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

Lawrence H. Reeve

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

1

Overview

- Hypothesis / Approach / Contribution
- Background
    - Unified Medical Language System (UMLS)
- Literature Review
- Approach
- Evaluation
- 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.
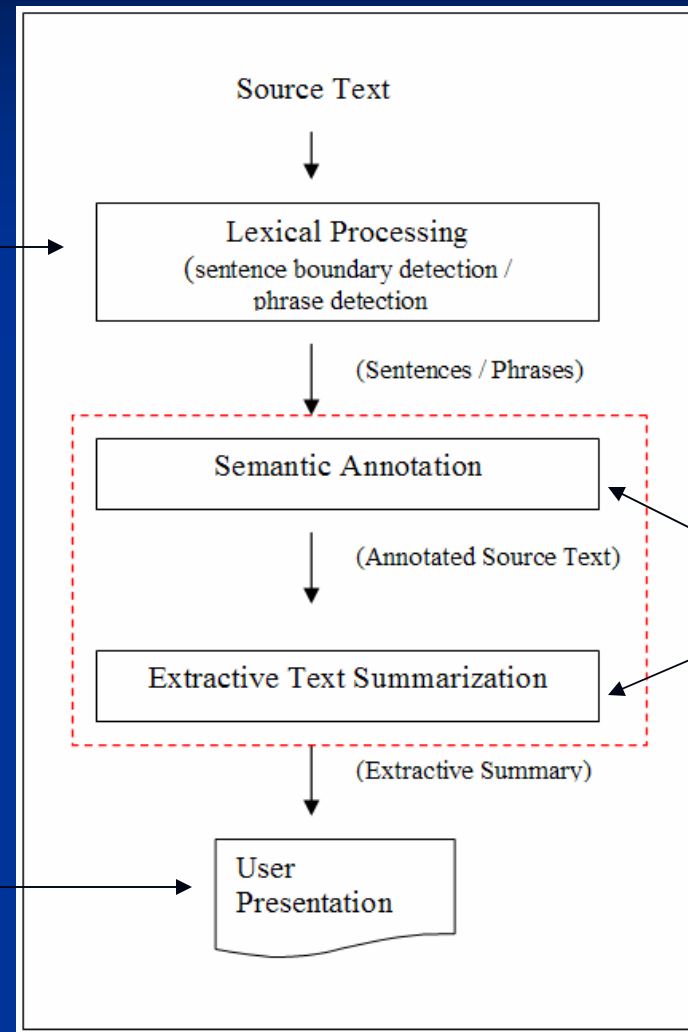
Research Questions

- **Biomedical Semantic Annotation**
  - Can biomedical semantic annotation time be improved for use in an online environment?

- **Biomedical Text Summarization:**
  - 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?

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
proposed research

Text File

Expected Contributions

- Biomedical Semantic Annotation:
  - 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
  - Competitive annotation time over existing approaches

- Biomedical Text Summarization
  - 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 of summarization

Expected Significance

- **Biomedical Semantic Annotation:**
  - Improved time to perform biomedical text annotation will allow use in an online environment
  - Support additional services: text summarization, semantic indexing and retrieval

- **Biomedical Text Summarization:**
  - New algorithms using domain-specific concepts improve system-generated summaries over existing summarization approaches

  - Personalization: concepts, rather than terms, can be used to generate custom biomedical summaries

Motivation

- Biomedical Semantic Annotation
  - *Purpose*: add machine-understandable meaning by finding domain-specific concepts within free-form texts

