BioChain: Concept Chains 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

BioChain: Concept Chains for Biomedical Text Summarization

KEYWORDS:

Biomedical text summarization, concept chaining, lexical chaining.

ABSTRACT

Searching for information on e-sources is difficult due to the large amount of data that exist. Searching for such an abundant amount of data manually is not plausible. This paper is about an intelligent text analysis, the purpose of which is to provide the life science community with a means to overcome the information overload problems. We propose *concept chaining* (*BioChain*) to link semantically-related *concepts* within biomedical text together, using methods from lexical chaining. The goal of BioChain is to produce a novel concept chaining methodology. BioChain is then applied to *biomedical text summarization*, using both abstracts and full-text.

1. INTRODUCTION

Searching for and managing information on the Web and other e-sources (electronic data sources) is difficult due to the large amount of data that exist. The World Wide Web made it possible for people and organizations to create and publish their work on the Web and e-sources, which created a lot of information. Searching for and/or processing such an abundant amount of data manually is not plausible. Therefore, we are living in a world that is overloaded with information. This information is scattered across the Internet and other e-sources, making it very hard for us to find what we need. This paper is about an intelligent text analysis methodology, the purpose of which is to provide the life science community with a means to overcome the information overload problems. Our work mainly focuses on text analysis of semantic relatedness of terms found in biomedical literature. We propose *concept chaining*, *BioChain*, to link semantically-related *concepts* within biomedical text together, using methods from lexical chaining. The biomedical lexical and semantic resources provided by the Unified Medical Language System (UMLS), such as Metathesaurus and Semantic Network, as well as the text-to-concept mapping tool MetaMap Transfer, are used to implement concept chaining in the biomedical domain. The goal of BioChain is to produce a novel concept chaining methodology using UMLS resources and the ideas of lexical chaining. The results of chaining text concepts based on semantic types are then applied to biomedical text summarization, using both abstracts and full-text.

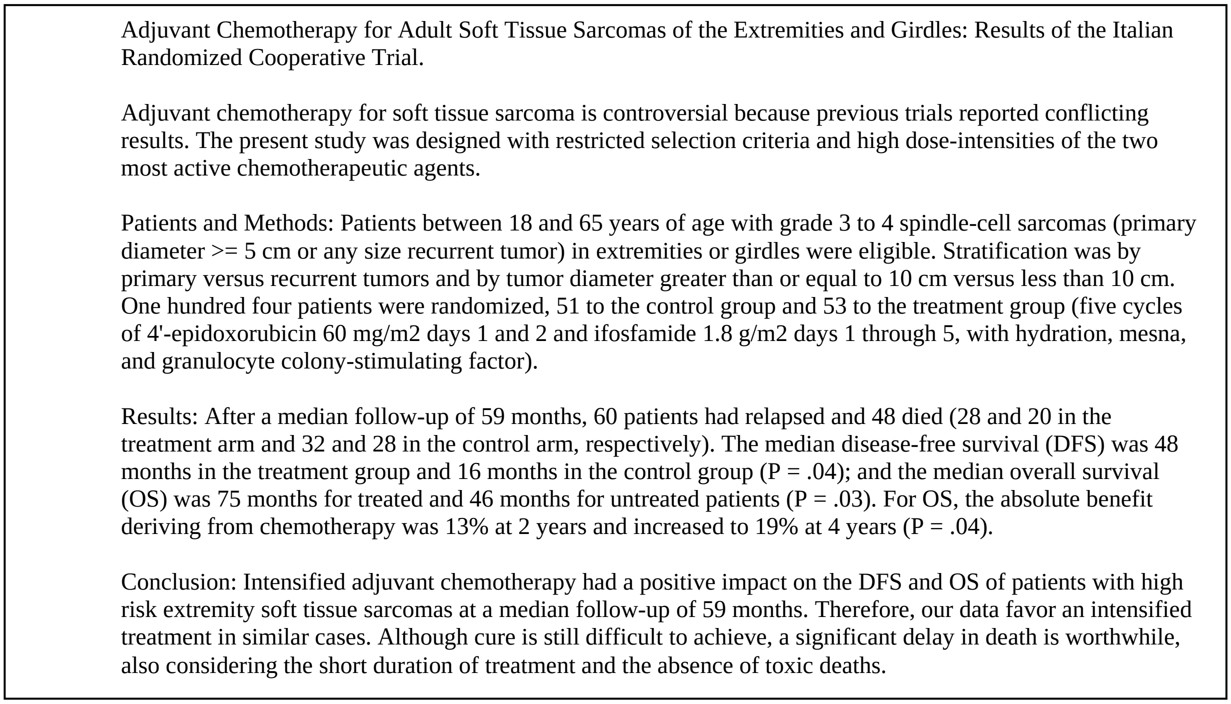
Oncologists are faced with the daunting task of evaluating the results of many published clinical trials in order to provide the best available cancer treatment to their patients. The PUBMED database alone contains 12 million citations from over 4,800 journals [1]. The U.S. National Institutes of Health Clinical Trials database contains information on over 13,500 trials [2]. Practicing oncologists must find the trial information related to their specialty, evaluate the study for its strength, and then possibly incorporate the new study information into their patient treatment efforts. The large number of clinical trials conducted and the data produced by them results in a classic case of information overload. To help alleviate this problem, BioChain is an effort to summarize each clinical trial results document into a few sentences to provide an indicative summary to medical practitioners. The summary is expected to allow the reader to gain a quick sense of what the study has found. The contribution of this work is to propose the use of concept chains rather than lexical chains based on terms to identify text themes. The use of concepts allows the expression of a single idea in multiple lexical forms. In addition, since the concepts are pre-defined, domain-specific filtering can be used to identify pertinent sections of a text. Our initial work on BioChain is described in [3]. This paper extends that work by explaining in detail concept chaining, and our evaluation methodology.

