Stratification was by primary versus recurrent tumors and by tumor diameter greater than or equal to 10 cm versus less than 10 cm. One hundred four patients were randomized, 51 to the control group and 53 to the treatment group (five cycles of 4'-epidoxorubicin 60 mg/m2 days 1 and 2 and ifosfamide 1.8 g/m2 days 1 through 5, with hydration, mesna, and granulocyte colony-stimulating factor).

Results: After a median follow-up of 59 months, 60 patients had relapsed and 48 died (28 and 20 in the treatment arm and 32 and 28 in the control arm, respectively). ***The median disease-free survival (DFS) was 48 months in the treatment group and 16 months in the control group (P = .04); and the median overall survival (OS) was 75 months for treated and 46 months for untreated patients (P = .03).*** For OS, the absolute benefit deriving from chemotherapy was 13% at 2 years and increased to 19% at 4 years (P = .04).

Conclusion: Intensified adjuvant chemotherapy had a positive impact on the DFS and OS of patients with high risk extremity soft tissue sarcomas at a median follow-up of 59 months. Therefore, our data favor an intensified treatment in similar cases. Although cure is still difficult to achieve, a significant delay in death is worthwhile, also considering the short duration of treatment and the absence of toxic deaths.

Source Text

Sentence extraction



The median disease-free survival (DFS) was 48 months in the treatment group and 16 months in the control group (P = .04); and the median overall survival (OS) was 75 months for treated and 46 months for untreated patients (P = .03).

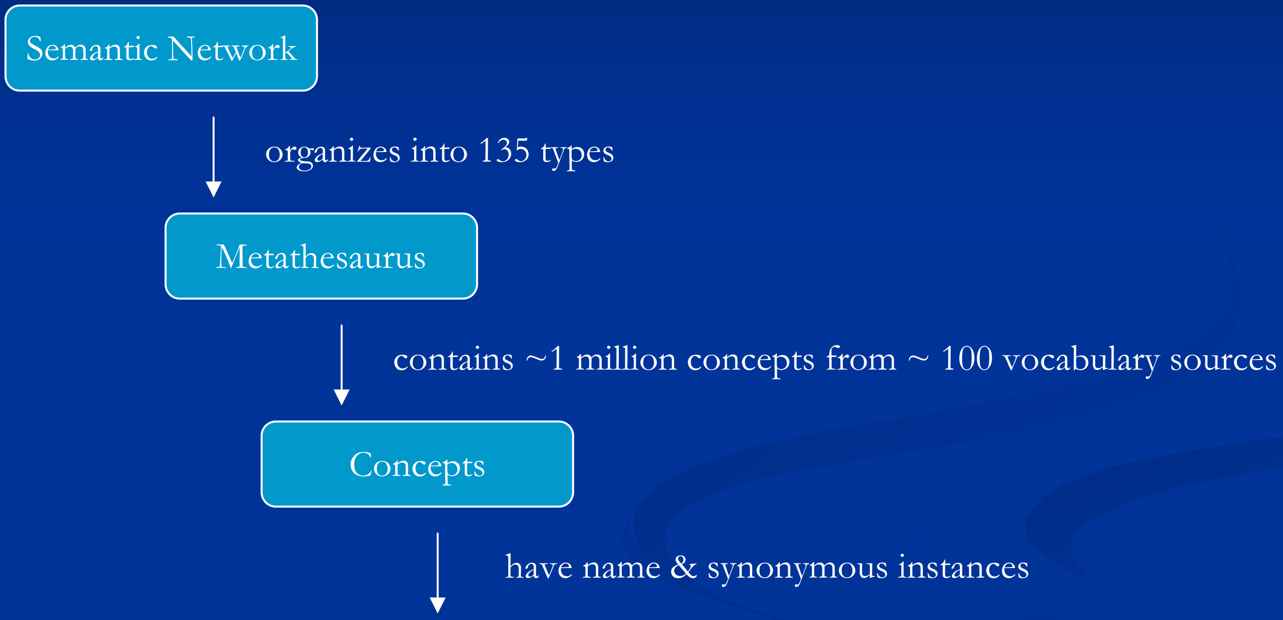
Summarized Text

Generating summaries in presence of author's abstract

- No ideal summary – dependent on user's information needs
- Abstract may be missing content from the full-text
- Question-answering systems for providing personalized information
- Automatic/semi-automatic summary generation for scaling commercial abstract services
- Evaluation of sentence selection methods that may be useful for use in multi-document summarization

UMLS

(Unified Medical Language System – U.S. National Library of Medicine)



Semantic Network

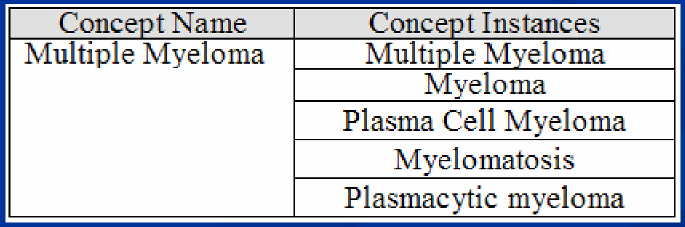
organizes into 135 types

Metathesaurus

contains ~1 million concepts from ~ 100 vocabulary sources

Concepts

have name & synonymous instances

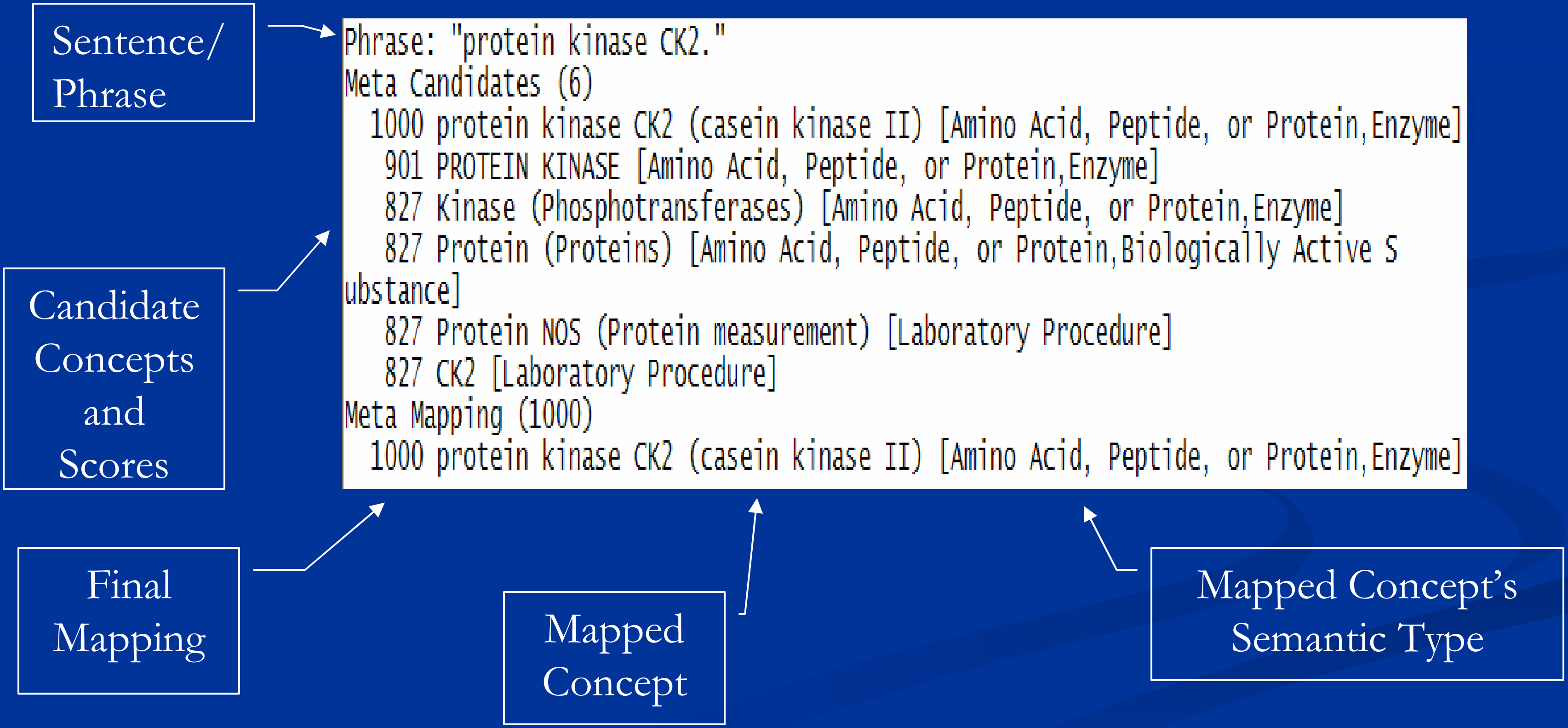


Concept Name	Concept Instances
Multiple Myeloma	Multiple Myeloma
	Myeloma
	Plasma Cell Myeloma
	Myelomatosis
	Plasmacytic myeloma

MetaMap

(U.S. National Library of Medicine)

- Maps free form text into UMLS concepts



Sentence/ Phrase

Candidate Concepts and Scores

Final Mapping

Mapped Concept

Mapped Concept’s Semantic Type

```
Phrase: "protein kinase CK2."
Meta Candidates (6)
  1000 protein kinase CK2 (casein kinase II) [Amino Acid, Peptide, or Protein,Enzyme]
   901 PROTEIN KINASE [Amino Acid, Peptide, or Protein,Enzyme]
   827 Kinase (Phosphotransferases) [Amino Acid, Peptide, or Protein,Enzyme]
   827 Protein (Proteins) [Amino Acid, Peptide, or Protein,Biologically Active S
ubstance]
   827 Protein NOS (Protein measurement) [Laboratory Procedure]
   827 CK2 [Laboratory Procedure]
Meta Mapping (1000)
  1000 protein kinase CK2 (casein kinase II) [Amino Acid, Peptide, or Protein,Enzyme]
```

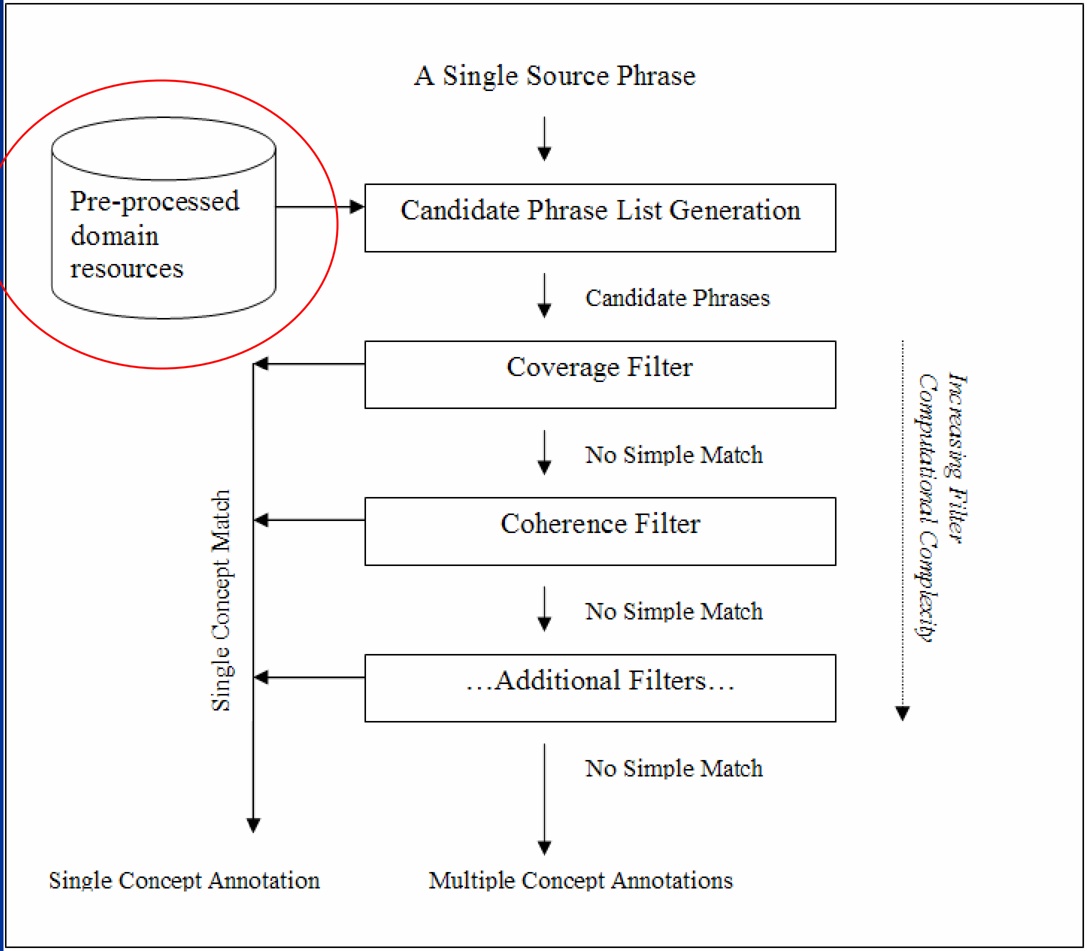
Evaluation Corpus

- Selected 24 papers from a database of ~1,200 randomized controlled trial publications

- Worked with three DUCoM student medical researchers
  - generated 3 extractive summaries at 20% compression to serve as model summaries

Approach:
Biomedical Semantic Annotation

Biomedical Annotator Design


A Single Source Phrase
Pre-processed domain resources
Candidate Phrase List Generation
Candidate Phrases
Coverage Filter
No Simple Match
Coherence Filter
No Simple Match
…Additional Filters…
No Simple Match
Single Concept Match
Increasing Filter Computational Complexity
Single Concept Annotation
Multiple Concept Annotations