  - *Uses*:
    - (a) synonymous phrase merging, and
    - (b) semantic filtering

  - *How*: Domain-specific resources: ontology, thesaurus
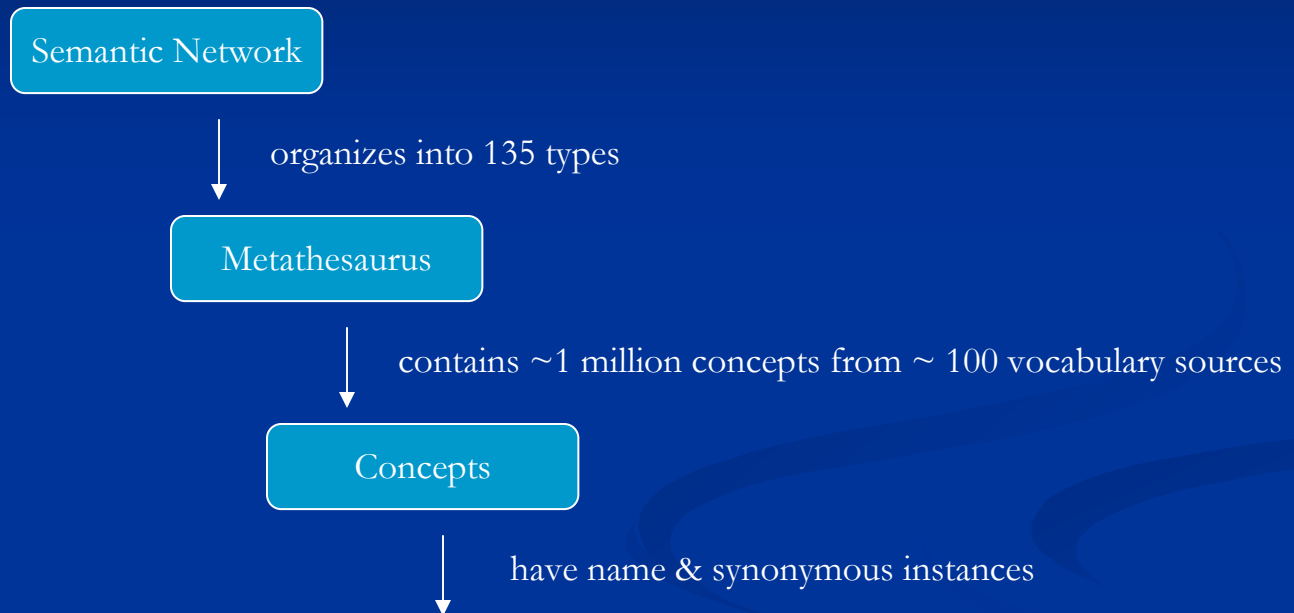

- Text Summarization
  - *Purpose*: Reduce amount of data required to understand main points of a text

  - *Uses*: Quick overviews of research, focus on certain details

  - *How*: Extract subset of sentences from source text

Overview

- Hypothesis / Approach / Contribution
- *UMLS Background*
- Literature Review
- Approach
- Evaluation
- Research Plan

Background - UMLS

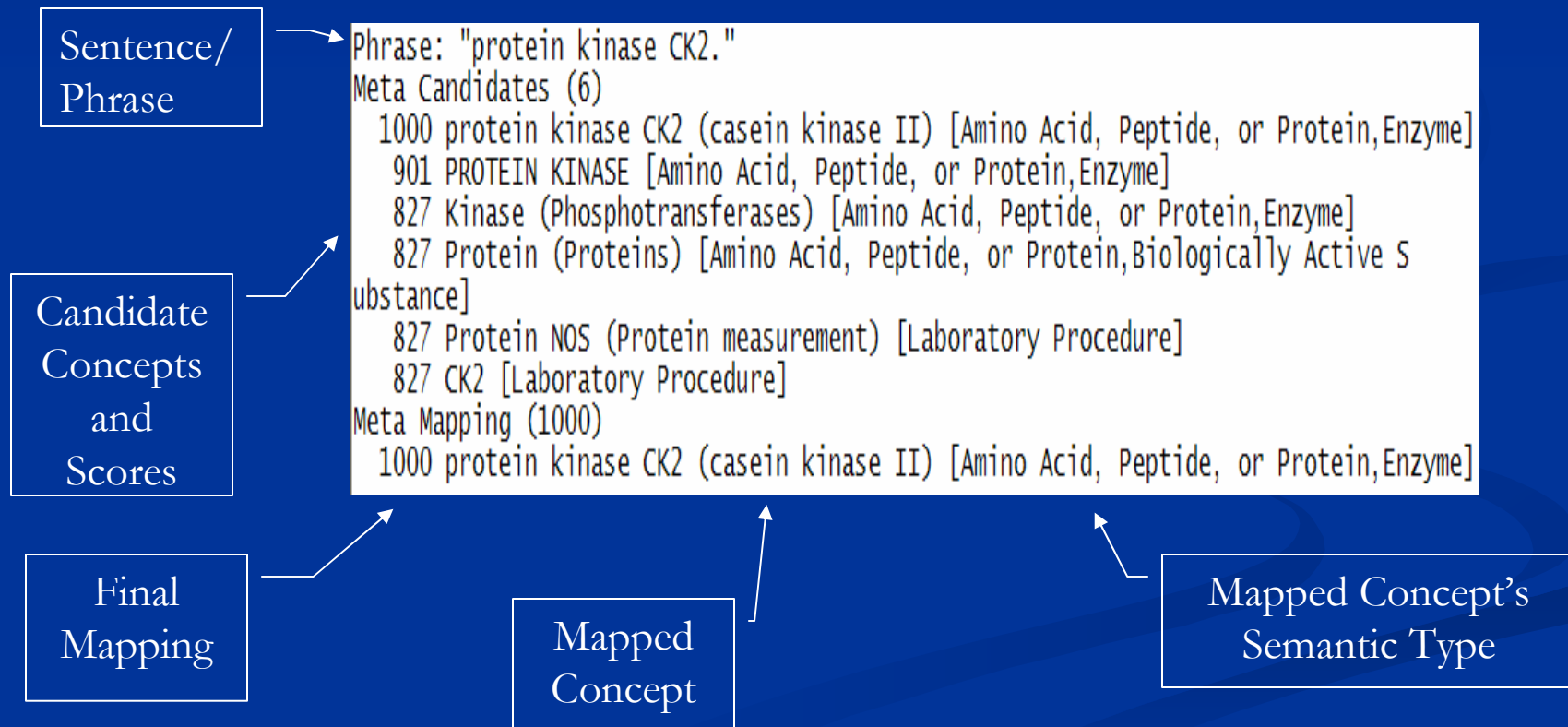## (Unified Medical Language System – U.S. National Library of Medicine)

