To appear in ***The 21st Annual ACM Symposium on Applied Computing 2006, Technical tracks on Bioinformatics* (*BIO 2006*), Dijon, France, April 23 -27, 2006.**

BioChain: Lexical Chaining Methods for Biomedical Text Summarization

Lawrence Reeve
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104 USA
lhr24@drexel.edu

Hyoil Han
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104 USA
hhan@cis.drexel.edu

Ari D. Brooks
College of Medicine
Drexel University
Philadelphia, PA 19102 USA
ari.brooks@drexelmed.edu

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. 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). Precision is measured at 0.90 and recall at 0.92. The resulting concept chains are used to summarize the text. We also evaluate generated summaries using existing summarization systems using sentence matching, and confirm the generated summaries are useful to a domain expert. Our results show that the proposed concept chaining is a promising methodology for biomedical text summarization.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing.*

General Terms
Algorithms, Measurement, Performance, Design.

Keywords
Text summarization, concept chaining, lexical chaining, biomedical text.

1. INTRODUCTION
Physicians and biomedical researchers need to master an ever increasing body of knowledge. While the Internet has made access to large databases of literature rapid and easy, summarization of the data remains difficult. There are many resources available to identify new knowledge once it is published. Once the articles are identified it remains the job of users to read through the abstract in order to determine if the information contained in the article is relevant and of good quality. Often, the abstract does not provide all the desired information making it essential to review the full article to make this decision. This process is time consuming, and if the search criteria are not specific enough, too many articles are identified and the task becomes prohibitively time consuming. This paper describes an important step toward automating the task of text summarization for document understanding. Eventually criteria for information type and measures of quality can be included to aid in the selection of the most relevant articles containing information of the best quality.

BioChain is an effort to summarize individual oncology clinical trial study publications into a few sentences to provide an indicative summary to medical practitioners or researchers. The summary is expected to allow the reader to gain a quick sense of what the clinical study has found. This work is being done as a joint effort between the Drexel University College of Information Science and Technology and College of Medicine. The College of Medicine has provided a database of approximately 1,200 oncology clinical trial documents that have been manually selected, evaluated and summarized. Our current goal is to develop approaches for summarizing single documents, with the ultimate goal of summarizing multiple documents into a single integrated summary in order to reduce the information overload burden on practicing physicians.

The rest of the paper is organized as follows. Section 2 details related work in the area of lexical chaining on which concept chaining is based. Section 3 describes the approach of chaining concepts to identify text themes. Section 4 presents the concept chaining process. Section 5 shows the results of evaluation. Section 6 summarizes the work.

2. RELATED WORK
Lexical chaining has been used for many years for text summarization. Lexical chaining is a method for determining

lexical cohesion among terms in text [8]. Lexical cohesion is a property of text that causes a discourse segment to "hang together" as a unit [9]. Lexical cohesion is important in computational text understanding for two major reasons: 1) providing term ambiguity resolution, and 2) providing information for determining the meaning of text [9]. Lexical chaining is useful for determining the "aboutness" of a discourse segment, without fully understanding the discourse. A basic assumption is the text must explicitly contain semantically related terms identifying the main concept. Lexical chains are an intermediate representation of source text, and are not used directly by an end-user. Instead, lexical chains are applied internally in some application; in our case, the application is text summarization for document understanding. We interchangeably use the term "document summarization" for "text summarization for document understanding."

Lexical chains for text summarization were first introduced by [9]. Their initial work described the approach, but did not implement it because electronic versions of a thesaurus were not available at the time. A thesaurus is used to relate words semantically; for example, through synonymy and hypernym/hyponym relationships. A machine implementation by [8] showed that the theoretical work by Morris/Hirst [9] could be practically realized for document summarization. While Barzilay/Elhadad proved the feasibility of computing lexical chains, their algorithm runs in exponential time. A linear time algorithm was later defined and implemented by [1]. A more recent implementation focuses on improving word sense disambiguation based on the idea of one sense per discourse [10]. All of these implementations use WordNet [2] as the knowledge source for identifying semantic relationships between terms. A computational model for semantic relationships between terms was developed by [2].