Pre-processing

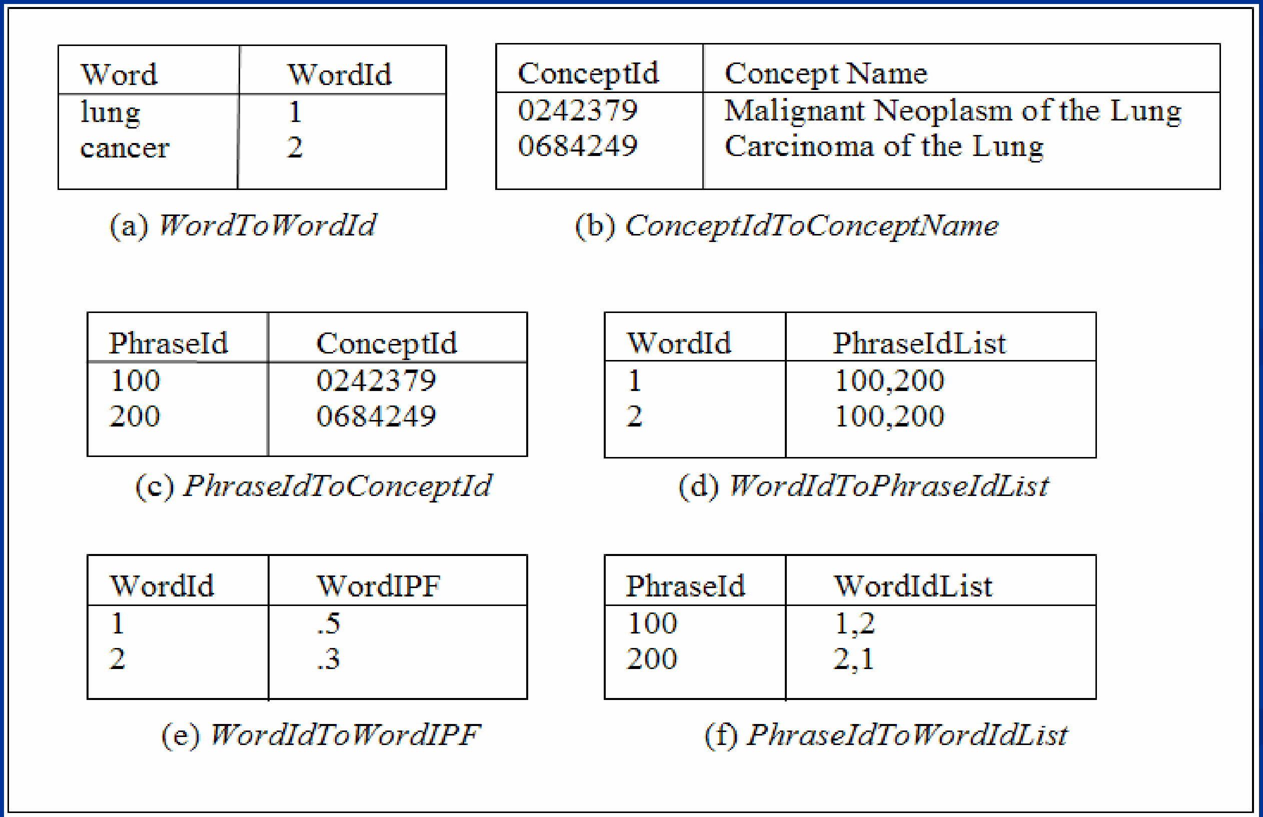
- Purpose:
  - Convert UMLS resources into a more efficient format
  - Pre-calculate phrase values
  - Convert words to base forms: {*eyes*, *eyed*, *eying*} → {*eye*}
- Rationale:
  - Reduce runtime calculations
  - Reduce runtime parsing of phrases, words
  - Produce a more compact format for in-memory usage

Pre-processing

- Phrase types:
  - *Source Phrase*: phrase from source text
  - *Concept Phrase*: phrase from UMLS concept
  - *Candidate Phrase*: concept phrases having words in common with a source phrase

Pre-processing: Table Generation

- Convert UMLS resources into a more efficient format
- Create efficient structures for in-memory lookup



Word	WordId
lung	1
cancer	2

(a) *WordToWordId*

ConceptId	Concept Name
0242379	Malignant Neoplasm of the Lung
0684249	Carcinoma of the Lung

(b) *ConceptIdToConceptName*

PhraseId	ConceptId
100	0242379
200	0684249

(c) *PhraseIdToConceptId*

WordId	PhraseIdList
1	100,200
2	100,200

(d) *WordIdToPhraseIdList*

WordId	WordIPF
1	.5
2	.3

(e) *WordIdToWordIPF*

PhraseId	WordIdList
100	1,2
200	2,1

(f) *PhraseIdToWordIdList*

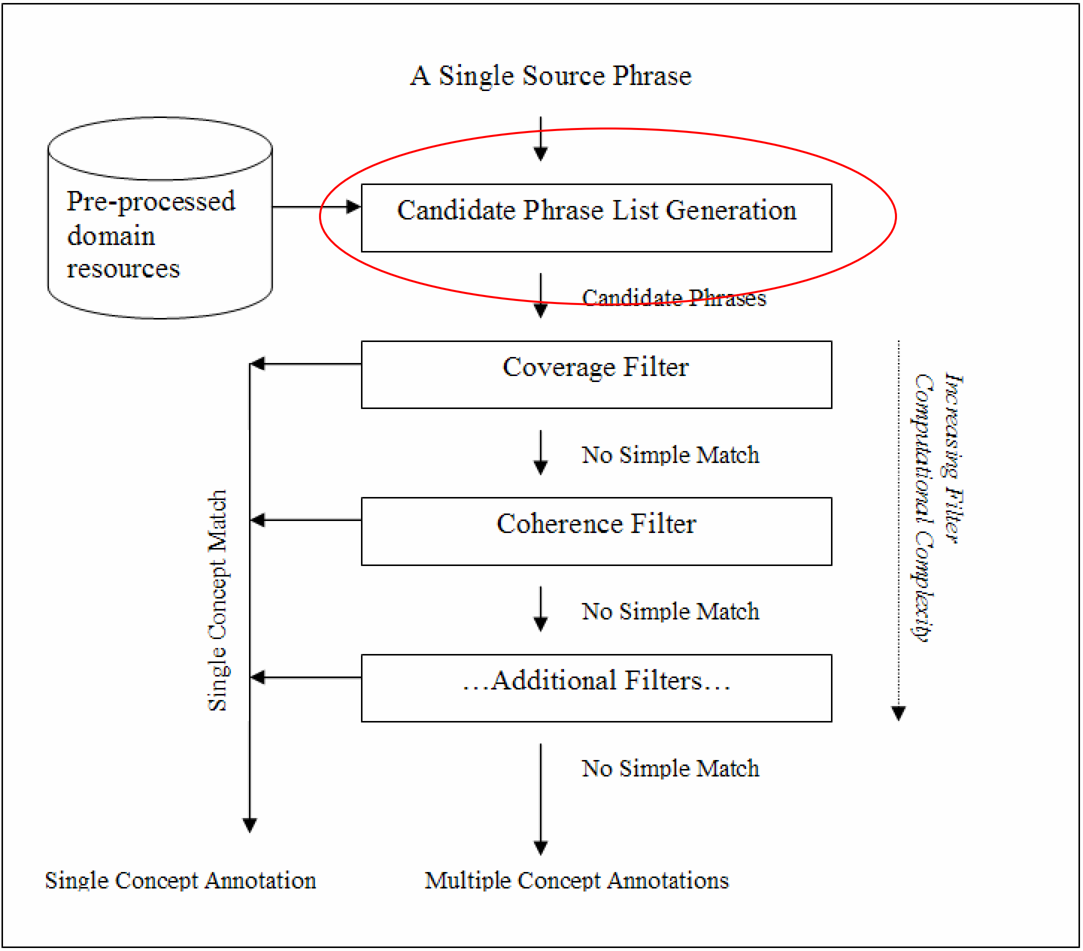
Pre-processing – IPF Calculation

- Calculate the IPF value of each UMLS word
- Gives indication of word importance based on its usage within all UMLS concept phrases

$$\boxed{\textit{inverse phrase frequency} = \log \frac{N}{n_i}}$$

N = total # of phrases in UMLS
n_i = total # phrases word *i* appears in

Candidate Phrase List Generation


A Single Source Phrase
Pre-processed domain resources
Candidate Phrase List Generation
Candidate Phrases
Coverage Filter
No Simple Match
Coherence Filter
No Simple Match
…Additional Filters…
No Simple Match
Single Concept Match
Increasing Filter Computational Complexity
Single Concept Annotation
Multiple Concept Annotations

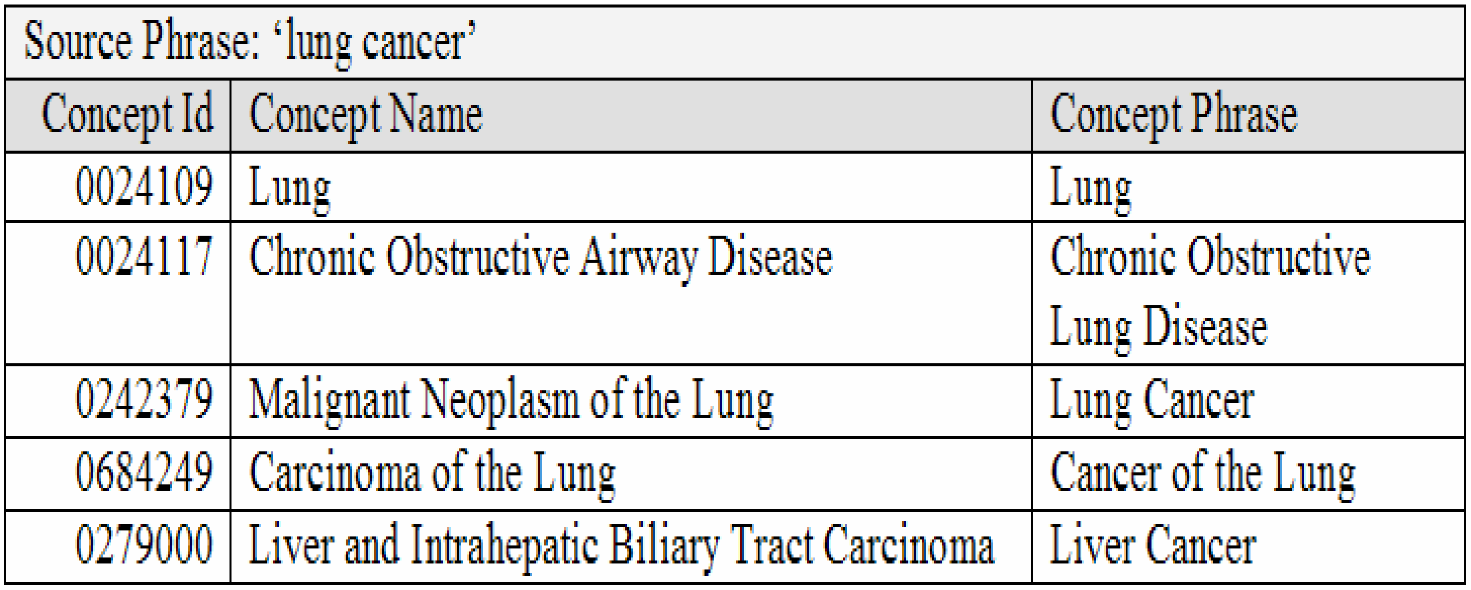
Candidate Phrase List Generation

- Generates a list of concept phrases having at least one word in common with a source phrase
  - It is a pool of possible concept matches
- Goal at each stage is to:
  - find a concept phrase match with a source phrase
  - reduce the size of candidate list passed to the next stage
- Two approaches: short and long source phrase

Candidate Phrase List Generation

(Short Source Phrase < 5 words)

- Find all concept phrases having at least one word in common with the source phrase



