College of Information Science and Technology

Biomedical Text Annotation and Summarization

Lawrence H. Reeve[1], Hyoil Han[1], Ari D. Brooks[2]
[1] Drexel University, College of Information Science and Technology
[2] Drexel University, College of Medicine

Abstract: A two-part text summarization system is described which addresses the scalability of utilizing the large volume of text-based information in the biomedical field. The contributions of the system are a) it utilizes biomedical concepts rather than terms to find the main points of a text, b) it uses two new concept-based algorithms to find important areas of a text for extracting sentences to form a summary, and c) it is supported by a new semantic annotation sub-system, which identifies biomedical concepts found in biomedical text documents. The semantic annotation sub-system uses a novel multiple-filter system architecture for online matching of concepts defined by a biomedical metathesaurus. The goal of semantic annotation is to show online text-to-concept mapping can be performed without a significant loss of precision as compared to current offline systems. An evaluation shows the text summarization algorithms using concepts outperform existing summarization systems, and the semantic annotation system performs twenty times faster than a state-of-the-art system with no significant loss of precision.

Figure 1: Text Summarization System and Semantic Annotation system

1. Biomedical Source Text

The source text used is a randomized controlled trial journal paper. Although the biomedical domain is used, any domain providing domain-specific resources can be used.

Biomedical Source Text

2a. Lexical Processing

Lexical processing identifies sentences and phrases within sentences. This task can be performed by any third-party component. In the evaluation system, lexical processing is provided by the National Library of Medicine's MetaMap application.

Lexical Processing

(Sentence Boundary and Phrase Detection)

2b. Domain-Specific Resources

The domain-specific resource provides a thesaurus for a particular domain. The thesaurus has a list of concepts and for each concept, a list of synonymous phrases.

Figure 2: UMLS Resource Organization

Domain-specific Resources

3. Semantic Annotation

Semantic Annotation maps phrases found in a source text to the best-matching concepts in the domain-specific resource.

UMLS provides a metathesaurus which specifies concepts and a list of synonymous phrases for each concept. Given a source text phrase, the task is to find which of the synonymous phrases (*candidate phrases*) best matches the source text phrase.

CONANN uses a multiple filter approach to remove unlikely candidate phrases. Current filters use inverse document frequency applied to phrases to measure *coverage* (overlapping terms), and skip-bigrams to measure *coherence* (term ordering).

4. Extractive Text Summarization

Extractive text summarization applies a score to each sentence, and then extracts the highest scoring sentences up to some user-specified limit to form a summary.

Our summarizers, *BioChain* and *FreqDist*, are designed to work with concepts, rather than terms. Biomedical terminology is highly variant, and using concepts allows for merging synonymous phrases within a source text, and for easy personalization of summaries by weighting some concepts more heavily than others.

Focus of research

CONANN – Semantic Annotator

Phrases annotated with domain concepts

BioChain, FreqDist – Extractive Text Summarizers

5. Generated Text Summary

To determine the summary size, the user specifies a percentage of the original source text to return (e.g., 20%) or a number of sentences. The summary can be tailored by weighting concepts to produce a personalized summary.

User Presentation

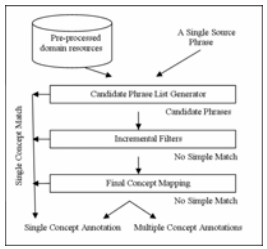

Figure 3: CONANN Architecture

Biochain:

Link related concepts together based on common UMLS Semantic Type. Find strongest chains and score sentences based on number of concepts belonging to strong chains.

Chain strength is measured using reiteration and chain length. Strong chains are chains which have a score greater than the average chain score + two standard deviations of all scores.

FreqDist (Frequency Distribution):
Find best matching distribution of concepts in source text when generating summary. The system-generated summary concept distribution should be as close as possible to the source-text concept distribution.
Five similarity functions were evaluated: Cosine, Dice, Euclidean, Simple Frequency, Vector Subtraction. Dice was found to perform best.

Motivation, Hypothesis and Method

Motivation: Generate short summaries of biomedical texts (randomized controlled trials in oncology) to allow physicians and researchers to assimilate more information in less time.

Approach: Identify and extract sentences from the source text to form a summary. The problem is how to identify important sentences within the full source text to extract while simultaneously reducing information redundancy.

Hypothesis: Domain-specific concepts can be used to identify important areas with a text which can then be extracted to form a summary.

Method: Annotate biomedical source texts with biomedical concepts and then use two novel algorithms which utilize concepts rather than terms to identify sentences which express the main themes of the text.