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

Background – UMLS - 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]
```

Overview

- Hypothesis / Approach / Contribution
- *UMLS Background*
- Literature Review ⬅
- Approach
- Evaluation
- Research Plan

Literature Review of Semantic Annotation

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.

Literature Review - Annotation

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

Literature Review - Annotation

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 (United States National Library of Medicine, 2005d) 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 each 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

Literature Review - Annotation

Literature Review of
Text Summarization

Summarization Model

- **Three phrase model** (Sparck Jones, 1999)
  - Interpretation
    - Text analysis

  - Transformation
    - Content Selection, Concept Generalization

  - Generation
    - Generates output from Transformation intermediate form

Summarization Methods

(Mani & Maybury, 1999)

- **Surface-level**
  - **Statistical analysis** (Luhn, 1958), (A. Nenkova & Vanderwende, 2005) , (Vanderwende & Suzuki, 2005)
  - **Cue Phrases** (Pollock & Zamora, 1975), (Kupiec, Pedersen, & Chen, 1999)
  - **Term Locations** (Edmundson, 1999) , (Teufel & Moens, 1999)

- **Entity-level**
  - **Lexical Chaining** (Barzilay & Elhadad, 1997) , (Morris & Hirst, 1991), (Silber & McCoy, 2002), (Galley & McKeown, 2003)

- **Discourse-level**
  - **Text structure** (Strzalkowski, Stein, & Wang, 1999)
    - Text format, threads of topics, rhetorical structure

Extractive Summarization Process

- Score each sub-unit of the text (e.g., sentence, phrase) using some method which identifies important text

- Find the highest-scoring sub-units

- Extract the sub-units
  - Order the sub-units in their order of appearance

Overview

- Hypothesis / Approach / Contribution
- *UMLS Background*
- Literature Review
- Approach ←
- Evaluation
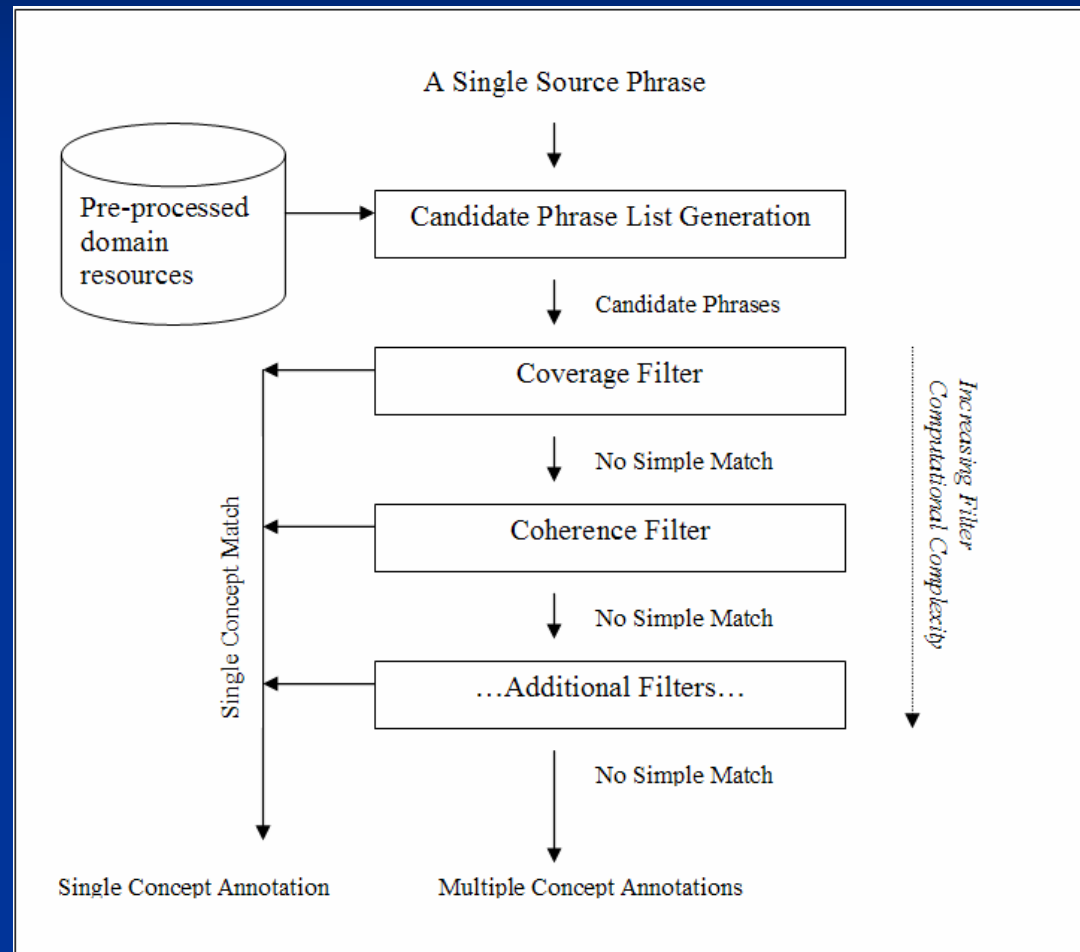- Research Plan

Approach:
Biomedical Semantic Annotation

(Can biomedical semantic annotation time be improved for use in an online environment?)

Preliminary Work

- Biomedical Annotator Design

- Developed Evaluation Corpus

- Publications:
  - 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.

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

  - 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.

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

Biomedical Annotator: Pre-processing

Pre-processing: Overview

- Purpose:
  - Reduce runtime calculations and parsing of UMLS phrases

- How:
  - Convert UMLS resources for efficient in-memory lookup

  - Convert words to base forms:
    - {*eyes, eyed, eying*} → {*eye*}

  - Calculate IPF values for all UMLS phrase words

Pre-processing: IPF Calculation

- Gives indication of word importance based on its usage within all UMLS concept phrases

    - Pre-processing calculates IPF value of each word in UMLS phrases

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

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

Biomedical Annotator:
Candidate Phrase List Generation

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, or
  - reduce the size of candidate list passed to the next stage

Candidate Phrase List Generation
(Short Source Phrase < 5 words)

- Find all concept phrases having at least one word in common with the source phrase
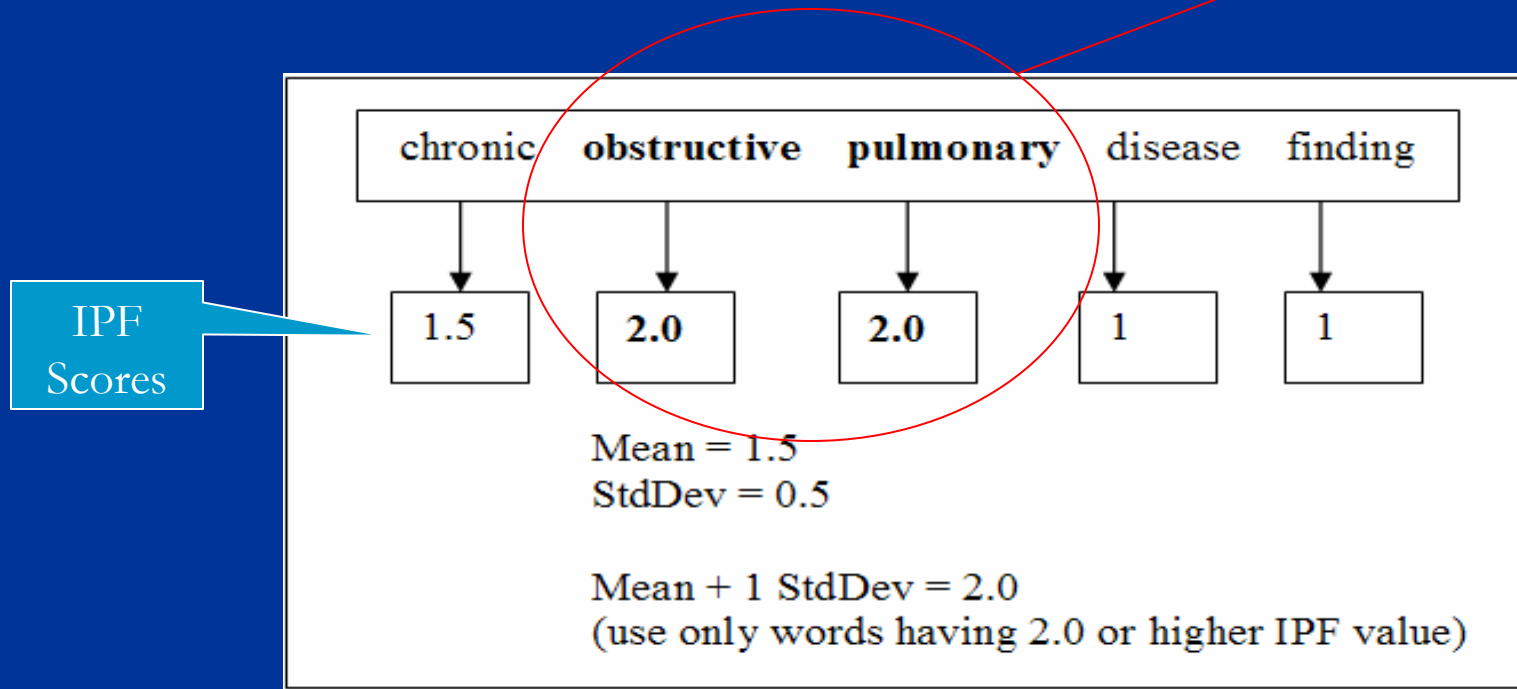
Source Phrase: 'lung cancer'		
Concept Id	Concept Name	Concept Phrase
0024109	Lung	Lung
0024117	Chronic Obstructive Airway Disease	Chronic Obstructive Lung Disease
0242379	Malignant Neoplasm of the Lung	Lung Cancer
0684249	Carcinoma of the Lung	Cancer of the Lung
0279000	Liver and Intrahepatic Biliary Tract Carcinoma	Liver Cancer

Approach – Biomedical Annotation

Candidate Phrase List Generation
(Long Source Phrase >= 5 words)

- Use only the important words to find candidate phrases
  - only the most likely candidate phrases will be placed into the candidate phrase list

Most significant words

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)

Approach – Biomedical Annotation

Biomedical Annotator:

Coverage Filter

Coverage Filter

- How well do words overlap between the source phrase and the UMLS concept phrase?

  - Use IPF to find best overlap of most meaningful words:

$$PhraseCoverageIPF = \sum_{i=1}^{N} IPF_i$$

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

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

Approach – Biomedical Annotation