The UMLS MetaMap Transfer application has been used for applications such as hierarchical indexing query expansion, user query categorization and data mining for clinical finding, molecular binding expressions, drug and disease relationships, and drugs and gene relationships [11]. To our knowledge, MetaMap Transfer output has not been used to identify text themes using concept chaining.

3. CONCEPT CHAINING

We propose to apply the concepts and methods of lexical chaining to biomedical text using concepts rather than terms. Lexical chaining approaches use linkages among word instances to identify semantically-related terms. The resulting linkages are used to identify the themes of text. Terms are typically linked together based on word senses [1]. WordNet [2] is often the lexical resource for identifying term relatedness, using relationship types such as synonymy, hypernymy, and hyponymy.

The BioChain approach uses concept chaining rather than lexical chaining. Concept chaining operates at the level of concepts rather than terms. The Unified Medical Language System (UMLS) [3] provides tools for mapping biomedical text into concepts and semantic types. This semantic mapping allows chaining together related concepts based on each concept's semantic type. The UMLS semantic network types are used as the head of chains, and the chains are composed of concept instances generated from noun phrases in the biomedical text. There are three primary UMLS resources used in the chaining process: Metathesaurus, Semantic Network, and MetaMap Transfer [7]. The Metathesaurus incorporates multiple source vocabularies from the various providers of healthcare terminology, such as SNOMED [4], so vocabulary coverage is very wide. The Metathesaurus contains concepts, names and relationships and links alternative names and views of the same concept together [5]. In addition, the UMLS Metathesaurus identifies relationships between different concepts, using relationship types such as concept co-occurrence, synonymy, and structure (such as parent, child, and sibling). The Semantic Network provides a categorization of almost all concepts in the UMLS Metathesaurus, as well as relationships between concepts in the Metathesaurus. The UMLS Semantic Network currently consists of 135 semantic types and 54 semantic relationship types [6]. The MetaMap Transfer application [7] implements text-to-concept mapping using concepts in the UMLS Metathesaurus and semantic types in the Semantic Network.

4. CONCEPT CHAINING PROCESS

Figure 1 shows the flow of concept chain processing. Biomedical text is first fed into the UMLS MetaMap Transfer application to identify biomedical concepts and their semantic types. The generated concepts are then mapped into chains based on their semantic type(s). It is possible for one concept to appear in multiple semantic types. This generally occurs when MetaMap Transfer cannot disambiguate noun phrases in the text. Chains which contain the core concepts of text, known as strong chains, are then identified. Finally, the most frequent concepts within strong chains are identified and used to find and extract sentences. Each stage in the process is detailed below. Due to space limitations, examples for each stage in BioChain are not shown.

4.1 Text-To-Concept Mapping

The UMLS MetaMap Transfer application is responsible for finding UMLS Metathesaurus concepts in biomedical text [7]. It processes text through a series of stages [11]. The text is first split into sections, sentences are identified, and words are tokenized. Lexical resources or patterns are used to identify entities such as dates and locations. The part-of-speech tagger tags each word with its part-of-speech. The parser breaks sentences into phrases. The variant generation step identifies variants of a phrase, such as acronyms, synonyms, and derivational and spelling variations. The candidate retrieval stage retrieves all UMLS Metathesaurus concepts containing the variants. The retrieved candidate concepts are then evaluated, scored, and a final mapping determined by the highest scoring concept.

4.2 Concept Chaining

Identified concepts are chained based on their semantic type(s) using an array [10]. A *concept chain* is created for each semantic type defined in the UMLS Semantic Network. Each entry in the array contains a list of concepts belonging to the semantic type. Each concept entry in a semantic chain contains the concept, sentence number, section number (roughly

paragraph), and source noun phrase. If a concept belongs to multiple semantic types (i.e., multiple concept chains), BioChain allows the concept to appear in multiple concept chains. Concept disambiguation is not explicitly implemented. One semantic type (i.e., concept chain) is usually stronger than the other, where strength is observed as the number of concepts in a chain. Concepts in weaker chains appear to be eliminated from consideration by their low score (see section 4.3 for scoring). For future work, we plan to implement a disambiguation stage and compare the generated chains.