BioChain: Concept Chains for Biomedical Text Summarization

Figure 1 shows an abstract of a biomedical document describing treatment for a form of cancer known as Sarcoma. Figure 2 shows a summary of the abstract constructed using BioChain. In this example, BioChain has determined that the strongest concepts in the abstract are quantitative, and has then extracted the sentence with the most concentrated set of quantitative concepts, which happens to be the study results on the effectiveness of a treatment. BioChain detects the strongest concepts in a text, and so the area where a sentence is extracted may not always be the study results section. For example, if the author of a clinical study focuses mostly on the chemicals used in a chemotherapy application, the extracted sentence will be about the chemicals used rather than the empirical result of using the chemicals for treatment. A way to get around this limitation is to run BioChain over particular sections of a document, such as the Results section. If just the results section is summarized, then the strongest concepts for a clinical trial text is expected to be the quantitative outcome, and a sentence describing this outcome is more likely to be selected. When summarizing a series of clinical trial papers, the user can be presented the title of the study along with a summary of the results section, allowing for selecting the most promising papers. The current approach of BioChain uses the entire text for summarization, mostly because we are interested in looking at the concept chains produced.

This project is being done in conjunction with Dr. Ari Brooks, an oncologist with the Drexel University College of Medicine. Dr. Brooks and his colleagues have provided a database of approximately 1,200 clinical trials 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.

Figure 1: Sample biomedical document abstract



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

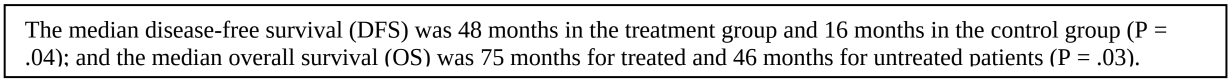
Adjuvant chemotherapy for soft tissue sarcoma is controversial because previous trials reported conflicting results. The present study was designed with restricted selection criteria and high dose-intensities of the two most active chemotherapeutic agents.

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.

Figure 2: Biomedical results extract produced by BioChain



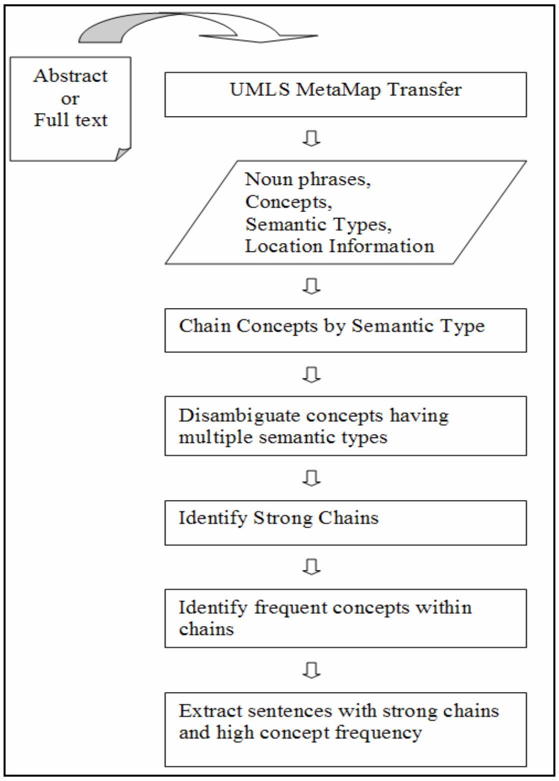
The median disease-free survival (DFS) was 48 months in the treatment group and 16 months in the control group (P = .04); and the median overall survival (OS) was 75 months for treated and 46 months for untreated patients (P = .03).

2. BioChain

The general idea for BioChain approach is to apply the concepts and methods of lexical chaining to biomedical text using concepts rather than terms. Typical lexical chaining approaches use linkages among term instances to identify semantically-related terms and identify the core themes of a text. Terms are typically linked together by word senses [4] using WordNet [5] as the lexical resource for identifying term relatedness using relationship types such as synonymy, hypernymy, and hyponymy.

BioChain approach uses concept chaining rather than lexical chaining. The Unified Medical Language System (UMLS) [6] provides tools for mapping biomedical text into concepts and semantic types. This mapping allows BioChain to operate at the level of concepts rather than terms. By identifying concepts from terms, BioChain can take advantage of a semantic type network predefined by domain experts. The semantic types are used as the head of chains, and the chains are composed of concept instances generated from noun phrases in the biomedical text. Figure 3 shows the flow of BioChain processing. Biomedical text is first fed into the UMLS MetaMap Transfer application, which identifies 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. BioChain performs its own disambiguation of concepts so that only one concept appears across all of the chains. That is, one concept appears in only one chain. Chains containing the core concepts of a text, known as strong chains, are then identified. Finally, the most frequent concepts within a strong chain are identified and used to find and extract sentences most representative of the frequent concepts.

Figure 3: BioChain Process


Abstract
or
Full text
UMLS MetaMap Transfer
Noun phrases,
Concepts,
Semantic Types,
Location Information
Chain Concepts by Semantic Type
Disambiguate concepts having multiple semantic types
Identify Strong Chains
Identify frequent concepts within chains
Extract sentences with strong chains and high concept frequency

There are three primary UMLS resources used in the chaining process: Metathesaurus, Semantic Network, and MetaMap Transfer. The Metathesaurus incorporates multiple source vocabularies from the various providers of healthcare terminology, such as SNOMED [7]. The Metathesaurus contains concepts, names and relationships and links alternative names and views of the same concept together [8]. In addition, the UMLS Metathesaurus identifies relationships between different concepts, using relationship types such as concept co-occurrence, synonymy, and parent, child, and sibling. The Semantic Network provides a categorization of 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 [9]. The MetaMap Transfer application [10] implements text-to-concept mapping using concepts in the UMLS Metathesaurus and semantic types in the Semantic Network.