Publications

Reeve, L., Han, H., & Brooks, A. D. (2006). Biomedical Text Summarization Using Concept Chains. *International Journal of Data Mining and Bioinformatics, 1*(4), 389-407.

Reeve, L. H., Han, H., & Brooks, A. D. (2007). The Use of Domain-Specific Concepts in Biomedical Text Summarization. To appear in the *Journal of Information Processing and Management, Special Issue on Summarization. Elsevier.*

Reeve, L. H. & Han, H. (2007). CONANN: An Online Biomedical Semantic Annotator. To appear in the *Proceedings of Data Integration in the Life Sciences*. June 27-29, 2007.

Summarization Evaluation

Basic Idea

- Generate model summaries from domain experts
- Generate system summaries
- Use ROUGE to compare system and model summaries

Semantic Annotation Evaluation

Intrinsic

- Use NLM's MetaMap application to generate baseline of singly-mapped noun phrases
- Pass this set of phrases to CONANN
- Compare precision and time-to-annotate

Extrinsic

- Compare MetaMap vs. CONANN in text summarization performance using FreqDist and SumBasic summarizers

Observations

- CONANN Annotator:
  - Performs 20x faster than state-of-the-art system
  - Small loss of precision, acceptable in a user task
- BioChain, FreqDist Summarizers:
  - Concepts perform closely with terms
  - Concepts can be used to personalize summaries

Semantic Annotation Results

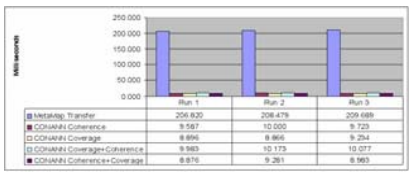

Figure 4: Average Phrase Annotation Time

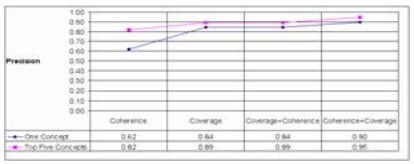

Figure 5: Annotator Precision

Summarization Method	*ROUGE-2 Score*	*ROUGE-SU4 Score*
FreqDist using MetaMap	0.1207	0.2200
FreqDist using CONANN	0.1192	0.2161
SumBasic using CONANN	0.1178	0.2098
SumBasic using MetaMap	0.1094	0.2003

Figure 6: ROUGE Scores (extrinsic evaluation)

Text Summarization Results

Summarizer	ROUGE-2	Summarizer	ROUGE-SU4
FreqDist-Term-Dice	0.13123	FreqDist-Term-Dice	0.23002
SumBasic-Term	0.12711	SumBasic-Term	0.22169
FreqDist-ConceptMMTX-Dice	0.12080	FreqDist-ConceptMMTX-Dice	0.21894
FreqDist-ConceptDUIST-Dice	0.11943	FreqDist-ConceptDUIST-Dice	0.21497
SumBasic-ConceptDUIST	0.11827	SumBasic-ConceptDUIST	0.20894
SumBasic-ConceptMMTX	0.10920	BiochainPlusFreqDist	0.19977
Lemur-MMR	0.10707	SumBasic-ConceptMMTX	0.19868
BiochainPlusFreqDist	0.10637	Lemur-MMR	0.19737
Biochain-MostFrequentStrongChainConcept	0.10419	Biochain-MostFrequentStrongChainConcept	0.19010
Biochain-AllStrongChainConcepts	0.09653	Biochain-AllStrongChainConcepts	0.18329
Mead	0.09160	Mead	0.17368
Baseline-Random	0.07913	Baseline-Random	0.16100
MSWord	0.07649	MSWord	0.14806
SweSum	0.07430	SweSum	0.14649
OTS	0.07364	OTS	0.14646
Baseline-Lead	0.07112	Baseline-Lead	0.13706
Author-Abstract	0.02951	Author-Abstract	0.06309
Expert-Summary	0.01462	Expert-Summary	0.02771

Figure 7: ROUGE-2 and ROUGE-SU4 scores for text summarization

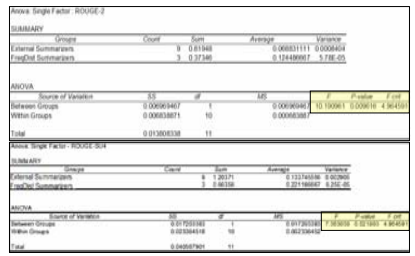

Figure 8: ANOVA: FreqDist vs. Other summarizers