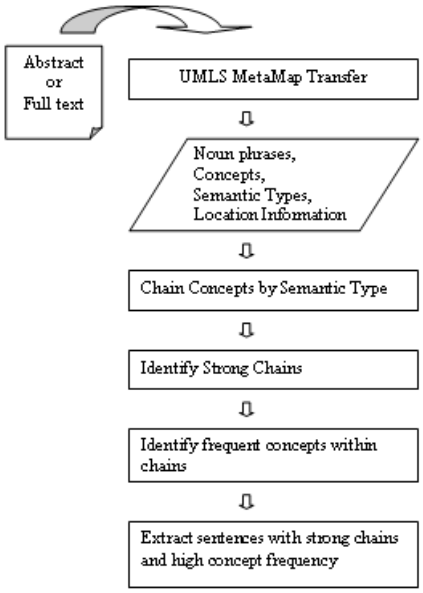

Figure 1: Concept Chaining Process

4.3 Identify Strong Chains

There has been no definitive measure for scoring chains, and the literature suggests changes in scoring methodology do not adversely impact chaining results [12]. The original lexical chain paper by [9] defines three types of strong chain features: 1) reiteration, 2) density, and 3) length. Reiteration is repetition of concepts throughout text. Density is physical proximity of concepts: concepts closer together are more likely to be related. Length is the number of concept instances within a chain. Our scoring method, shown in Figure 2, includes a combination of features as proposed by [12] and Barzilay/Elhadad [8]. Our domain expert identified the semantic types important within the oncology clinical trial domain. A chain is scored as zero if not in the list shown in Figure 4.

Once all chains are scored, strong chains, which identify the semantic types occurring most often, are computed. Lexical chaining research generally uses two standard deviations above the mean of all chain scores [8], as shown in Figure 3.

$$Score(Chain) = Frequency\ of\ most\ frequent\ concept * Number\ of\ distinct\ concepts$$

Figure 2: Chain scoring

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

Figure 3: Strong chain identification

4.4 Identify Frequent Concepts and Summarize

Summarization identifies sentences most likely capture the main ideas of text. BioChain uses the sentence extraction method to generate a summary [13]. The top-*n* sentences in text are extracted, using *n* as an upper bound on the number of sentences to select. Top sentences are identified by sorting strong chains into ascending order based on their score, and then identifying the most frequent concepts within each chain. Then sentences that include the most frequent concepts are extracted and consist of a summary. Multiple concepts having the same frequency count are considered equal, and sentences from each concept are extracted.

UMLS Semantic Type	UMLS Semantic Type Name
T37	Injury or Poisoning
T51	Event
T52	Activity
T61	Therapeutic or Preventative Procedure
T62	Research Activity
T67	Phenomena or Process
T81	Quantitative Concept
T169	Functional Concept
T170	Intellectual Product
T191	Neoplastic Process

Figure 4: Important Semantic Types for oncology clinical trials

5. EVALUATION

Evaluating lexical chains is difficult because it is unclear how to evaluate their quality independent of the application in which they are used [10]. The basic subjective question is: how does one know the quality of a chain? Two types of quantitative evaluation were performed. The first compares the generated summary against three existing summarization systems. The second compares a human summary (abstract) against the full text and defines measures of precision and recall. In addition to a quantitative evaluation, we used a domain expert to review the quality of the generated summaries, and received positive feedback. We also considered using ROUGE [14]. ROUGE measures a summary against several human-generated summaries, which were not available for our clinical trial texts.

Summaries generated from concept chains were compared against three existing systems. Two systems are commercially available: Microsoft Word summarization feature [15] and Copernic Summarizer [16], and one is a research system: SweSum [17]. The Copernic Summarizer uses a keyphrase extraction approach [18], while SweSum uses a term frequency approach in combination with a lexical resource [17]. The Microsoft Word summarization method is not known. The number of matching sentences is compared. The compression rate is 25% of the original source text. Compression rate

indicates the percentage of sentences from the source text which should be extracted in order to build a summary. For example, if the source text is 100 sentences and the compression rate is 25%, then a maximum of 25 sentences will be extracted to produce a summary. The compression rate is user-definable, and allows for controlling the length of a summary. Table 1 compares the abstract and full-text of two clinical trial research papers. The Document Id column shows an internal document tracking number, the Filtering column is whether or not the chains use the restricted semantic types in Figure 4, the Cancer Type column shows the type of cancer discussed in the source text, and Concept Chain Sentence Count column displays how many sentences were generated by BioChain. Filtering and non-filtering were both reviewed since the other systems perform no domain-specific filtering. Intuitively, we expected that the unfiltered summary would match more closely with the other systems. In one paper filtering helped in finding similar sentences with other systems, while in another paper filtering reduced similarity. In general, the Microsoft Word and SweSum have the most number of sentences in common with BioChain for full-text, while Copernic Summarizer is more similar to BioChain for abstracts. For accurate comparisons, we are planning a study utilizing medical staff for manual comparison among systems in Table 1.