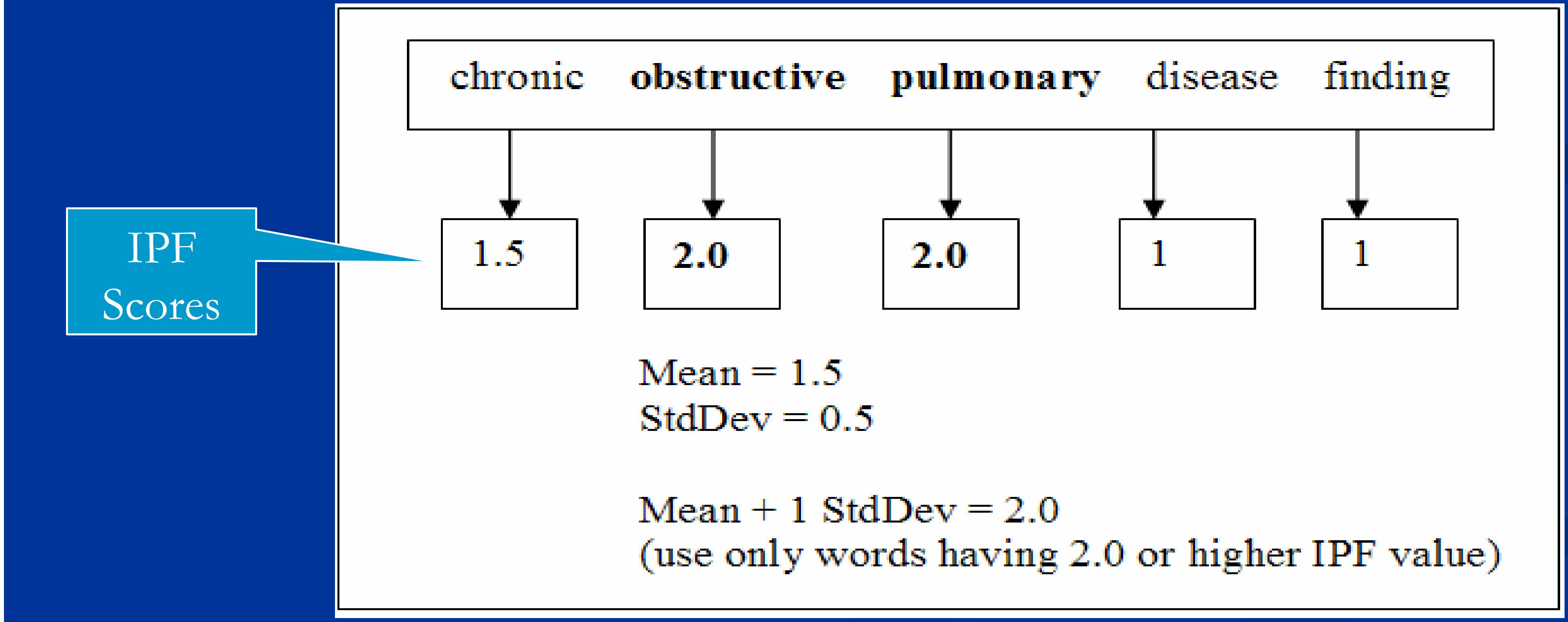
Source Phrase: ‘lung cancer’

Concept Id	Concept Name	Concept Phrase
0024109	Lung	Lung
0024117	Chronic Obstructive Airway Disease	Chronic Obstructive Lung Disease
0242379	Malignant Neoplasm of the Lung	Lung Cancer
0684249	Carcinoma of the Lung	Cancer of the Lung
0279000	Liver and Intrahepatic Biliary Tract Carcinoma	Liver Cancer

Candidate Phrase List Generation

(Long Source Phrase >= 5 words)

- Find all concept phrases having at least one word in common with the source phrase's most important words
  - By using only important words, only the most likely candidate phrases will be placed into the candidate phrase list



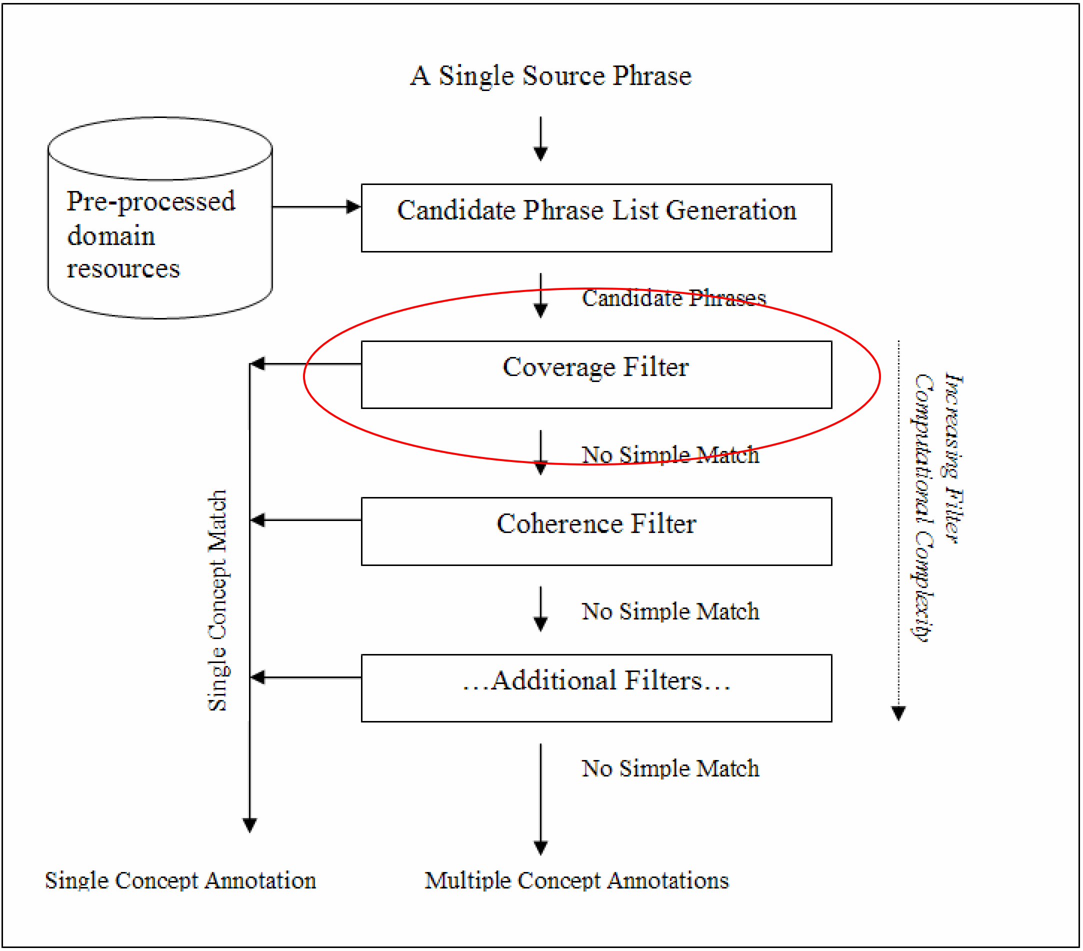
IPF Scores

chronic	**obstructive**	**pulmonary**	disease	finding
1.5	**2.0**	**2.0**	1	1

Mean = 1.5
StdDev = 0.5

Mean + 1 StdDev = 2.0
(use only words having 2.0 or higher IPF value)

Coverage Filter


A Single Source Phrase
Pre-processed domain resources
Candidate Phrase List Generation
Candidate Phrases
Coverage Filter
No Simple Match
Coherence Filter
No Simple Match
…Additional Filters…
No Simple Match
Single Concept Match
Increasing Filter Computational Complexity
Single Concept Annotation
Multiple Concept Annotations

Coverage Filter

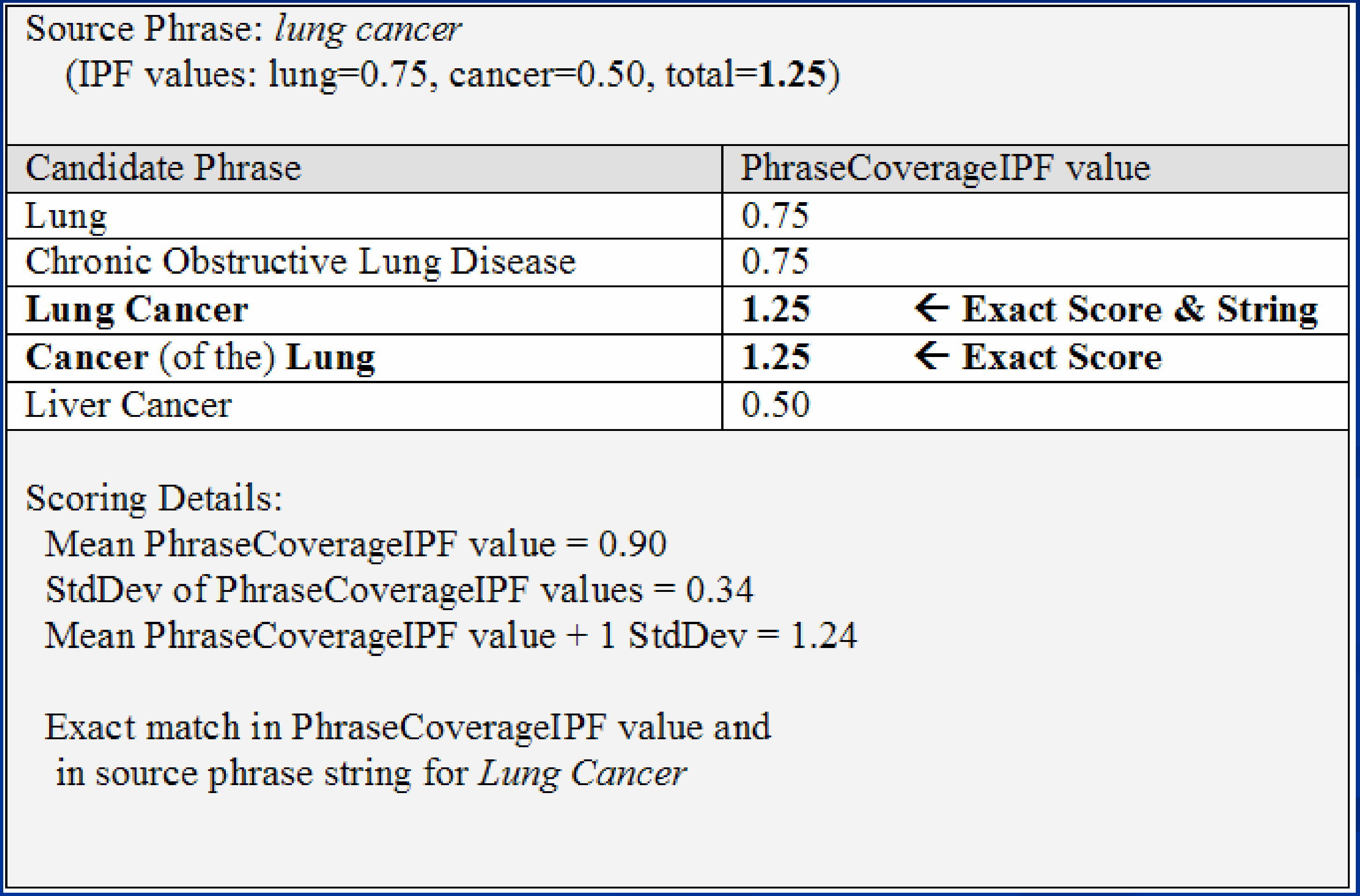
- Find the candidate phrase subset having the best coverage of words with the source phrase
  - Common, important words = better match

$$\boxed{\text{PhraseCoverageIPF} = \sum_{i=1}^{N} \text{IPF}_{i}}$$

IPF = Inverse Phrase Frequency values of word
PhraseCoverageIPF = sum of common word IPF values
N = total # of common words between source phrase and candidate phrase
i = common word instance between source phrase and candidate phrase

Coverage Filter

(Example – Exact Match)



Source Phrase: *lung cancer*

(IPF values: lung=0.75, cancer=0.50, total=**1.25**)

Candidate Phrase	PhraseCoverageIPF value
Lung	0.75
Chronic Obstructive Lung Disease	0.75
Lung Cancer	**1.25** ← **Exact Score & String**
Cancer (of the) **Lung**	**1.25** ← **Exact Score**
Liver Cancer	0.50

Scoring Details:

Mean PhraseCoverageIPF value = 0.90

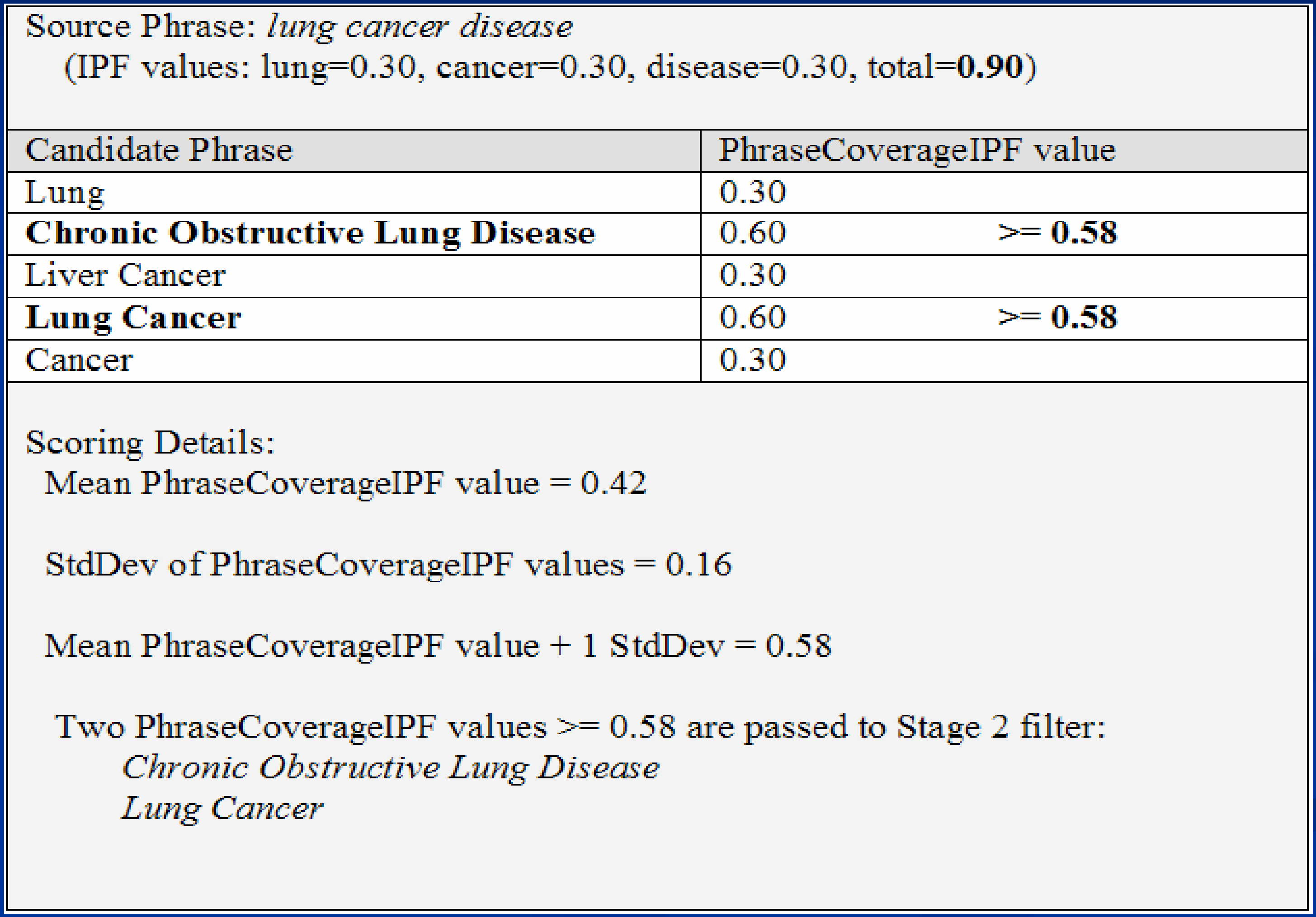
StdDev of PhraseCoverageIPF values = 0.34