2.1 Text-To-Concept Mapping

The UMLS MetaMap Transfer application is responsible for finding UMLS Metathesaurus concepts in text [10]. It processes text through a series of stages, shown in Figure 4 as excerpted from [10], and also described by [11]. The input is biomedical text. For this project, documents identified as Dr. Brooks as being clinical papers in oncology are obtained from PubMed in abstract form; the full-text is retrieved from the paper’s copyright holder. The format of the full-text is usually in PDF format, and must be converted to text-only format. This requires removing graphics (such as figures and tables) as well as their captions. The idea is to get only the sequential paragraph text, and not any supporting text. Once the input is available in plain-text format, it is first split into sections (titles, subtitles, paragraphs, etc), sentences are identified, and words are tokenized. Lexical lookup uses lexical resources or patterns to identify entities such as dates and locations. The part-of-speech tagger tags each word with various parts-of-speech tags. 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 Metathesaurus concepts containing the variants. The retrieved candidate concepts are then evaluated, scored, and a final mapping determined by the highest scoring concept. It is possible for a phrase to map to multiple concepts. In addition, a concept may belong to multiple semantic types, and due to ambiguity, the final identification of semantic type may result in multiple semantic types. Figure 5, excerpted from [10], shows the output when MetaMap Transfer analyzes the phrase “Obstructive Sleep Apnea”.

Figure 4 – UMLS MetaMap Transfer process (Excerpted from [10])

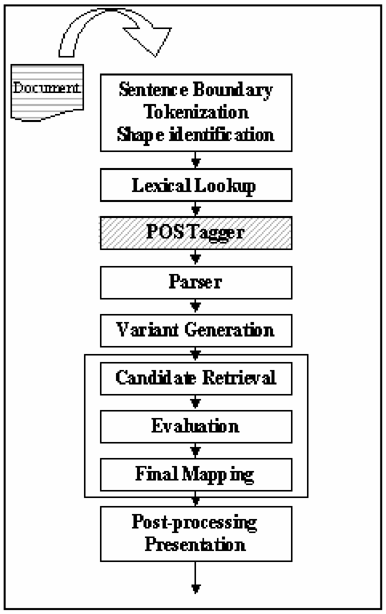
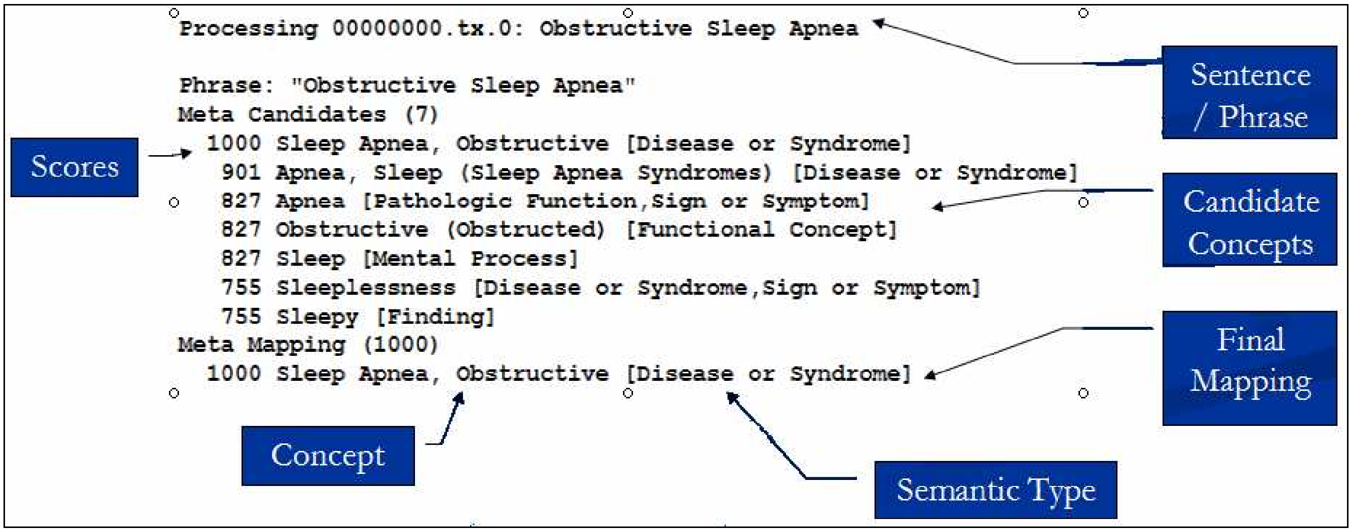


Figure 5:MetaMap Transfer Output (Excerpted from [10])



```
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]
```

2.2 Concept Chaining

Once MetaMap Transfer has evaluated the text and generated concepts and semantic types for each phrase, concepts can be chained. The data structure used to chain concepts is a simple array indexed by the semantic type. Each entry in the array contains a list of concepts belonging to the semantic type. The idea for using an array for linking related concepts is inspired by the algorithm proposed by [12] for lexical chaining in constant time. Before concept chaining occurs, the list of semantic types is read in from the UMLS file “SRDEF.” Each semantic type is identified by a unique identifier, such as T081 - Quantitative Concept. The numeric portion of identifier is used as an index into the chaining array. For example, identifier T081 is mapped to position 81 in the array. BioChain walks through the text document and extracts each sentence from the MetaMap Transfer output. Each concept MetaMap Transfer finds in the sentence is added to the semantic type chain(s) specified by MetaMap Transfer. Each entry in the semantic chain contains the concept, sentence number, the section number (roughly paragraph), and noun phrase. Figure 6 shows a semantic type chain generated from the abstract shown in Figure 1. The semantic type is ‘Quantitative Concept,’ and its type number is T081. The chain is ordered by occurrence of phrases as they are encountered in the text. The concept shown is the UMLS Metathesaurus concept as determined by MetaMap Transfer. For the abstract in Figure 1, approximately 21 such chains are constructed from the 135 possible semantic types defined by the UMLS Semantic Network.