Biomedical Annotator:

Coherence Filter

Coherence Filter

- Measures word *order* , not just *overlap*
  - Idea: Do the overlapping words make sense when put together?

- Measure using Skip-bigrams
  - Permits inexact string matching, but enforces word order

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

Approach – Biomedical Annotation

Coherence Filter

(Measured Using 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

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

Coherence Filter
(Example Evaluation Showing Significance of Word Order)
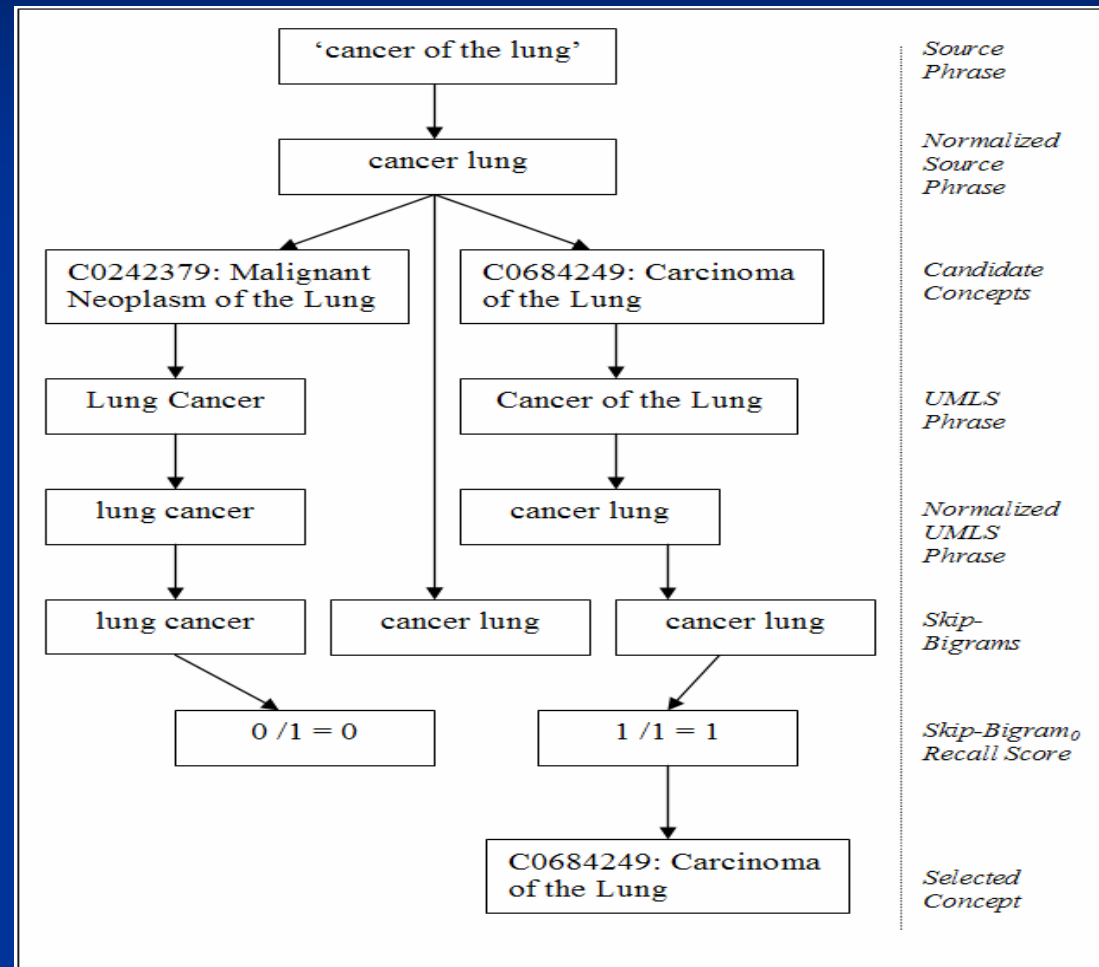
Possible Additional Filters

- ## 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

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

Approach:
Biomedical Text Summarization

1) Can the UMLS Semantic Network and UMLS Metathesaurus be used in the biomedical domain to identify important areas within a biomedical text?

2) Can existing text summarization approaches be adapted in new ways to utilize domain-specific concepts?

Preliminary Work

- Developed Evaluation Corpus

- Implemented and Evaluated Concept Chaining Summarizer:
  - 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 (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.

- Implemented and Evaluated Frequency Distribution Summarizer:
  - Lawrence Reeve, Hyoil Han and Ari D. Brooks. (2007) The Use of Domain-Specific Concepts in Biomedical Text Summarization, Journal of Information Processing and Management. Refereed. Accepted.

  - 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.
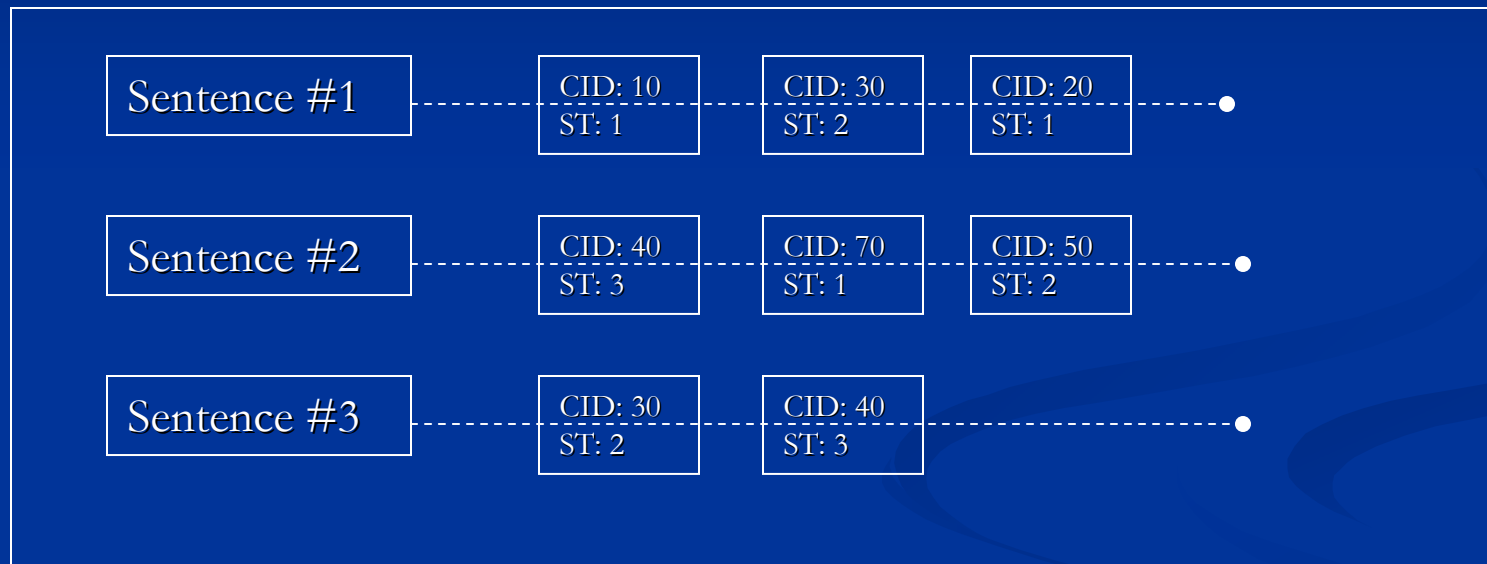
Approach – Biomedical Annotation

Concept Chaining Summarizer

Concept Chaining Process

1) Chain together related concepts using UMLS Semantic Network

2) Identify strongest chains

3) Find strongest concepts within strongest chains

4) Score each sentence based on number of strong concepts it contains

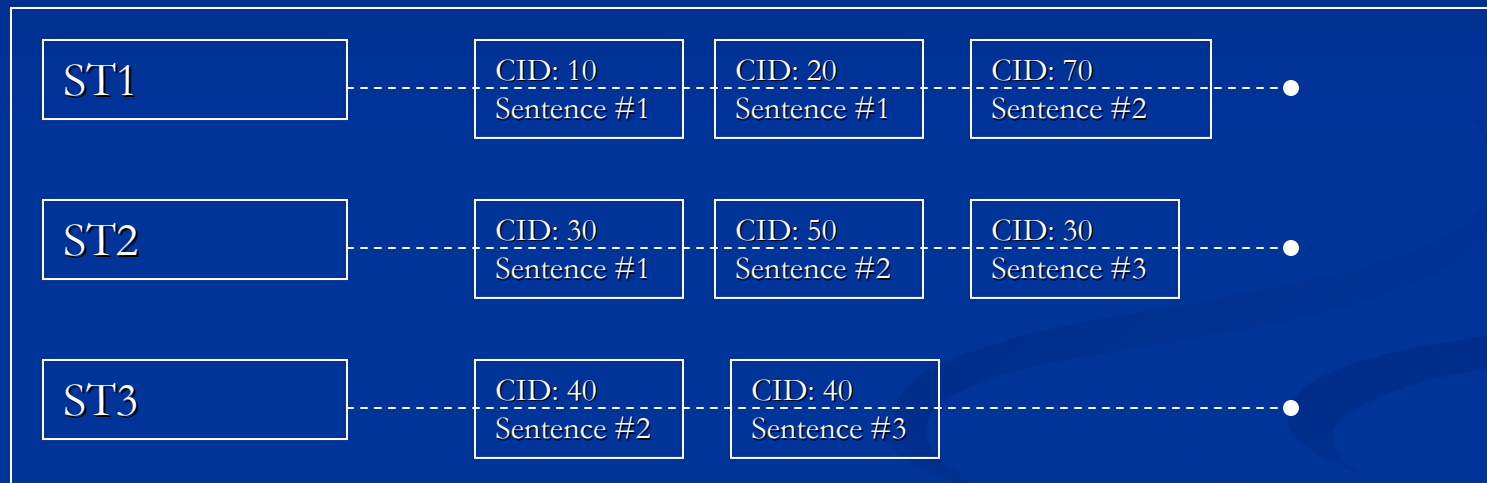5) Extract highest-scoring sentences

Concept Chain - Example

Source Text:

Sentence #1	CID: 10 ST: 1	CID: 30 ST: 2	CID: 20 ST: 1	●
Sentence #2	CID: 40 ST: 3	CID: 70 ST: 1	CID: 50 ST: 2	●
Sentence #3	CID: 30 ST: 2	CID: 40 ST: 3		●

CID = UMLS Concept Id
ST   = UMLS Semantic Type

Concept Chain – Example cont'd

Concept Chains (by Semantic Type):

