



# BioChain: Lexical Chaining Methods for Biomedical Text Summarization

Lawrence Reeve<sup>1</sup>, Hyoil Han<sup>1</sup> and Ari D. Brooks<sup>2</sup>

College of Information Science and Technology<sup>1</sup>, College of Medicine<sup>2</sup>, Drexel University

## Topics

- Introduction
- Biomedical Text Summarization
  - Goal, Need & Approach
  - BioChain: Chaining of Biomedical Concepts
  - Text Summarization Process
- Evaluation
- Conclusions

## Introduction

- Finding for information from e-sources is difficult
  - Volume of text makes manual search and retrieval implausible
- Concept chaining (*BioChain*) to link semantically-related concepts within biomedical text together is proposed.
- BioChain is a novel concept chaining methodology.
  - BioChain is then applied to *biomedical text summarization*
    - Reduces the amount of text a user must read without loss of main ideas of the full-text
- Lexical cohesion is a property of text that causes a discourse segment to "hang together" as a unit.
- Lexical cohesion is important in computational text understanding
  - provides term ambiguity resolution, and
  - Provides information for determining the meaning of text.
- Lexical chaining is a method for determining lexical cohesion among terms in a text.
- Lexical chaining is useful for determining the *aboutness* of a discourse segment, without fully understanding the discourse.

## Goal

- Answer questions, such as: "What's the latest, best information on cancer treatment?"
  - Current focus is on oncology clinical trial papers
    - Using a database of ~1,200 manually processed papers produced by oncology domain experts
  - Current goal: Summarize a single clinical trial paper
  - Long-term goal: Summarize multiple clinical trial documents

## Source Text Input

- Abstract or full text from PubMed
  - Need to identify noun phrases within each sentence
    - Concepts are derived from noun phrases using vocabulary in metathesaurus
  - Conversion from PDF-formatted files:
    - Columns, Captions, Bibliography, Reference numbers, Images of documents, Text tables

## Approach

- Apply methods/concepts from lexical chaining:
  - Cluster (chain) words together based on semantic-relatedness
    - Words are chained together based on word 'senses' (concepts)
- Lexical Chaining...
  - identifies lexical cohesion
    - property causing sentences to 'hang together' (Morris & Hirst, 1991)
  - captures core themes of a text (aboutness)
  - is an intermediate format
- Example: (Doran et al., 2004)
  - "The house contains an attic. The home is a cabin."
  - Lexical Chain: dwelling → {house, attic, home, cabin}

## Need for summarization

- There is no 'ideal' summary
  - Depends on reader's information need
  - Author's abstract is one view of an ideal summary
  - May want alternative summaries
- Abstract may be missing ideas from the full-text
- Use in question-answering systems to provide personalized information
- Semi-automatic generation of abstracts for use in abstract services
- Evaluate sentence selection methods for use in multi-document summarization

## Concept Chaining Implemented Using UMLS

- UMLS (Unified Medical Language System)
  - Developed by National Library of Medicine
  - Resources used in BioChain:
    - Metathesaurus: Maps terms into concepts
    - Semantic Network: Organizes related concepts
    - MetaMap Transfer Application: text-to-concept mapping tool

## MetaMap Transfer (by NLM)

- Maps noun phrases to UMLS Metathesaurus concepts and UMLS Semantic Types (from MetaMap by National Library of Medicine)

```
Processing 00000000.tx.0: Obstructive Sleep Apnea
Phrase: "Obstructive Sleep Apnea"
Meta Candidates (7)
1000 Sleep Apnea, Obstructive [Disease or Syndrome]
901 Apnea, Sleep (Sleep Apnea Syndromes) [Disease or Syndrome]
827 Apnea [Pathologic Function,Sign or Symptom]
827 Obstructive (Obstructed) [Functional Concept]
827 Sleep [Mental Process]
755 Sleeplessness [Disease or Syndrome,Sign or Symptom]
755 Sleepy [Finding]
Meta Mapping (1000)
1000 Sleep Apnea, Obstructive [Disease or Syndrome]
```

## Abstract

Lexical chaining is a technique for identifying semantically-related terms in text. We propose *concept chaining* to link semantically-related concepts within biomedical text together. The resulting concept chains are then used to identify candidate sentences useful for extraction. The extracted sentences are used to produce a summary of the biomedical text. The concept chaining process is adapted from existing lexical chaining approaches, which focus on chaining semantically-related terms, rather than semantically-related concepts.