Figure 6: Semantic type chain based on Figure 1 abstract

```
T081 - Quantitative Concept: 14.0
        phrase: high dose intensities
                concept: High dose (qualifier value), sentence#3, section#1
        phrase: primary diameter cm
                concept: cm (qualifier value),  sentence#5, section#2
        phrase: size recurrent tumor
                concept: Size (attribute), sentence#5, section#2
        phrase: One hundred four patients
                concept: One (qualifier value), sentence#6, section#2
        phrase: median disease-free survival
                concept: Disease-Free Survival, sentence#9, section#3
        phrase: median disease-free survival
                concept: Median Statistical Measurement, sentence#9, section#3
        phrase: median overall survival
                concept: survival aspects,  sentence#10, section#3
        phrase: median overall survival
                concept: Median Statistical Measurement, sentence#10, section#3
        phrase: absolute benefit
                        concept: benefits, sentence#11, section#3
```

2.3 Chain Disambiguation

Some concepts have multiple semantic types. This usually occurs when MetaMap Transfer is unable to disambiguate the concept. For example, the concept *Ifosfamide* belongs to several semantic types: 1) Organophosphorus Compound, 2) Pharmacologic Substance, and 3) Gene or Genome. The key problem is to determine which semantic type is best in context. Disambiguation is required because the idea of chaining concepts is to find the theme of a document, and without concept disambiguation, the wrong theme could be identified. In the current BioChain approach, concept disambiguation is not explicitly implemented. With the documents examined to date, it appears that when concepts fall into multiple semantic types, one semantic type is usually stronger than the other, where strength is observed as the number of concepts in a chain. Since chain scoring is computed as a measure of chain length (number and type of concepts), the weaker concepts appear to be eliminated from consideration by their low score. We cannot always count on this behavior, so we plan to implement a disambiguation stage.

There are two ideas for performing concept disambiguation. The first is to use chain length to remove concepts. When concepts fall into more than one semantic type, the concept that occurs in the longest chain is kept, while the same concept is removed from all other semantic type entries. The second approach is to use additional UMLS information to judge concept-concept strength. The UMLS Metathesaurus provides detailed information on concept co-occurrence and parent, child, and sibling relationships with other concepts. The idea is to find out which concepts in a chain co-occur more often together, or are more related together. A scoring function is applied, where each relationship type is given a numeric value indicating its usefulness in determining semantic relatedness between concepts. The latter is the approach to disambiguation done often in lexical chaining, as suggested in the papers by [4], [12], and [13], among others.

2.4 Identifying Strong Chains

Once chains of concepts based on semantic types have been constructed and disambiguated, they are scored to give an indication of their contribution to the text. Morris and Hirst [14] defined three types of features likely to contribute to determining strong chains. The features are 1) reiteration, 2) density, and 3) length. Reiteration is the repetition of concepts throughout a text. Density relates to the physical proximity of concepts: the closer they are together the more likely they are related and not just random occurrences. Length is the number of concept instances within a chain.

There has been no definitive measure of scoring chains, and some suggest that changes in scoring methodology do not adversely impact chaining results [15]. For BioChain, several methods were implemented. The first scoring method uses the Barzilay and Elhadad method [16]:

```
Score(Chain) = Length * Homogeneity, where:
  Length = number of occurrences of members in the chain, and
  Homogeneity =  1 – (number of distinct noun occurrences / Length)
```

The second approach uses word distance scoring, inspired by the observation of [14] that the further away terms are, the less likely they are to be related:

```
Score(Chain): distance = Termdistance$_2$ – Termdistance$_1$
            if (distance >= 0 && distance <= 3)
                  score = 1.0
            else if (distance >= 4 && distance <= 6)
                  score = 0.5
            else if (distance >= 7 && distance <= 9)
                  score = 0.25
```

A final scoring method that we developed also produced reasonable output for BioChain used a combination of methods as proposed by [15] and Barzilay/Elhadad [16], and take advantage of reiteration and homogeneity:

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

Once all chains are scored, the strongest chains can then be determined. Strong chains identify the ‘best’ semantic types in text; that is, those semantic types which occur most often in the text, as identified by the concepts derived from noun phrases in the text. Lexical chaining research generally uses two standard deviations above the mean of the scores computed for every chain in the document [16]:

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

Figure 7 shows the top semantic type chains generated from the abstract shown in Figure 1. The top chains displayed are all chains whose score is greater than zero. The strongest chains are shown in

Figure 8. Although the lexical chaining standard is to use two standard deviations, for short text such as an abstract, a limited number of chains can result, as shown in Figure 8, where only one chain is considered strong. To get a sense of the other chains that might be strong, chains with scores exceeding one standard deviation are also displayed in Figure 8.