ST1		CID: 10 Sentence #1	CID: 20 Sentence #1	CID: 70 Sentence #2	
ST2		CID: 30 Sentence #1	CID: 50 Sentence #2	CID: 30 Sentence #3	
ST3		CID: 40 Sentence #2	CID: 40 Sentence #3		

CID = Concept Id
ST   = Semantic Type

Strong Chain Identification

- The strongest chain(s) indicate the set of concepts most representative of the source text

- 1) Compute Chain score: $Score(Chain) = Frequency\ of\ most\ frequent\ concept * number\ of\ distinct\ concepts$
    - Uses reiteration and length

- 2) Find the strongest chains based chain scores:

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

(Barzilay & Elhadad, 1997)

Approach – Biomedical Summarization – Concept Chaining

Strong Concept Identification

- Perform frequency count on concepts within strongest chain
  - concepts with highest frequency are top concepts for each chain

- Score each sentence based on number of top concepts it contains

Concept Frequency Distribution Summarizer

Concept Frequency Distribution

- Match distribution of concepts in source text to summary

  - System-generated summary concept distribution should be close to source-text concept distribution

- Main problem: how to determine concept distribution similarity between the two

Approach – Biomedical Summarization

Concept 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

Approach – Biomedical Summarization

Concept Distribution Similarity

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

- Model system-generated summary and full-text as vector of concepts

- Evaluate 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

Overview

- Hypothesis / Approach / Contribution
- *UMLS Background*
- Literature Review
- Approach
- Evaluation ⬅
- Research Plan

Evaluation:

Corpus

Evaluation Corpus

- Selected 24 random papers from a database of ~1,200 oncology randomized controlled trial publications Brooks, A. D., & Sulimanoff, I. (2002).

  - Why 24 papers?
    - ROUGE research suggests significance @ 24 samples with 3 references (Lin, C. 2004)
    - Manual processing constraints on selecting more

Evaluation Corpus

- **Model Summaries:**
  - Author's abstract +
  - Three DUCoM student medical researchers
    - Generated an extractive summary of each paper

  - Why four model summaries per paper?