Mean PhraseCoverageIPF value + 1 StdDev = 1.24

Exact match in PhraseCoverageIPF value and in source phrase string for *Lung Cancer*

Coverage Filter

(Example – Inexact Match)



Source Phrase: *lung cancer disease*
(IPF values: lung=0.30, cancer=0.30, disease=0.30, total=**0.90**)

Candidate Phrase	PhraseCoverageIPF value	
Lung	0.30	
Chronic Obstructive Lung Disease	0.60	**>= 0.58**
Liver Cancer	0.30	
Lung Cancer	0.60	**>= 0.58**
Cancer	0.30	

Scoring Details:

Mean PhraseCoverageIPF value = 0.42

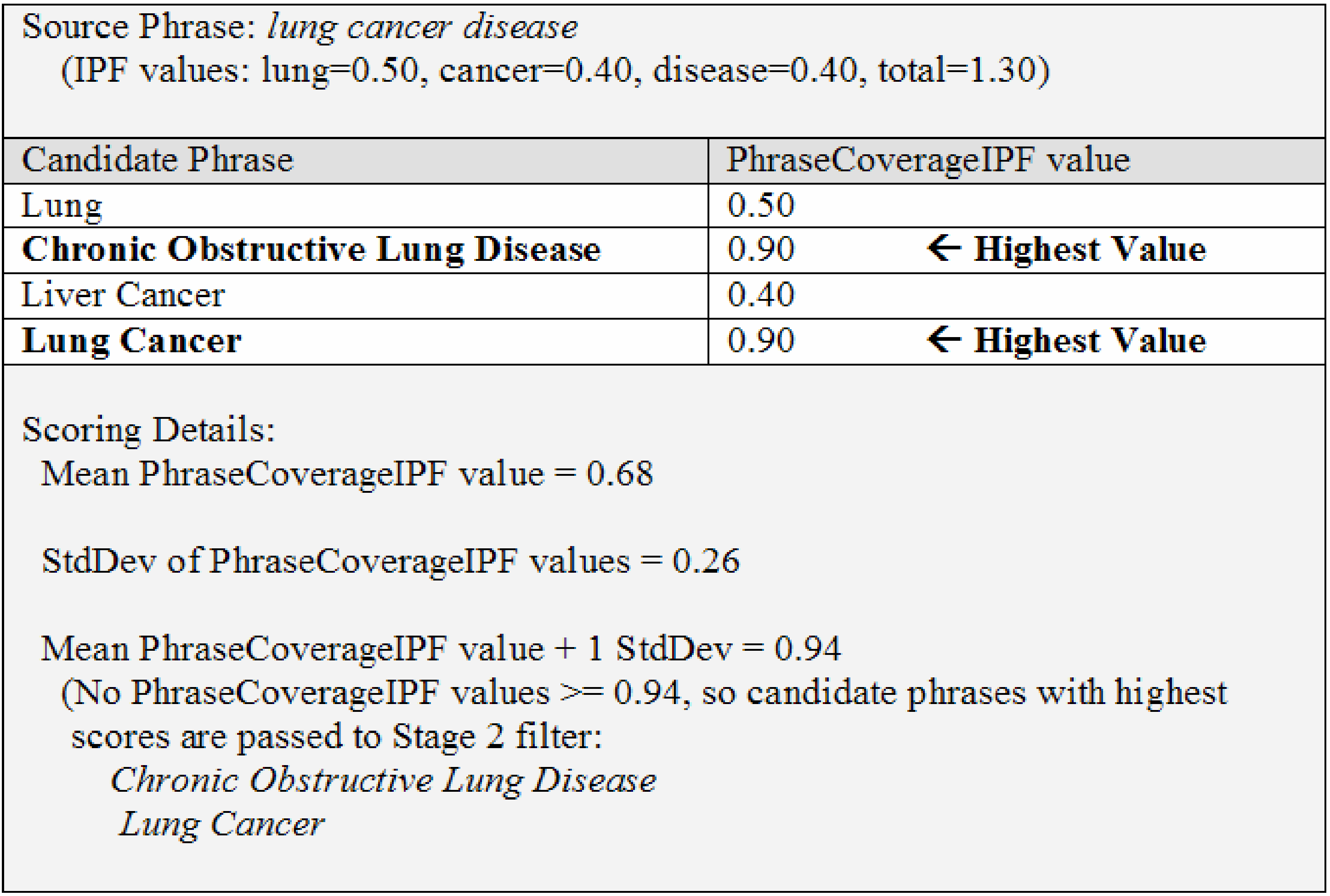
StdDev of PhraseCoverageIPF values = 0.16

Mean PhraseCoverageIPF value + 1 StdDev = 0.58

Two PhraseCoverageIPF values >= 0.58 are passed to Stage 2 filter:
Chronic Obstructive Lung Disease
Lung Cancer

Coverage Filter

(Example – Inexact Match, Highest Values)



Source Phrase: *lung cancer disease*
(IPF values: lung=0.50, cancer=0.40, disease=0.40, total=1.30)

Candidate Phrase	PhraseCoverageIPF value
Lung	0.50
Chronic Obstructive Lung Disease	0.90 ← **Highest Value**
Liver Cancer	0.40
Lung Cancer	0.90 ← **Highest Value**

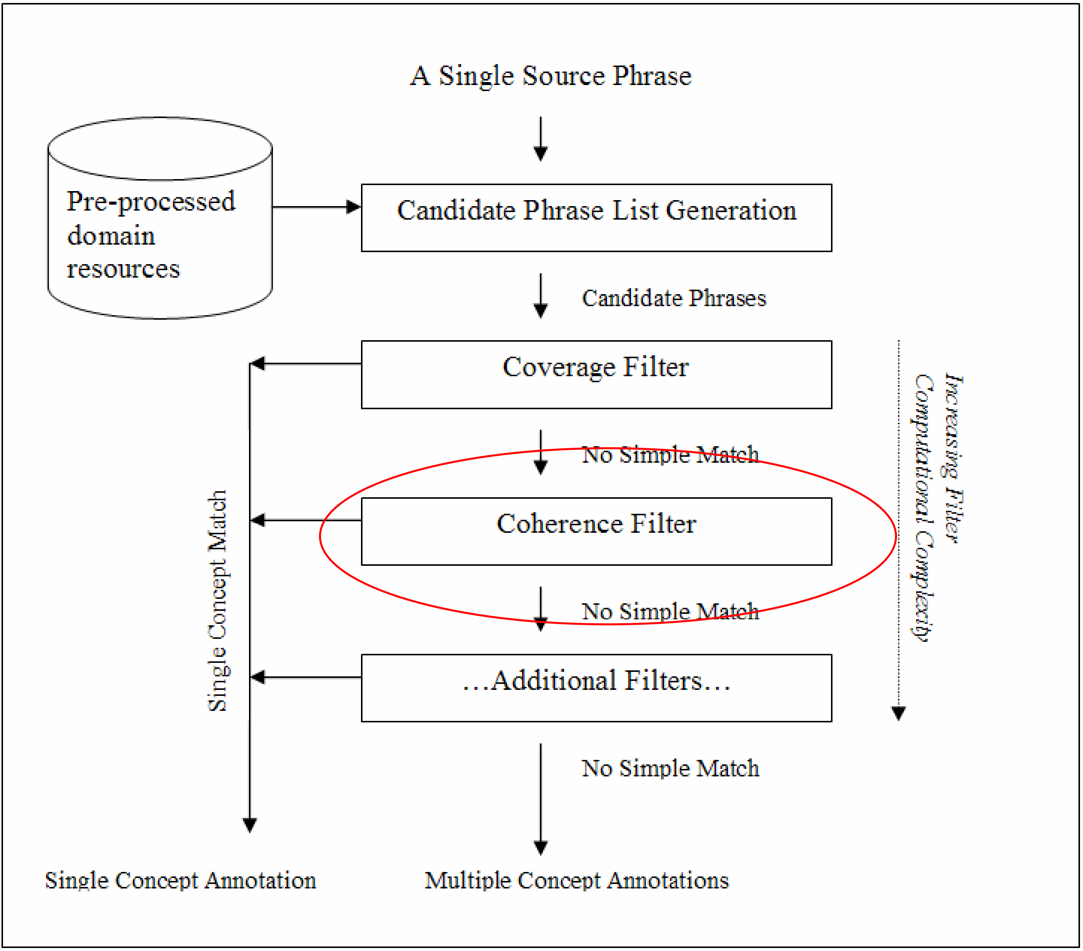
Scoring Details:

Mean PhraseCoverageIPF value = 0.68

StdDev of PhraseCoverageIPF values = 0.26

Mean PhraseCoverageIPF value + 1 StdDev = 0.94
(No PhraseCoverageIPF values >= 0.94, so candidate phrases with highest scores are passed to Stage 2 filter:
Chronic Obstructive Lung Disease
Lung Cancer

Coherence Filter


A Single Source Phrase
Pre-processed domain resources
Candidate Phrase List Generation
Candidate Phrases
Coverage Filter
No Simple Match
Coherence Filter
No Simple Match
…Additional Filters…
No Simple Match
Single Concept Match
Increasing Filter Computational Complexity
Single Concept Annotation
Multiple Concept Annotations

Coherence Filter

- Find the candidate phrase subset having the best word ordering in common with the source phrase
  - Coverage filter – finds concept phrases with common source phrase words
  - Coherence filter - finds concept phrases having common words with a source phrase, in a common order
    - Do the common words make sense when put together?

- Measured using Skip-bigrams
  - Bi-grams which allow for intervening words to be omitted (skipped)
  - Permits inexact string matching, but enforces word order

Coherence Filter

(Skip-bigram Example)

- Can enumerate all skip-bigrams for a phrase, or…
- Specify a maximum gap size to reduce computational complexity

Phrase: ‘peripheral plasma cell myeloma’

peripheral plasma
peripheral cell
peripheral myeloma
plasma cell
plasma myeloma
cell myeloma

Complete Skip-bigram List

peripheral plasma
plasma cell
cell myeloma

Gap-Zero Skip-bigram List

peripheral plasma
peripheral cell
plasma cell
plasma myeloma
cell myeloma

Gap-One Skip-bigram List

Coherence Filter

(Measuring Skip-bigrams)

- Compare source phrase skip-bigrams with candidate phrase skip-bigrams using traditional precision/recall measures

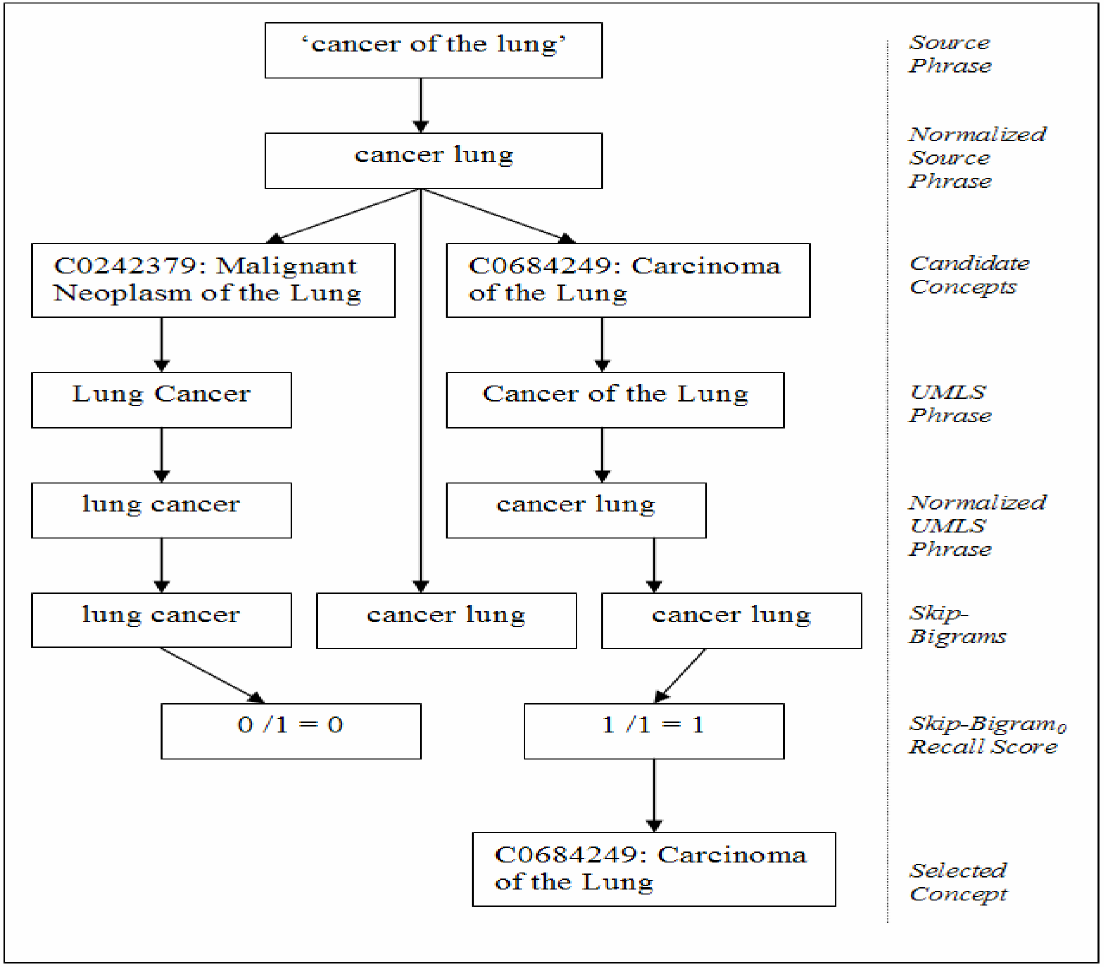
- Recall used in Coherence Filter as candidate phrase score
  - shown in machine translation to correlate well with human evaluations

$$\text{Precision} = \frac{\text{CommonSkipBigramsWithinGap}(\text{SourcePhrase}, \text{CandidatePhrase})}{\text{CountSkipBigramsWithinGap}(\text{SourcePhrase})}$$

$$\text{Recall} = \frac{\text{CommonSkipBigramsWithinGap}(\text{SourcePhrase}, \text{CandidatePhrase})}{\text{CountSkipBigramsWithinGap}(\text{CandidatePhrase})}$$