Figure 7: Top semantic type chains generated from the abstract in Figure 1.

```
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
```

Figure 8: Strong chains

```
Two standard deviations :
        Avg score:    1.6666666666666667
        Std Dev:      3.0671497204093914
        Strong Score: 7.80096610748545
        T081-Quantitative Concept: 14.0
One standard deviation:
        Avg score:    1.6666666666666667
        Std Dev:      3.0671497204093914
        Strong Score: 4.733816387076058
        T081-Quantitative Concept: 14.0
        T061-Therapeutic or Preventive Procedure: 6.0
        T169-Functional Concept: 6.0
```

2.5 Sentence Extraction

Once strong chains are identified, sentences can be extracted in order to generate a summary. There are two different approaches to generating summaries from text: extractive and abstractive [17]. The *extractive* approach extracts sentences or parts of sentences verbatim from text, and is the most common way to perform summarization. Although this is a simple method, there are problems with performing extraction using this approach. The first is how to order the extracted sentences. An intuitive ordering method is to order sentences based on their order of appearance in text. A problem with any ordering method is that the resulting summary may not have a logical flow and appear as simply sentences joined together without cohesiveness. Another problem with the extractive approach is anaphora resolution. References such as pronouns make sense within the scope of the entire text, but when extracted the references are lost. For example, an extracted sentence in the summary containing the phrase “The results they provided” loses the reader when the word “they” cannot be resolved. The second and substantially more difficult approach is called *abstractive*, and involves generating summary text using natural language processing techniques. The current approach of BioChain uses the extractive approach because of its simplicity and its usefulness in evaluating chaining algorithm performance.

BioChain identifies sentences to be extracted by sorting the strongest chains into ascending order based on their score, and then identifying the top concepts within each chain. The top concepts are identified by performing a frequency count on concepts within chains. The concept with the highest frequency is considered to be the top concept. Multiple concepts having the same frequency count (a tie) are considered equal and sentences from each are extracted. Each concept, when originally chained to a semantic type, included the sentence number it originated from. The sentence number is used to extract a sentence from the text. Figure 9 shows the extracted sentence from the abstract shown in Figure 1. The strongest chain is “T081 – Quantitative Concept” and the strongest concept in the chain is “Median Statistical Measurement.”

Figure 9: Extracted sentences

T081-Quantitative Concept
Concept: Median Statistical Measurement, sentence#9
Sentence: The median disease-free survival (DFS) was 48 months in the treatment group and 16 months in the control group (P = .04);

Concept: Median Statistical Measurement, sentence#10
Sentence: and the median overall survival (OS) was 75 months for treated and 46 months for untreated patients (P = .03).

3. EVALUATION

Evaluating lexical chains is difficult because it is unclear how to evaluate their quality independent of the application in which they are used [12]. The basic question is: how does one know the quality of a chain? This is a subjective question. Several authors have proposed various ways of evaluating lexical chains empirically. Silber and McCoy [4] compared the lexical chains generated over the full text and its summary. Two characteristics were measured: 1) matching word senses in both full text and summary lexical chains, and 2) matching of the lexical chains from the full text and the summary. Galley and McKeown [12] used the SEMCOR semantic concordance corpus [18] to evaluate noun sense disambiguation accuracy. Other approaches evaluate the performance of applications using lexical chains. The idea is that lexical chains are the key contributor to system performance. Barzilay and Elhadad [16] evaluated the generated document summary output generated by their system based on lexical chains. [13] compared the topic detection ability of their system on pairs of generated summary texts.

To evaluate BioChain, two types of evaluation were performed. The first compares the performance of BioChain summarization against existing summarization systems. The second compares the abstract against the full text and defines measures of precision and recall.

In addition to a quantitative evaluation, we used a domain expert to evaluate the quality of the extracted sentences in order to produce a qualitative evaluation. We have reviewed the work with a domain expert and initial results indicate the summaries are useful, but more detail is needed. The domain expert identified the concepts most important to the generation of clinical summaries. For example, qualitative concepts are excluded because only the results of the clinical trial are desired, not the author interpretations. A filter was added to set the score of all chains not belonging to the key clinical trial concepts to zero. A zero score prevents any sentences from being extracted from the ignored concepts. Figure 10 shows the UMLS concepts used to extract sentences. We are continuously working with our domain expert to further refine the output of BioChain.

Figure 10: Concepts used to identify sentences for extraction

Concept ID	Concept 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

We also considered measuring the chain effectiveness using an extrinsic approach, as discussed in [13]. That is, we would measure the effectiveness of the generated summary and conclude that the summarization score measures the concept chain effectiveness. The assumption is the more effective a concept chain, the better the generated summary. The Document Understanding Conferences uses a tool called ROUGE [19], which measures unigram co-occurrence as an indicator of summary effectiveness. The drawback of ROUGE we ran into is that it requires several human-generated summaries of a full-text be available to measure the generated summary. That is, the sentences in an abstract should be the same as some sentences in its full text. For this project, we had one human summary available (i.e., the abstract of full-text), and it did not include sentences from the full-text. For this reason, we were not able to use ROUGE.

3.1 Comparison to Other Summarization Systems

To provide a comparison of BioChain summarization output against existing systems, we compared BioChain against two commercial systems: Microsoft Word summarization feature [20] and

Copernic Summarizer [21], and one research system: SweSum [22]. The key metric will be the matching number of sentences extracted at compression rates (based on original document size) of 25%. The Microsoft Word summarization uses a term frequency approach, the Copernic Summarizer uses a keyphrase extraction approach [23], and SweSum uses a term frequency approach in combination with a lexical resource [22]. The idea is to get a sense how close concept-based sentence identification compares with other approaches.