    - More references give a consensus
    - References in DUC 2001: 3, 2002: 2, 2003-2006: 4
    - ROUGE research used 3 references (Lin, C. 2004)

Evaluation:

Biomedical Semantic Annotation

Semantic Annotation Evaluation

- Compare proposed annotation system to a gold standard:
    - Verify concept identification accuracy is competitive
    - See the time required to perform annotation

Semantic Annotation Evaluation

- Two types of evaluation:
  - 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 against gold standard of MetaMap:
    - Speed: annotation time per phrase
    - Accuracy:

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

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

Approach – Annotation Evaluation

Semantic Annotation Evaluation

- **Extrinsic**
  - Use summarization to evaluate annotation output
    - FreqDist and modified SumBasic
    - Idea: If summarization performance using annotated concepts improves, improvement is due to better identification of concepts

Evaluation:

Biomedical Text Summarization

Summarization Evaluation

- Compare proposed system's summary performance to existing systems summarizer performance

- Review results to see if proposed system's summary generation performance is competitive or improved with existing system summarizers

Summarization Evaluation

- Approach:
  - 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

- Use ROUGE (Recall Oriented Understudy for Gisting Evaluation) (Lin & Hovy, 2003)
  - Compares system summaries to model summaries
  - Results based on n-gram overlap between models and system-summary

- 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)

Approach – Summarization Evaluation

Overview

- Hypothesis / Approach / Contribution
- *UMLS Background*
- Literature Review
- Approach
- Evaluation
- Research Plan

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

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.*

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.
- Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, Spain. 10-18. Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. Proceedings of the AMIA Symposium 2001, 17-21.
- Brooks, A. D., & Sulimanoff, I. (2002). Evidence-based oncology project. *Surgical oncology clinics of north america* (pp. 3-10)