Coherence Filter

(Example Evaluation)



'cancer of the lung'

Source Phrase

cancer lung

Normalized Source Phrase

C0242379: Malignant Neoplasm of the Lung

C0684249: Carcinoma of the Lung

Candidate Concepts

Lung Cancer

Cancer of the Lung

UMLS Phrase

lung cancer

cancer lung

Normalized UMLS Phrase

lung cancer

cancer lung

cancer lung

Skip-Bigrams

0 /1 = 0

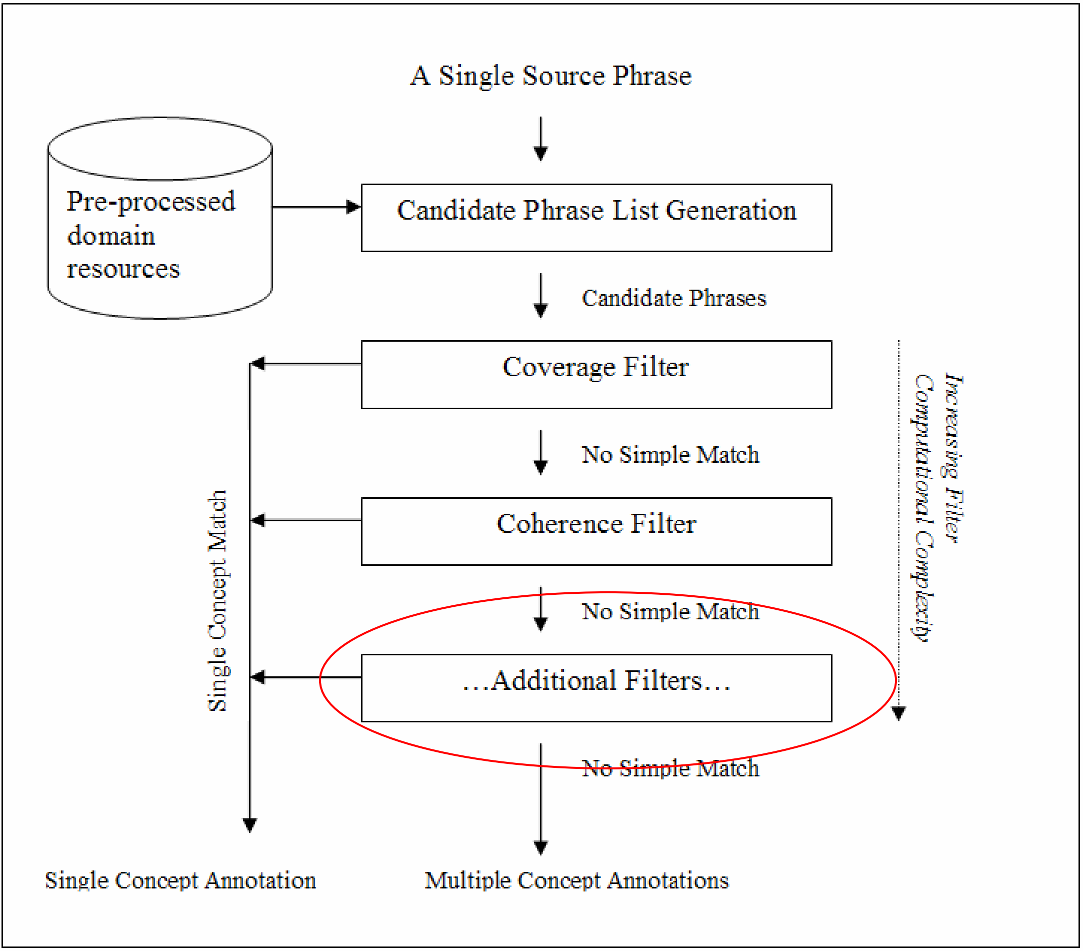
1 /1 = 1

Skip-Bigram$_0$ Recall Score

C0684249: Carcinoma of the Lung

Selected Concept

Possible Additional Filters


A Single Source Phrase
Pre-processed domain resources
Candidate Phrase List Generation
Candidate Phrases
Coverage Filter
No Simple Match
Coherence Filter
No Simple Match
…Additional Filters…
No Simple Match
Single Concept Match
Increasing Filter Computational Complexity
Single Concept Annotation
Multiple Concept Annotations

Possible Additional Filters

- Concept Disambiguation
  - Annotate exact source phrase matches in source text, then iteratively disambiguate other source phrases based on concept co-occurrence probability
- Language Modeling
  - Use all synonymous phrases for a concept and generate a language model for each concept
  - Compare source phrase language model to concept language model

Approach:
Biomedical Text Summarization

Summarization Approaches

- Entity-level
  - Graph-based approaches
  - Concept Chaining (BioChain)
- Surface-level
  - Statistical approaches
  - Concept Frequency Distribution (FreqDist)
- Discourse-level
  - Text structure

Concept Chaining Summarizer

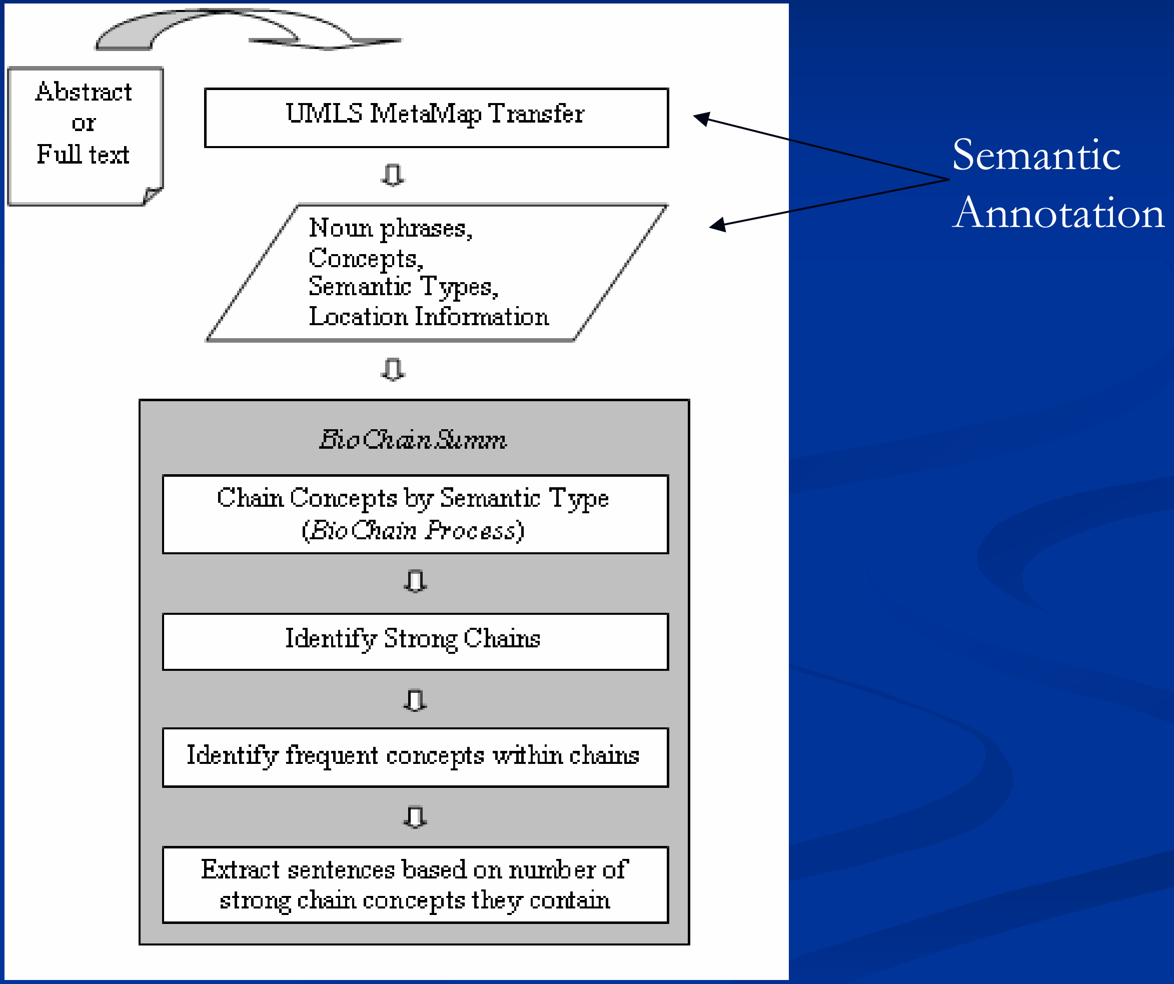
Concept Chaining

- Chains together related concepts
  - Concepts related by UMLS Semantic Network

- Identifies important sentences based on the strongest chains
  - Strength based on concept frequency within chain

- Important sentences are then extracted to form a summary

Summarization Process Using Concept Chaining



Abstract or Full text

UMLS MetaMap Transfer

Noun phrases, Concepts, Semantic Types, Location Information

BioChainSumm

Chain Concepts by Semantic Type (*BioChain Process*)

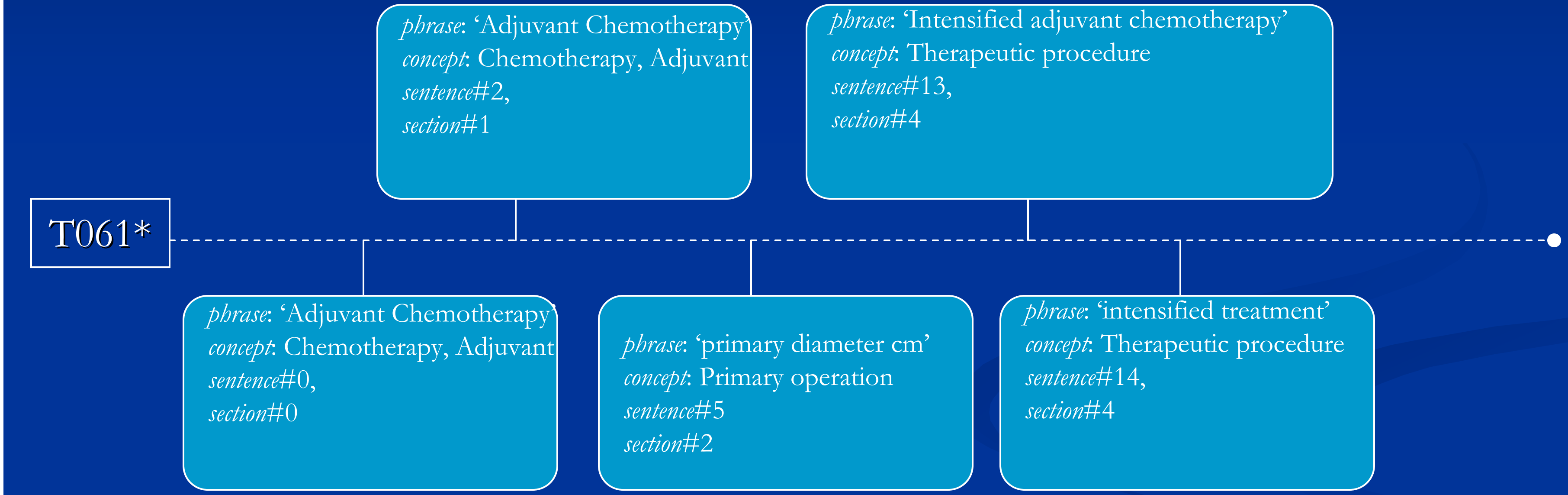
Identify Strong Chains

Identify frequent concepts within chains

Extract sentences based on number of strong chain concepts they contain

Semantic Annotation

Concept Chain - Example



T061*

phrase: ‘Adjuvant Chemotherapy’
concept: Chemotherapy, Adjuvant
sentence#2,
section#1

phrase: ‘Intensified adjuvant chemotherapy’
concept: Therapeutic procedure
sentence#13,
section#4

phrase: ‘Adjuvant Chemotherapy’
concept: Chemotherapy, Adjuvant
sentence#0,
section#0

phrase: ‘primary diameter cm’
concept: Primary operation
sentence#5
section#2

phrase: ‘intensified treatment’
concept: Therapeutic procedure
sentence#14,
section#4