Figures 11a and 11b show the results of comparing the summarization of the abstract and full-text for two documents using four systems. The Document Id column shows an internal document tracking number, the Cancer Type column shows the type of cancer discussed in the source text, and BioChain Sentence column displays how many sentences were generated by BioChain with 25% compression. The compression rate refers to how many sentences will be extracted based on the total number of sentences in the source text. Due to sentence boundary detection issues in natural language processing, not all systems agree on where a sentence begins and ends. Therefore, not all systems will generate the same number of sentences in the summary. For this evaluation, the sentences BioChain generated are compared to the sentences generated by other systems. The comparison is done using simple string matching. The number of sentences matching BioChain-produced sentences are counted. There are three columns, one for each summarization system, in Figures 11a and 11b showing matching sentence counts.

Figure 11a shows the number of matching sentences if concept filtering is not used, while figure 11b shows the number of matching sentences if concept filtering is used. We tested both scenarios since the other systems perform no domain-specific filtering. Intuitively, we expected that the unfiltered BioChain would match more closely with the other systems. This was true of the colorectal cancer paper, but not true of the sarcoma cancer paper. That is, in one paper filtering helped in finding similar sentences with other systems, while in another paper filtering reduced similarity. In general, the frequency-based approaches have the most number of sentences in common with BioChain. All summarization systems take their default parameters.

Figure 11a: Comparison of BioChain sentence output with other summarization systems using *unfiltered* chaining

			# (%) Sentences Matched with BioChain		
Document Id	Cancer Type	# BioChain Sentences After Compression	Microsoft Word 2002 (Frequency)	SweSum (Frequency)	Copernic Summarizer (Keyphrase)
1001-Abstract	Colorectal	3	2 (66%)	2 (66%)	1 (33%)
1197-Abstract	Sarcoma	3	2 (66%)	2 (66%)	3 (100%)
1001-Full-Text	Colorectal	37	7 (19%)	4 (11%)	6 (16%)
1197-Full-Text	Sarcoma	47	25 (53%)	23 (49%)	19 (40%)

Figure 11b: Comparison of BioChain sentence output with other summarization systems using *filtered* chaining

			# (%) Sentences Matched with BioChain		
Document Id	Cancer Type	# BioChain Sentences After Compression	Microsoft Word 2002 (Frequency)	SweSum (Frequency)	Copernic Summarizer (Keyphrase)
1001-Abstract	Colorectal	3	1 (33%)	1 (33%)	2 (66%)
1197-Abstract	Sarcoma	3	2 (66%)	2 (66%)	3 (100%)
1001-Full-Text	Colorectal	37	13 (35%)	7 (19%)	5 (14%)
1197-Full-Text	Sarcoma	47	16 (34%)	13 (28%)	11 (23%)

3.2 Concept Chain Evaluation

Another approach for evaluation compares concept chains generated from a main text and also from its abstract, using the general approach described by [4]. The source of data for generating concept chains is MEDLINE clinical trial papers previously analyzed by Dr. Ari Brooks and his colleagues. The abstract of each paper is treated as the summary of the main text of a paper. The general idea is that the abstract is a human summary of the full text. It is expected that the main concepts of the full text should be reflected in the main concepts of the abstract. The two metrics proposed by [4] will be used:

1) Recall: Percentage of strong chains from the full-text that have at least one concept in the summary.
2) Precision: Percentage of concept instances in the abstract that have at least one instance in the strong chains in the full-text.

The notion of *strong chains* and chain scoring was defined by [16], and is discussed in section 2.4.

Figure 12 shows the precision and recall for 24 documents from the clinical trials database. The sample papers cover four different cancer types: breast, cervical, colorectal, and sarcoma, and are filtered using the concepts shown in figure 10. Using this collection, the 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 [4] because they are in a different domain with different lexical resources, our evaluation is based on their approach. Silber/McCoy [4] indicate their lexical chaining implementation has an average recall of 0.83 and an average precision of 0.85.

The average number of strong chains is three. Since the number of semantic types in UMLS is 135, strong chains represent on average 2% of the semantic types. The average number of unique UMLS concepts in an abstract is eight, indicating that coverage of the filtered concepts shown in figure 10 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 (abstract) to the full-text.

Figure 12: Precision/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)

4. RELATED WORK

The MetaMap application has been used for applications such as: hierarchical indexing query expansion, user query categorization and data mining for clinical findings, molecular binding expressions, drug and disease relationships, and drugs and gene relationships [11]. No work has been identified which uses the ideas of chaining together Metathesaurus concepts to semantic types in order to identify text themes, or using MetaMap output to perform such chaining. Since our work is divided into two areas, chaining and summarization, the previous work in these two areas is discussed.

4.1 Lexical Chaining

Lexical chaining has been used for many years for text summarization. Lexical chaining is a method for determining lexical cohesion among terms in a text [16]. Lexical cohesion is a property of text that causes a discourse segment to “hang together” as a unit [14]. 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 [14]. Lexical chaining is useful for determining the *aboutness* of a discourse segment, without fully understanding the discourse. As a basic assumption, the text must explicitly contain semantically related terms identifying the main concept. For example, if a text is about a political candidate and does not contain terms signifying the person is a candidate, lexical chaining cannot identify the fact the person is a political candidate.

Lexical chains for text summarization were first introduced by Morris and Hirst [14]. 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 Barzilay and Elhadad [16] showed the theoretical work by Morris/Hirst could be practically realized for document summarization. While Barzilay/Elhadad proved the feasibility of computing lexical chains, their implementation ran in exponential time, making its mainstream use unlikely. A linear time algorithm was later defined and implemented by [4]. A more recent implementation by Galley and McKeown [12] focused on improving word sense disambiguation based on the idea of one sense per discourse. All of these implementations use WordNet [5] as the knowledge source for identifying semantic relationships between terms.

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 this case, the application is document summarization. In document summarization, either single document or multiple document, the top-n-sentences approach is the most common. The summarization system identifies the sentences that most likely capture the main idea of a document or set of documents. The top sentences are then output, up to some limit either on the number of sentences or the number of characters, in order to produce a summary.