- 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.
- Edmundson, H. P. (1999). New methods in automatic extracting. In I. Mani, & M. T. Maybury (Eds.), (pp. 23-42). Cambridge, MA: MIT Press.
- Galley, M., & McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence,* Acapulco,Mexico. 1486-1488.
- 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.
- Kupiec, J., Pedersen, J., & Chen, F. (1999). A trainable document summarizer. In I. Mani, & M. T. Maybury (Eds.), (pp. 55-60). Cambridge, MA: MIT Press.
- Lin, C. (2004). Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? *Proceedings of the NTCIR Workshop 4,* Tokyo, Japan.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.
- Mani, I., & Maybury, M. T. (1999). In Mani I., Maybury M. T. (Eds.), *Advances in automatic text summarization.* Cambridge, MA: MIT Press.
- Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics, 17*(1), 21-43.

References

- 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.

- Nenkova, A., & Vanderwende, L. (2005). *The impact of frequency on summarization* No. MSR-TR-2005-101). Redmond, Washington: Microsoft Research.

- Pollock, J. J., & Zamora, A. (1975). Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences, 15*(4), 226-232.

- Silber, G. H., & McCoy, K. F. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics, 28*(4), 487-496.

- Sparck Jones, K. (1999). Automatic summarizing: Factors and directions. In I. Mani, & M. T. Maybury (Eds.), *Advances in automatic text summarization* (pp. 2-12). Cambridge, MA: MIT Press.

- 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.

- Strzalkowski, T., Stein, G., & Wang, J. B, wise. A robust practical text summarizer. 1999. *Advances in Automatic Text Summarization,* , 137–154.

- Teufel, S., & Moens, M. (1999). Argumentative classificatin of extracted sentences as a first step towards flexible abstracting. In I. Mani, & M. T. Maybury (Eds.), (pp. 155-171). Cambridge, MA: MIT Press.

- United States National Library of Medicine. (2005d). *Unified medical language system (UMLS)*.http://www.nlm.nih.gov/research/umls/

- Vanderwende, L., & Suzuki, H. (2005). Frequency-based summarizer and a language modeling extension. *Proceedings of the MultiLingual Summarization 2005 Workshop,*

- 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.