*T061 = Therapeutic or Preventive Procedure (UMLS Semantic Type)

Concept Chains - Characteristics

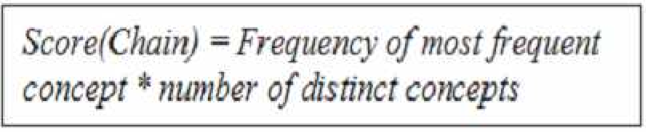
- Concept Chains:
  - Chains with at least one concept:
    - Average: 60, Minimum: 49, Maximum: 73
  - Chains with no concepts:
    - Average: 74, Minimum: 62, Maximum: 86
- Concepts with chains
  - All Concepts in chain:
    - Average: 11, Mininum: 1, Maximum: 56
  - Distinct Concepts in chain:
    - Average: 5, Mininum: 1, Maximum: 46

Concept Chain Scoring

- Want to find chains which contain concepts most discussed in the source text
- Score each chain to give its overall importance in describing the source text
- Lexical chaining research identified three factors for chain strength:
  - Reiteration: more repetion is better
  - Density: shorter distance between concepts is better
  - Length: longer chain length is better

Concept Chain Scoring

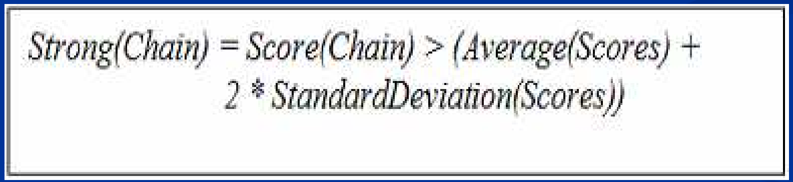
- Chain score:



*Score(Chain) = Frequency of most frequent concept * number of distinct concepts*

  - Uses reiteration and length

- Find the strongest chain(s) based on each chain's score:



*Strong(Chain) = Score(Chain) > (Average(Scores) + 2 * StandardDeviation(Scores))*

  - The strongest chain(s) indicate which chain's have concepts most representative of the source text

Strong Chains – Example

- Chains with score > 0.0:
  - T081-Quantitative Concept, score: 14.0
  - T061-Therapeutic or Preventive Procedure, score: 6.0
  - T169-Functional Concept, score: 6.0
  - T079-Temporal Concept, score: 4.0
  - T080-Qualitative Concept, score: 4.0
  - T082-Spatial Concept, score: 4.0
  - T073-Manufactured Object, score: 2.0
  - T109-Organic Chemical, score: 2.0
  - T170-Intellectual Product, score: 2.0
  - T121-Pharmacologic Substance, score: 1.0

Strong chains: (*2 StdDev*)

Avg score: 1.667 (includes all zero score chains)

Std Dev: 3.067

Strong Score: 7.801

T081-Quantitative Concept: 14.0

Strong chains: (*1 StdDev*)

Avg score: 1.667 (includes all zero score chains)

Std Dev: 3.06714

Strong Score: 4.734

T081-Quantitative Concept: 14.0

T061-Therapeutic or Preventive Procedure: 6.0

T169-Functional Concept: 6.0

Sentence Scoring

- Need to find sentences which summarize source text the best
- Top chains identify what is discussed most
- Perform frequency count on concepts within top chain(s)
  - concept(s) with highest frequency is top concept for each chain
- Score each sentence based on number of top concepts it contains

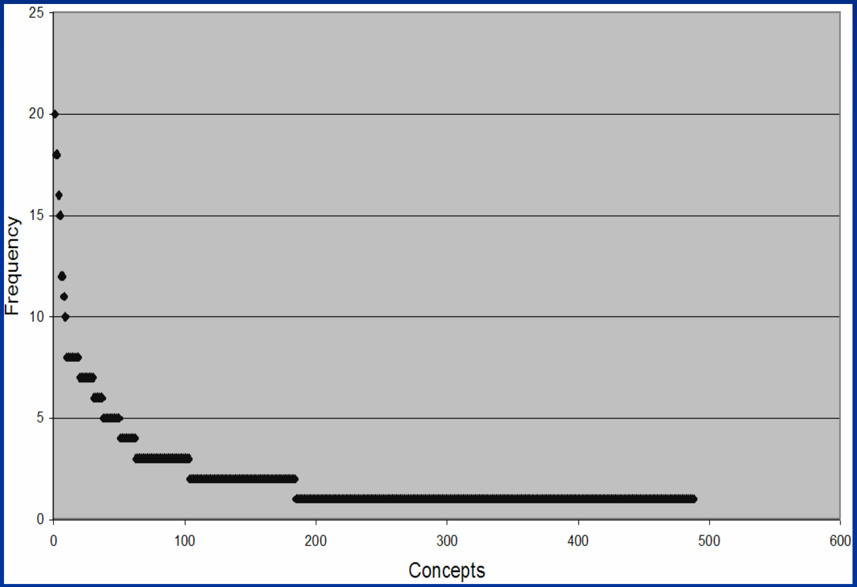
Frequency Distribution Summarizer

Concept Frequency Distribution

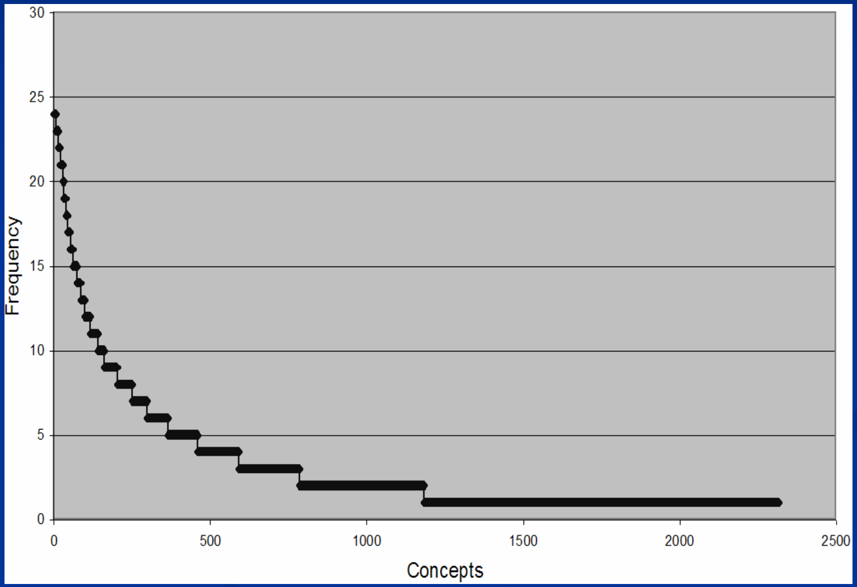
- Match distribution of concepts in source text to summary
  - Summary concept distribution should be close to concept distribution in source text
- Main problem: how to determine concept distribution similarity between the two

Concept Frequency Distribution

(Concepts sorted by descending frequency)



Distribution of 488 discovered biomedical concepts within the paper abstracts



Distribution of 2,317 discovered biomedical concepts within the full-text of the papers

Frequency Distribution Summarization Algorithm (FreqDist)

- 1. Build frequency model of full-text

- 2. Iteratively select sentences from full-text:
  - Take a sentence, add it to current summary
    - This is a candidate summary
  - Compare candidate summary distribution model for similarity to full-text distribution model
  - Select sentence which, when added to summary, best models the full-text

- 3. Repeat #2 until desired summary size reached

Concept Distribution Similarity

- Problem: how to determine similarity between summary and full-text?

- Modeled summary and full-text as vector of concepts

- Evaluated five similarity functions for the vectors:
  - *Cosine*: Calculate angle between the two vectors
  - *Dice*: Consider vector membership commonality
  - *Euclidean* Distance: Sum of squared distances
  - *Unit Item Frequency*: Fast simulation of cosine
  - *Vector Subtraction*: Subtract the two vectors

Evaluation:

Biomedical Semantic Annotation

Semantic Annotation Evaluation

- Two types:
  - Intrinsic:
    - Compare concept annotation output to a gold standard
      - i.e., concepts mapped by annotator to concepts mapped by MetaMap
    - Measures how well proposed annotator is to known annotator
  - Extrinsic:
    - Use the concept annotation output in a task (i.e., summarization)
    - Evaluate the task to see if there is improvement
    - Can determine if annotation has improved beyond the gold standard

Semantic Annotation Evaluation

- Intrinsic
  - Data Set:
    - Generate unique noun phrases from evaluation corpus
    - Use MetaMap to generate concept(s) for each phrase
  - Evaluate vs. MetaMap:
    - Speed: annotation time per phrase
    - Accuracy:

$$\text{Precision} = \frac{\#\text{ of correct concepts}}{\text{total }\#\text{ of concepts mapped}}$$

$$\text{Recall} = \frac{\#\text{ of correct concepts}}{\text{total }\#\text{ of MetaMap concepts}}$$

Semantic Annotation Evaluation

- Extrinsic
  - Use summarization to evaluate annotation output
    - FreqDist and SummBasic
    - Assumption: If summarization performance using concepts improves, improvement is due to better identification of concepts

Evaluation:
Biomedical Text Summarization

Summarization Evaluation

- Basic idea:
  - Generate ideal summaries from domain experts, use these as model summaries
  - Generate system summaries and compare to model summaries using ROUGE tool
  - Also use external summarizers to put our work in perspective
- Resources:
  - A corpus of 24 randomly selected biomedical texts
  - Three domain experts generated extractive *model* summaries for each paper @20%
  - Eight external system summarizers also generated extractive summaries @20%

Evaluation

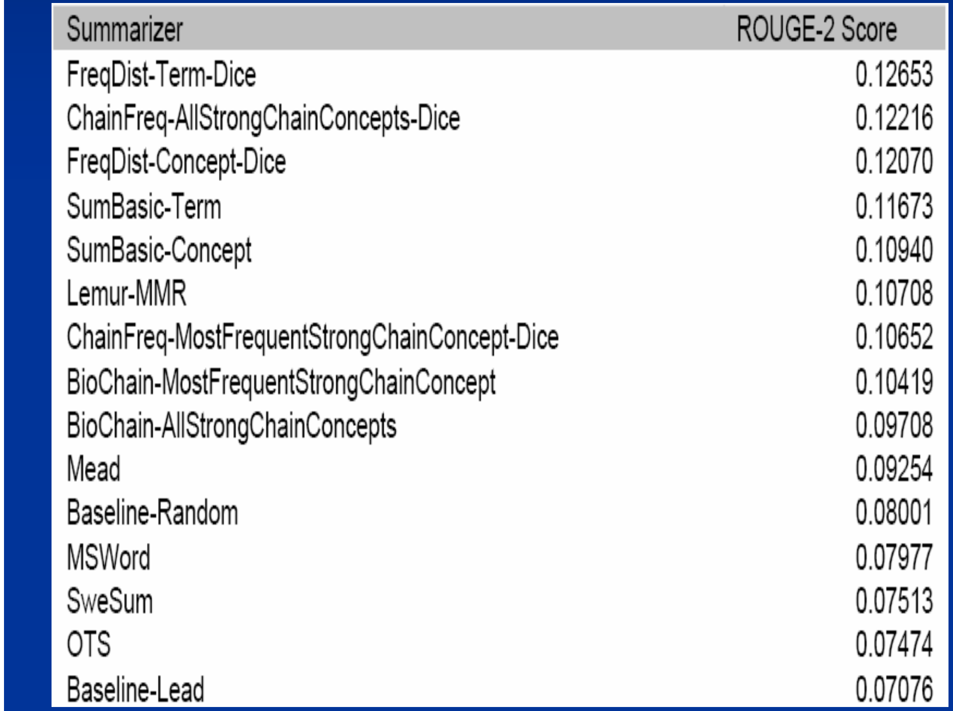
- Used ROUGE (Recall Oriented Understudy for Gisting Evaluation)
  - Compares system summaries to model summaries
  - Results based on n-gram overlap
- ROUGE-2
  - Bigram co-occurrence
- ROUGE-SU4
  - Skip bigram with distance of 4
    - (bigrams with no more than 4 intervening words)
- Same metrics as used in DUC-2005

Evaluation - Summarizers

- Baseline
  - LEAD: first 20% of sentences in text
  - RANDOM: randomly select 20% of sentences in text
- Frequency-based:
  - AutoSummarize (part of Microsoft Word)
  - Open Text Summarizer (OTS)
- Multiple Feature:
  - MEAD: (uses sentence position, sentence length, clustering)
  - SWESUM: (uses sentence position, higher weights for number values)
- Reduce Information Redundancy
  - Lemur MMR
  - SumBasic (frequency-based; probabilistic; state of the art)