To measure chaining performance, a human summary (paper abstract) is compared against the full-text. The main concepts of the full text should be reflected in the main concepts of the abstract. The two metrics proposed by [1] were used:

Recall: Percentage of strong chains from the full-text that have at least one concept in the summary.

Precision: Percentage of concept instances in the abstract that have at least one instance in the strong chains in the full-text.

Table 2 shows the precision and recall for 24 documents from the oncology clinical trials collection, and is based on the format presented by [1]. Column 1 is an internal document tracking number, and column 2 is the type of cancer that each paper is about. Columns 3-6 are derived from the output of BioChain analysis. Column 3 lists the number of strong chains found in the full-text. Column 4 is the total number of unique concepts found within the abstract. Column 5 is the number of strong chains having at least one concept in common with the abstract, defined as recall. Column 6 is the number of concepts in the abstract having membership in at least one strong chain, defined as precision. Average recall is 0.92 and the average precision is 0.90. We conclude that the abstract, treated as a human generated summary, accurately represents the concepts in the full-text. Although direct comparisons are not possible with the work of Silber/McCoy [1] because they are in a different domain with different lexical resources, our evaluation is based on their approach. Silber/McCoy report average recall of 0.83 and an average precision of 0.85. The average number of strong chains is 3, which is approximately 2%-3% of the 135 semantic types in UMLS. The average number of unique UMLS concepts in an abstract is eight, indicating coverage of the filtered concepts shown in Figure 4 is approximately 80% on average.

We also composed a diversity test where the abstract of one paper is compared against the full-text of another paper based on the same cancer type. Our initial concern was that the concept filtering was so narrow that all abstracts and papers on the same topic would show high precision and recall. The test shows recall is 0.33 and precision is 0.00, indicating the diverse abstract and full-text are not good matches, and that the evaluation method is a good indicator of matching a human generated summary (i.e., abstract) to the full-text.

6. CONCLUSION

Using UMLS resources, a concept chaining methodology was proposed and developed. Concept chaining applies lexical chaining methods to link semantically-related concepts within biomedical text into chains. The strongest chains are identified and used to extract sentences in order to form a summary of the text. The resulting concept chains from the full-text are evaluated against the concepts of a human summary (i.e., the paper's abstract). Precision is measured at 0.90 and recall at 0.92. Our results show that the proposed concept chaining is an excellent methodology for biomedical text summarization. Although this method can be generally applied, the domain was focused on oncology clinical trial texts. Domain-specific filtering on the chain was performed. Our future plans are to 1) implement concept disambiguation and 2) improve sentence extraction. In addition, our ultimate goal is to summarize the results of multiple clinical trial texts.

7. REFERENCES

[1] G.H. Silber and K.F. McCoy, "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization," *Computational Linguistics*, vol. 28, 2002.
[2] C. Fellbaum, *WORDNET: An Electronic Lexical Database*, The MIT Press, 1998.
[3] United States National Library of Medicine, "Unified Medical Language System (UMLS)," 2005.
[4] SNOMED International, "SNOMED Clinical Terms," 2005.
[5] United States National Library of Medicine, "UMLS Metathesaurus Fact Sheet," 2004.
[6] United States National Library of Medicine, "UMLS Semantic Network Fact Sheet," 2004.
[7] United States National Library of Medicine, "MetaMap Transfer," 2005.
[8] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization," in Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, 1997, pp. 10-18.
[9] J. Morris and G. Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text," *Computational Linguistics*, vol. 17, pp. 21-43, 1991.
[10] M. Galley and K. McKeown, "Improving Word Sense Disambiguation in Lexical Chaining," in Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, 2003, pp. 1486-1488.
[11] A.R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," in Proceedings of the AMIA Symposium 2001, 2001, pp. 17-21.
[12] W.P. Doran, N.S. Stokes, J. Dunnion and J. Carthy, "Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization," in Proceedings of the 5th International conference on Intelligent Text Processing and Computational Linguistics, 2004.