4.2 Document Summarization

There are three stages for summarizing text: topic identification, interpretation, and summary generation [24]. Topic identification assigns scores to a pre-defined unit of text, such as words, sentences, paragraphs, and so forth. The scores measure one or more features in a text. If more than one feature is utilized, the individual scores are combined to produce a single score. The text units are ranked, and the highest-scoring are extracted to produce a summary. Some features that have been used for summarization include text unit position within a document, cue phrases, frequency measures of text units, cohesivness, as well as combinations of these features. Hovy reports that performance studies of the various features show frequency measueres perform between 15%-35% precision and recall, while cohesive approaches typically range from 30%-60% [24]. Precision and recall are measured by comparing sentences extracted by a machine vs. sentences extracted by a person. BioChain uses a cohesive approach based on the methods of lexical chaining, but with the advantage of having a domain-specific resources. The second stage, interpretation, produces a summary by using natural language generation to generate a different text containing the important themes of the source text [24]. That is, text units such as sentences are not extracted. BioChain uses an extractive approach, and so does not incorporate stage 2. The final stage, summary generation, is used to plan how to generate the text of a final summary. This includes, for example, how to order the highest-ranking sentences. In BioChain, the highest-scoring sentences are output in the order they appear in the original source text.

4.2.1 Results from DUC 2004

There has been much work done in the Document Understanding Conferences (DUC) [25]. DUC provides an annual forum for researchers to extend text summarization technology. In DUC 2004, several approaches for identifying sentences for extraction were used, and we survey several of the approaches here. The News Story system uses a C5.0 to generate a decision tree to predict words in the source text that should be part of summary [26]. They use eight text features: 1) TF of word in document, 2) IDF of term in external news corpus, 3) position of word from start of document, 4) Lexical cohesion score between word and document, 5) noun flag, 6) verb flag, 7) adjective flag, and 8) noun or proper noun phrase flag. Their findings are that TF, word position and IDF have greatest impact on summary quality. In addition, they concluded that lexical cohesion adds little as a feature in decision tree classification.

The LAKE system uses a keyphrase extraction approach that is used to identify candidate sentences [27]. LAKE begins by extracting all uni-grams, bi-grams, tri-grams, and four-grams and filter them with part-of-speech patterns. A Naïve Bayes classifier trained using manual keyphrases is then used to identify relevant keyphrases. The resulting keyphrases are scored using two features: 1) Keyphrase TF*IDF, and 2) distance of keyphrase from the start of document. Their results scored in the middle of all 2004 DUC submissions. The authors feel their system can be improved by finding additional features that capture the semantic properties of keyphrases. One possibility they mention is to compute lexical chains and using keyphrase membership in a chain as a feature.

The KMS system describes a system where a text is decomposed into a parse tree format [28]. The parse tree is then used to identify noun phrases and score them based on a frequency analysis of terms in the noun phrases in addition to the occurrence of words in a DUC topic specification. Their performance fell in an acceptable range, and the authors observe that in general their frequency-based approach performs better than systems based on other approaches.

Finally, the GISTexter system uses a frequency-based method to identify sentences to extract [29]. GISTexter computes a weight for each term in a collection based on term frequency in a relevant set of documents. This weight is then used to score each sentence. The top scoring sentences are then extracted. DUC generally imposes limits on the lengths of extracted summaries. GISTexter used the following method to generate a summary:

1) Put highest ranked sentence as summary
2) If summary not long enough, then add next sentence that has more information than the summary has in common with it
3) Iterate until summary length exceeded or no more sentences
4) Perform summary compression, using several different approaches
    a. If (summary length - last sentence length) > 600 characters, remove last sentence
    b. If summary length > 665 characters, truncate at 665 characters

The system performed among the top systems. The authors found the best approach for summary compression is truncation - 4c (listed above). This approach rated the second highest score of all systems.

4.2.2 Biomedical Document Summarization

A survey of biomedical document summarization was done by [17]. The authors identify the following methods currently used to perform medical text summarization:

Extractive: Extractive approaches, as previously mentioned, take sentences from the source text and re-use them in the generated summary. There are two approaches used in this technique: statistical and graph. The statistical approach ranks each sentence and extracts the highest ranking sentence. The scoring is done in many ways, such as term frequency, keyphrase identification, and noun phrase frequency, as demonstrated by the systems in DUC 2004 (described above). The graph approach generates a tree representation of a text, and the most salient nodes in the tree are identified. The tree representation can be based on paragraph similarity, cohesion relationships between terms, and rhetorical structure relationships.

Abstractive: The abstractive approaches rely on natural language generation to summarize a text. The first abstractive approach uses a predefined template and the fields in the template are filled-in from information contained in the source text. The second approach uses a syntactical analysis of the source text to identify key components of each candidate sentence to form new sentences from existing sentences.

Multimedia: Many forms of medical communication include video, audio, and graphics presentations. These forms of communication have been analyzed to perform summarization on each type of multimedia, but the approaches are not directly usable for text processing, and are not discussed further.

Cognitive model based: Cognitive-based approaches try to simulate the methods of human summarizers to produce a summary of a source text. The authors mention a system that uses 79 agents based on over one hundred human strategies to produce a summary. The agents work in combination with a knowledge base, a domain-specific ontology, and rhetorical structure information to produce a summary.

5. FUTURE WORK

The remaining work for BioChain includes 1) sentence extraction, and 2) concept disambiguation. For sentence extraction, we are considering several additional approaches. The first is to find clusters of sentences with a semantic chain, rather than identifying the top concept and its sentences. The idea is that sentences in close proximity to one another are most representative of a particular semantic type. Another approach for identifying sentences for extraction is to identify concept relationship types within a semantic type (chain) and assign weights to each relationship type. Relationship types include synonymy, siblings, parent, and child. Each concept is scored based on its contribution to the chain, as measured by its relationship to other concepts.