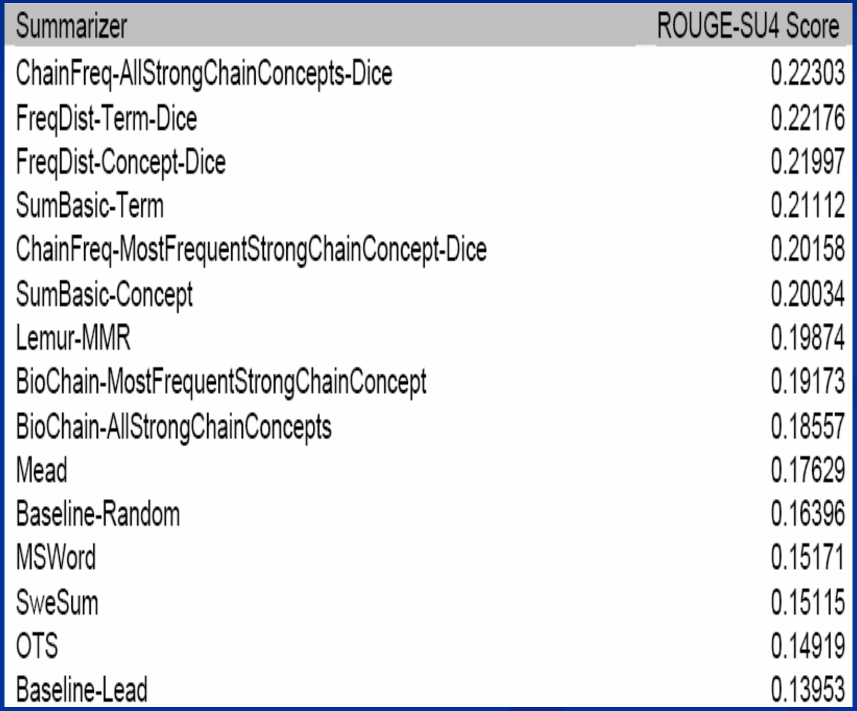
ROUGE Scores for Each Summarizer

ROUGE-2 Scores



Summarizer	ROUGE-2 Score
FreqDist-Term-Dice	0.12653
ChainFreq-AllStrongChainConcepts-Dice	0.12216
FreqDist-Concept-Dice	0.12070
SumBasic-Term	0.11673
SumBasic-Concept	0.10940
Lemur-MMR	0.10708
ChainFreq-MostFrequentStrongChainConcept-Dice	0.10652
BioChain-MostFrequentStrongChainConcept	0.10419
BioChain-AllStrongChainConcepts	0.09708
Mead	0.09254
Baseline-Random	0.08001
MSWord	0.07977
SweSum	0.07513
OTS	0.07474
Baseline-Lead	0.07076

ROUGE-SU4 Scores



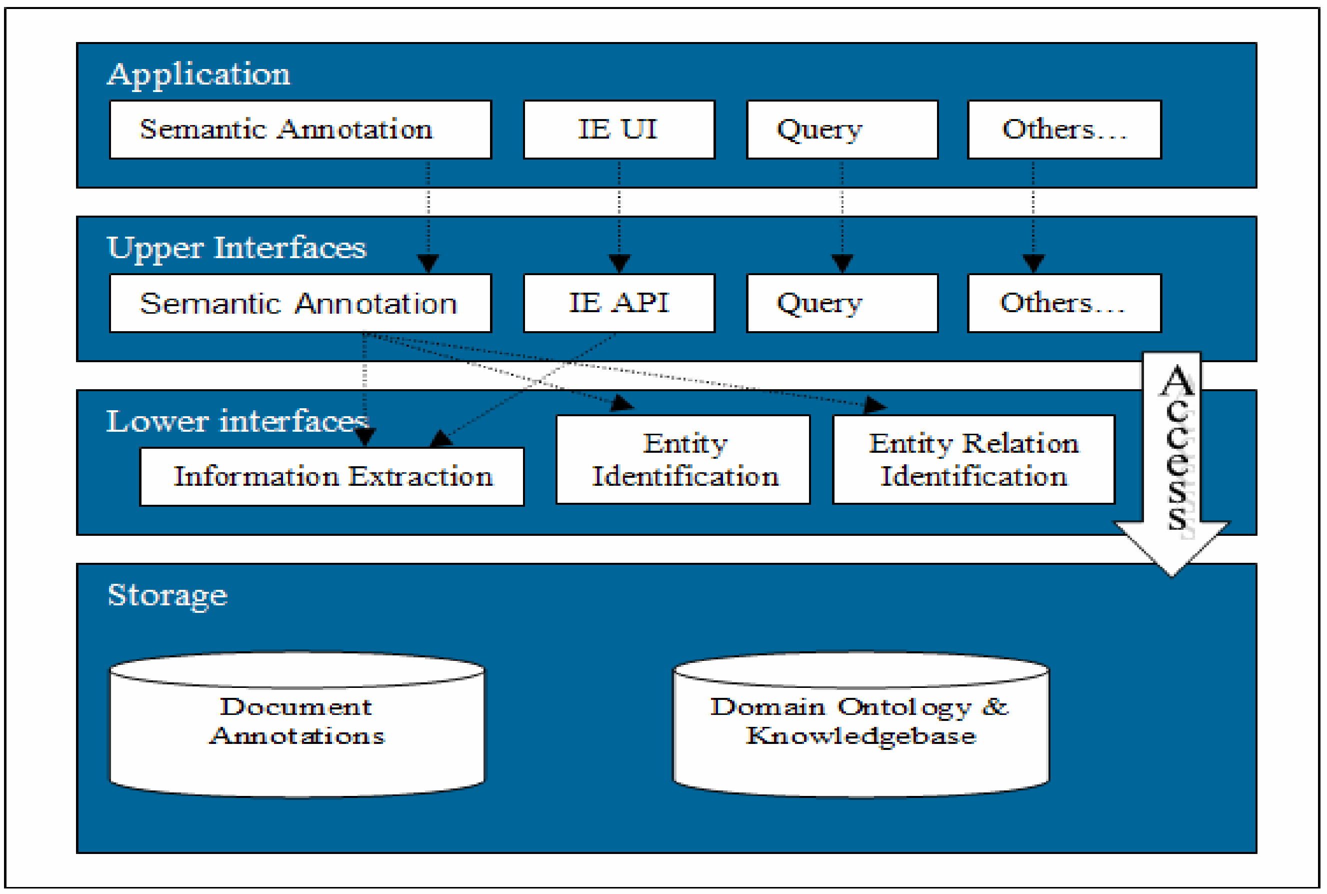
Summarizer	ROUGE-SU4 Score
ChainFreq-AllStrongChainConcepts-Dice	0.22303
FreqDist-Term-Dice	0.22176
FreqDist-Concept-Dice	0.21997
SumBasic-Term	0.21112
ChainFreq-MostFrequentStrongChainConcept-Dice	0.20158
SumBasic-Concept	0.20034
Lemur-MMR	0.19874
BioChain-MostFrequentStrongChainConcept	0.19173
BioChain-AllStrongChainConcepts	0.18557
Mead	0.17629
Baseline-Random	0.16396
MSWord	0.15171
SweSum	0.15115
OTS	0.14919
Baseline-Lead	0.13953

Observations

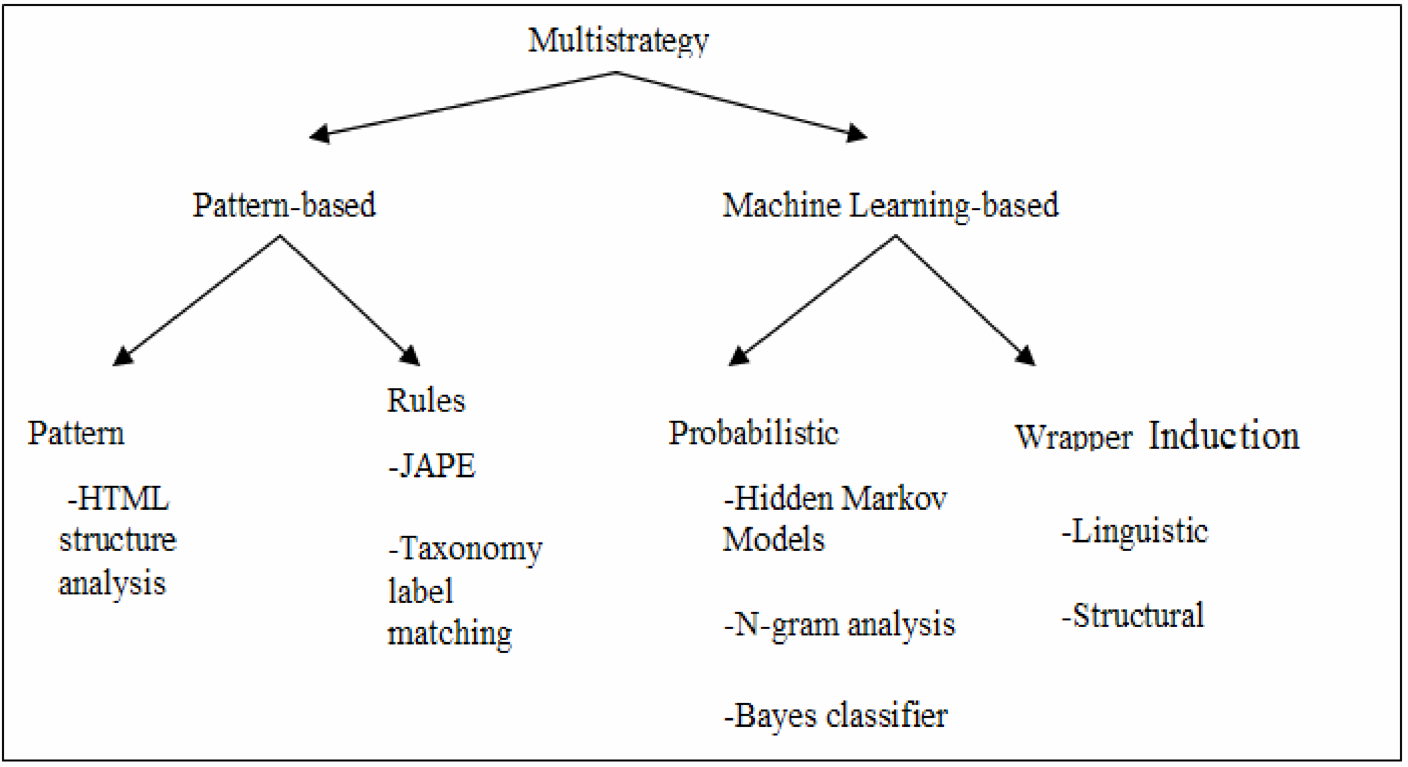
- Random summarizer performs relatively well!
- LEAD is worst performing
  - Unlike news genre, where LEAD is competitive
- Use of DICE (membership) for FreqDist works best
- Use of terms and concepts perform closely
  - But concepts can allow for easier tailoring of summary

Literature Review:
Semantic Annotation

Semantic Annotation Architecture



Semantic Annotation Platform Classification



Multistrategy

Pattern-based

Machine Learning-based

Pattern

-HTML structure analysis

Rules

-JAPE

-Taxonomy label matching

Probabilistic

-Hidden Markov Models

-N-gram analysis

-Bayes classifier

Wrapper Induction

-Linguistic

-Structural

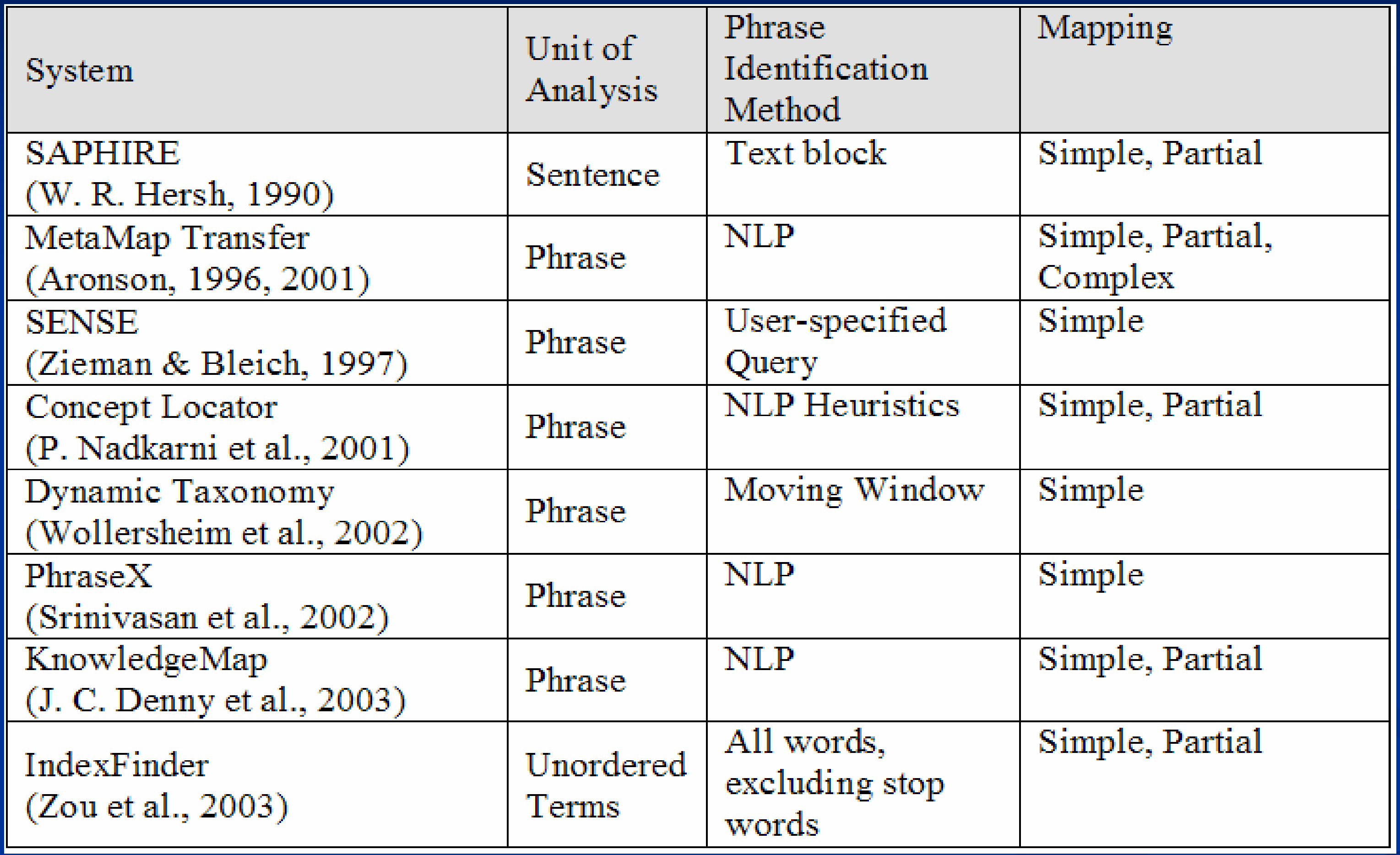
Biomedical Annotation Process

- Construct a unit of analysis by generating subsets of words in the source text (e.g., phrase, sentence)
- (Optional) Normalize the source text unit using UMLS {{387 United States National Library of Medicine 2006;}} by (a) removing possessives, (b) replacing punctuation with spaces, (c) removing stop words, (d) convert words to lower-case, (e) breaking a string into constituent words, and (f) sorting words into alphabetical order
- For each word in the input phrase, build a set of all concepts containing the word;
- Find the intersection of the concept sets;
- (Optional) Find the best matching concept based on the common word membership between the source text and concept text

Source & Concept Match Types

- *None*: there is no match of terms;
- *Simple*: there is an exact match between the terms in the text of the source and the UMLS concept text;
- *Partial*: one or more terms do not match exactly;
- *Complex*: two or more sets of terms map to distinct concepts.