To appear in ***The 21st Annual ACM Symposium on Applied Computing 2006, Technical tracks on Bioinformatics (BIO 2006)*, Dijon, France, April 23 -27, 2006.**

[13] S.D. Afantenos, V. Karkaletsis and P. Stamatopoulos, "Summarization from Medical Documents: A Survey "*Artificial Intelligence in Medicine*, vol. 33, pp. 157-177, 2005.
[14] C. Lin, "Recall-Oriented Understudy for Gisting Evaluation (ROUGE)," 2005.
[15] Microsoft Coporation, "Microsoft Word 2002," 2002.
[16] I. Copernic Technologies, "Copernic Summarizer," 2005.
[17] H. Dalianis, "SweSum - A Text Summarizer for Swedish," NADA, KTH., Stockholm, Sweden, Tech. Rep. TRITA-NA-P0015, 2000.

Table 1: Comparison of generated sentence output with other summarization systems.

				# (%) Sentences Matched with Concept Chaining		
Document Id	Filtered?	Cancer Type	Concept Chain Sentence Count	Microsoft Word 2002 (Frequency)	SweSum (Frequency)	Copernic Summarizer (Keyphrase)
1001-Abstract	Yes	Colorectal	3	2 (66%)	2 (66%)	1 (33%)
1197-Abstract	Yes	Sarcoma	3	2 (66%)	2 (66%)	3 (100%)
1001-Full-Text	Yes	Colorectal	37	7 (19%)	4 (11%)	6 (16%)
1197-Full-Text	Yes	Sarcoma	47	25 (53%)	23 (49%)	19 (40%)
1001-Abstract	No	Colorectal	3	1 (33%)	1 (33%)	2 (66%)
1197-Abstract	No	Sarcoma	3	2 (66%)	2 (66%)	3 (100%)
1001-Full-Text	No	Colorectal	37	13 (35%)	7 (19%)	5 (14%)
1197-Full-Text	No	Sarcoma	47	16 (34%)	13 (28%)	11 (23%)

Table 2: Precision and Recall of Concept Chains: Abstract vs. Full-text

Document Id	Cancer Type	Total Strong Chains in Full-Text	Total Concepts in Abstract	Strong Chains with Corresponding Concepts in Summary (Recall)	Concepts in Abstract with Corresponding Strong Chains in Full-Text (Precision)
0162	Breast	3	9	2 (0.67)	9 (1.00)
0234	Breast	2	7	2 (1.00)	7 (1.00)
0271	Breast	2	4	1 (0.50)	4 (1.00)
0312	Breast	2	8	1 (0.50)	4 (0.50)
0872	Breast	2	9	2 (1.00)	8 (0.89)
0954	Breast	3	11	3 (1.00)	11 (1.00)
1001	Colorectal	4	19	4 (1.00)	19 (1.00)
1108	Cervical	3	10	3 (1.00)	10 (1.00)
1110	Cervical	3	6	3 (1.00)	6 (1.00)
1111	Cervical	3	5	3 (1.00)	5 (1.00)
1115	Cervical	2	18	1 (0.50)	12 (0.67)
1117	Cervical	4	14	4 (1.00)	12 (0.86)
1118	Cervical	4	9	4 (1.00)	9 (1.00)
1122	Cervical	4	9	4 (1.00)	7 (0.78)
1132	Cervical	3	7	3 (1.00)	7 (1.00)
1154	Breast	4	9	4 (1.00)	8 (0.89)
1197	Sarcoma	4	12	3 (0.75)	12 (1.00)
UNK1	Breast	4	20	4 (1.00)	19 (0.95)
UNK2	Breast	1	9	1 (1.00)	8 (0.89)
UNK3	Breast	3	7	3 (1.00)	5 (0.71)
Averages		60	202	55 (0.92)	182 (0.90)
Diverse test: 0162 Abstract & 0954 Full-text	Breast/Breast	3	9	1 (0.33)	0 (0.00)