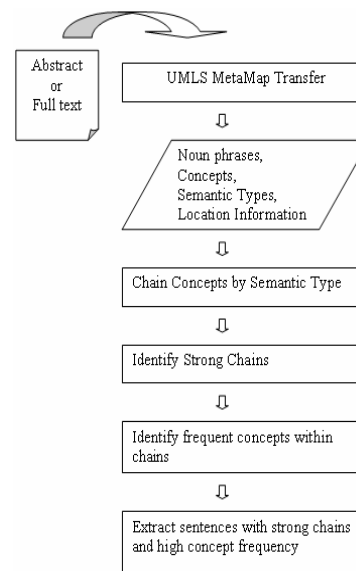


Figure 1 Concept Chaining Process

## BioChain: Concept Chaining

- Use semantic network to link together related concepts:
  - Ex: T081 - Quantitative (semantic type)
    - High dose (concept)
    - cm (concept)
    - Size (concept)
    - Median Statistical Measurement (concept)
- MetaMap Transfer:
  - Noun phrase → concept → semantic type
- BioChain:
  - Semantic type → concept, concept, concept
  - Each concept instance in semantic type entry contains original noun phrase, Sentence number, and Section (paragraph) number

**Publications:** Lawrence Reeve, Hyoil Han, and Ari D. Brooks, *BioChain: Lexical Chaining Methods for Biomedical Text Summarization*, *Appears in The 21st Annual ACM Symposium on Applied Computing 2006, Technical tracks on Bioinformatics (BIO 2006)*, Dijon, France, April 23-27, 2006.

The Unified Medical Language System (UMLS) Metathesaurus and Semantic Network are used as semantic resources. The UMLS MetaMap Transfer tool is used to perform text-to-concept mapping. The goal is to propose *concept chaining* and develop a novel concept chaining system for the biomedical domain using UMLS lexicon and the ideas of lexical chaining. The resulting concept chains from the full-text are evaluated against the concepts of a human summary (the paper's abstract).

## Chain Scoring

- Our scoring method
  - Take advantage of reiteration and length
  - Score (Chain) =  $\frac{\text{Frequency of most frequent concept} * \text{number of distinct concepts}}{\text{number of distinct concepts}}$
- Strong chains identify 'best' semantic types in text
  - Lexical chaining research generally uses:
    - two standard deviations above the mean of the scores computed for every chain in the document (Barday and Elhadad, 1997)

## Strong Chains – Example

- Top chains:
  - 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

## Text Summarization Process: Identifying Top Concepts & Sentence Extraction

- Part of sentence extraction process
- Get top chains (top semantic types) based on chain strength
- Perform frequency count on concepts with chains
  - concept(s) with highest frequency is top concept
- Use extractive approach

## Evaluation

- Avg precision=0.90, recall=0.92
- Avg # of strong chains in full-text is 3
  - Represents 2% of all possible semantic types
- Avg unique UMLS concepts in abstract is 8
  - Avg 80% coverage of concepts in filter
- Diversity test (abstract and paper do not match)
  - precision=0.00, recall=0.33

## Conclusions

- Text summarization for biomedical literature (specifically oncology)
- Use lexical chaining approaches with existing UMLS resources to identify the 'aboutness' of a text using concepts vs terms
- Extract sentences containing strongest concepts within a strong (semantic type) chain
- Result is an indicative summary of what text is about
- Evaluation shows concept chaining (BioChain) is strong between human summary and full-text