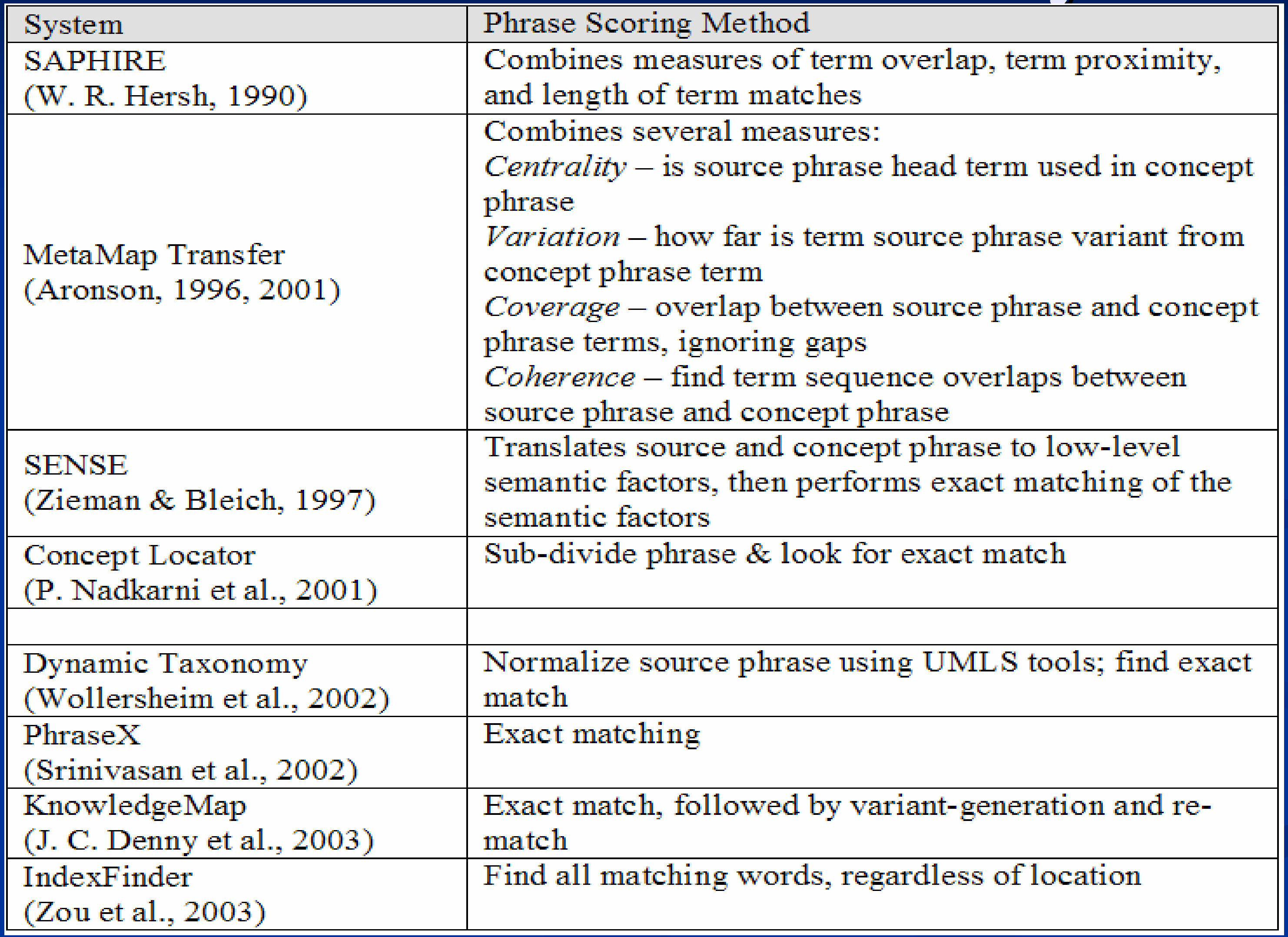
Biomedical Annotation Systems



System	Unit of Analysis	Phrase Identification Method	Mapping
SAPHIRE (W. R. Hersh, 1990)	Sentence	Text block	Simple, Partial
MetaMap Transfer (Aronson, 1996, 2001)	Phrase	NLP	Simple, Partial, Complex
SENSE (Zieman & Bleich, 1997)	Phrase	User-specified Query	Simple
Concept Locator (P. Nadkarni et al., 2001)	Phrase	NLP Heuristics	Simple, Partial
Dynamic Taxonomy (Wollersheim et al., 2002)	Phrase	Moving Window	Simple
PhraseX (Srinivasan et al., 2002)	Phrase	NLP	Simple
KnowledgeMap (J. C. Denny et al., 2003)	Phrase	NLP	Simple, Partial
IndexFinder (Zou et al., 2003)	Unordered Terms	All words, excluding stop words	Simple, Partial

None: no match of terms; *Simple*: exact match; *Partial*: one or more terms do not match exactly; *Complex*: two or more sets of terms map to distinct concepts.

Biomedical Annotation Systems



System	Phrase Scoring Method
SAPHIRE (W. R. Hersh, 1990)	Combines measures of term overlap, term proximity, and length of term matches
MetaMap Transfer (Aronson, 1996, 2001)	Combines several measures: *Centrality* – is source phrase head term used in concept phrase *Variation* – how far is term source phrase variant from concept phrase term *Coverage* – overlap between source phrase and concept phrase terms, ignoring gaps *Coherence* – find term sequence overlaps between source phrase and concept phrase
SENSE (Zieman & Bleich, 1997)	Translates source and concept phrase to low-level semantic factors, then performs exact matching of the semantic factors
Concept Locator (P. Nadkarni et al., 2001)	Sub-divide phrase & look for exact match
Dynamic Taxonomy (Wollersheim et al., 2002)	Normalize source phrase using UMLS tools; find exact match
PhraseX (Srinivasan et al., 2002)	Exact matching
KnowledgeMap (J. C. Denny et al., 2003)	Exact match, followed by variant-generation and re-match
IndexFinder (Zou et al., 2003)	Find all matching words, regardless of location

Proposed Research

- Classified as probabilistic (ML-based)
- Support Simple and Partial matches
  - Support for complex to follow (additional stage)
- Phrase is unit of analysis
- Can use any method for finding phrases
- Follows general annotation process
  - Except uses a pipeline approach

Literature Review: Text Summarization

Summarization Model

- Three phrase model
  - Interpretation
    - Text analysis
  - Transformation
    - Content Selection, Concept Generalization
  - Generation
    - Generates output from Transformation intermediate form

Summarization Methods

- Surface-level
  - Statistical analysis, cue phrases, term locations
- Entity-level
  - Graph-based (e.g., thesaural relations, lexical chaining)
- Discourse-level
  - Text structure
- Hybrid
  - e.g., surface-level + discourse-level

Proposed Research

- Surface-level
  - Concept-frequency
- Entity-level
  - Concept-chaining
- Discourse-level
  - Section weighting
    - (e.g., Introduction, Methodology, Discussion, Conclusion)
- Hybrid
  - Combine concept-frequency with section weighting

Preliminary Work

Publications

- 2 Journal Papers (refereed)
- 2 Book Chapters (refereed)
- 4 Conference Papers (refereed)
- 1 Bulletin Paper

Publications

2007:

- Lawrence Reeve, Hyoil Han and Ari D. Brooks (2007). *Biomedical Text Summarization Using Concept Chains*, International Journal of Data Mining and Bioinformatics. *Refereed. Accepted.*

- Lawrence Reeve, Hyoil Han and Ari D. Brooks. *The Use of Domain-Specific Concepts in Biomedical Text Summarization,* Journal of Information Processing and Management. *Refereed. Accepted.*

2006:

- Lawrence H. Reeve, Hyoil Han, Saya V. Nagori, Jonathan C. Yang, Tamara A. Schwimmer, and Ari D. Brooks (2006). *Concept Frequency Distribution in Biomedical Text Summarization.* Proceedings of the 15th Conference on Information and Knowledge Management. *Refereed* - 15% acceptance.

- Lawrence Reeve, Hyoil Han, and Ari D. Brooks (2006). *BioChain: Using Lexical Chaining Methods for Biomedical Text Summarization.* Proceedings of the 21st Annual ACM Symposium on Applied Computing, Bioinformatics track. *Refereed* - 32% acceptance.

- Lawrence Reeve and Hyoil Han (2006). *A Comparison of Semantic Annotation Systems for Text-based Web Documents.* Web Semantics and Ontology, David Taniar and J. Wenny Rahayu (Eds.), Idea Group Publishing. *Refereed.*

Publications, continued

2005:

- Lawrence Reeve and Hyoil Han (2005). *Survey of Semantic Annotation Platforms*. Proceedings of the 20th Annual ACM Symposium on Applied Computing, Web Technologies and Applications track. Presentation. *Refereed* - 37% acceptance.

- Lawrence Reeve and Hyoil Han (2005). *Semantic Annotation for Semantic Social Networks Using Community Resources*. AIS SIGSEMIS Bulletin, Vol. 2, Issue (3&4), pp: 52-56.

- Lawrence Reeve, Hyoil Han, and Chaomei Chen (2005). *Information Visualization and the Semantic Web*. Visualizing the Semantic Web, Vladimir Geroimenko and Chaomei Chen (Eds.), Springer. *Refereed.*

2004:

- Lawrence Reeve (2004). *Adapting the TileBar Interface for Visualizing Resource Usage.* Proceedings of the 30th International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems. Presentation. *Refereed.*

Research Plan

- Fall 2006
  - Defend dissertation proposal
  - Implement semantic annotator for biomedical text
- Winter 2007
  - Evaluate semantic annotator
  - Implement additional summarizers for summarization evaluation
  - Participate in DUC 2007
- Spring 2007
  - Continue work on semantic annotation and evaluation
  - Determine characteristics of biomedical text, optimal summary size, sections drawn from
  - Participate in DUC 2007
  - Begin writing final dissertation
- Summer 2007
  - Based on sections drawn from, modify concept chaining and frequency distribution summarizers to modify scoring and evaluate
  - Continue final dissertation writing
  - Defend final dissertation

References

- Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. *Proceedings of the AMIA Symposium 2001,* 17-21.
- Aronson, A. R. (1996). *MetaMap: Mapping text to the UMLS metathesaurus.* Unpublished manuscript.
- Denny, J. C., Smithers, J. D., Miller, R. A., & Spickard, A. (2003). "Understanding" medical school curriculum content using KnowledgeMap. *Journal of the American Medical Informatics Association, 10*(4), 351-362.
- Denny, J. C., Irani, P. R., Wehbe, F. H., Smithers, J. D., & Spickard, A.,3rd. (2003). The KnowledgeMap project: Development of a concept-based medical school curriculum database. *Proceedings of the Annual AMIA Symposium,* , 195-199.
- Hersh, W. R., & Greenes, R. A. (1990). SAPHIRE--an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Computers and Biomedical Research, an International Journal, 23*(5), 410-425.
- Nadkarni, P. M. (1997). Concept locator: A client-server application for retrieval of UMLS metathesaurus concepts through complex boolean query. *Computers and Biomedical Research, an International Journal, 30*(4), 323-336.
- Srinivasan, S., Rindflesch, T. C., Hole, W. T., Aronson, A. R., & Mork, J. G. (2002). Finding UMLS metathesaurus concepts in MEDLINE. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium,* , 727-731.
- Wollersheim, D., Rahayu, W., & Reeve, J. (2002). Evaluation of index term discovery in medical reference text. Paper presented at the *Proceedings of the International Conference on Information Technology and Applications,* Bathhurst, NSW, Australia.
- Zieman, Y. L., & Bleich, H. L. (1997). Conceptual mapping of user's queries to medical subject headings. *Proceedings : A Conference of the American Medical Informatics Association / AMIA Annual Fall Symposium. AMIA Fall Symposium,* , 519-522.
- Zou, Q., Chu, W. W., Morioka, C., Leazer, G. H., & Kangarloo, H. (2003). IndexFinder: A method of extracting key concepts from clinical texts for indexing. *Proceedings of the AMIA Annual Symposium,* 763-767.