Concept disambiguation can also be implemented to see if it has any effect on the generated chains. Currently it appears concepts placed in weaker semantic types are automatically rejected by chain strength. Implementing concept disambiguation will allow us to determine if this result is due to the particular data we are using.

6. SUMMARY

We proposed *concept chaining*, *BioChain*, to link semantically-related *concepts* within biomedical text together, using methods from lexical chaining. The biomedical lexical and semantic resources provided by the Unified Medical Language System (UMLS), such as metathesaurus and semantic network, as well as the text-to-concept mapping tool MetaMap Transfer, are used to implement concept chaining in the biomedical domain. The goal of BioChain is to produce a novel concept chaining methodology using UMLS resources and the ideas of lexical chaining. The results of chaining text concepts based on semantic types are then applied to biomedical text summarization, using both abstracts and full-text. Lexical chains are an indicator for finding the main concepts of a document or set of documents. By finding the main concepts, the most central sentences of a document or set of documents can be identified and extracted in order to generate a summary. Lexical chaining work to date has focused on the use of lexical chains for the general domain, where the main lexical resource is typically WordNet. The BioChain project has taken these core ideas from lexical chaining, and applied them to chain concepts based on their semantic types. The use of concepts rather than words is possible due to the availability of the UMLS MetaMap Transfer tool. MetaMap Transfer performs text-to-concept mapping. To the best of our knowledge, this is the first use of MetaMap Transfer to be used for linking generated concepts together for text summarization. However, MataMap Transfer is very slow and we plan to develop our own text-to-concept mapping methodology to replace it.

The base algorithm has been implemented and used to extract sentences from an abstract and its corresponding full text. For chaining, evaluation has determined that the top semantic types in the abstract are represented in the full text. Our future plans are to 1) implement concept disambiguation, 2) improve sentence extraction and 3) develop our own text-to-concept mapping methodology for the real-time text summarization.

7. REFERENCES

[1] United States National Library of Medicine, "PubMed," 2005.

[2] United States National Library of Medicine, "ClinicalTrials.gov," 2005.

[3] L. Reeve, H. Han and A.D. Brooks, "BioChain: Using Lexical Chaining Methods for Biomedical Text Summarization," in Proceedings of the 21st Annual ACM Symposium on Applied Computing, Bioinformatics track, *To appear*.

[4] G.H. Silber and K.F. McCoy, "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization," *Computational Linguistics,* vol. 28, pp. 487-496, 2002.

[5] C. Fellbaum, *WORDNET: An Electronic Lexical Database,* The MIT Press, 1998.

[6] United States National Library of Medicine, "Unified Medical Language System (UMLS)," 2005.

[7] SNOMED International, "SNOMED Clinical Terms," 2005.

[8] United States National Library of Medicine, "UMLS Metathesaurus Fact Sheet," 2004.

[9] United States National Library of Medicine, "UMLS Semantic Network Fact Sheet," 2004.

[10] United States National Library of Medicine, "MetaMap Transfer," 2005.

[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] 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.

[13] W. Doran, N. Stokes, J. Carthy and J. Dunnion, "Comparing Lexical Chain-based Summarisation Approaches Using an Extrinsic Evaluation," in Proceedings of the Global WordNet Conference (GWC 2004), 2004, pp. 112-117.

[14] 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.

[15] 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 CICLing-2004, 2004,

[16] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization," in In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, 1997, pp. 10-18.

[17] S.D. Afantenos, V. Karkaletsis and P. Stamatopoulos, "Summarization from Medical Documents: A Survey," *Journal of Artificial Intelligence in Medicine,* vol. 33, pp. 157-177, 2005.

[18] G. Miller, C. Leacock, R. Tengi and R.T. Bunker, "A Semantic Concordance," in Proceedings of ARPA Workshop on Human Language Technology, 1993, pp. 303-308.

[19] C. Lin, "Recall-Oriented Understudy for Gisting Evaluation (ROUGE)," vol. 2005, April 13. 2005.

[20] Microsoft Coporation, "Microsoft Word 2002," 2002.

[21] I. Copernic Technologies, "Copernic Summarizer," 2005.

[22] H. Dalianis, "SweSum - A Text Summarizer for Swedish," NADA, KTH., Stockholm, Sweden, Tech. Rep. TRITA-NA-P0015, 2000.

[23] P. Turney, "Learning algorithms for keyphrase extraction," *Information Retrieval,* vol. 2, pp. 303-336, 2000.

[24] E.H. Hovy, "Automated Text Summarization," *The Oxford Handbook of Computational Linguistics,* R. Mitkov Ed. Oxford: Oxford University Press, 2005, pp. 583-598.

[25] National Institute of Standards and Technology (NIST), "Document Undertanding Conferences," July 5, 2005.

[26] W. Doran, N. Stokes, E. Newman, J. Dunnion, J. Carthy and F. Toolan, "News Story Gisting at University College Dublin," in Proceedings of the Document Understanding Conference (DUC-2004), 2004.

[27] E. D'Avanzo, B. Magnini and A. Vallin, "Keyphrase Extraction for Summarization Purposes," in Proceedings of the 2004 Document Understanding Conference, 2004.

[28] K.C. Litkowski, "Summarization Experiments in DUC 2004," in Proceedings of the 2004 Document Understanding Conference, 2004.

[29] F. Lacatusu, A. Hickl, S. Harabagiu and L. Nezda, "Lite-GISTexter at DUC 2004," in Proceedings of the 2004 Document Understanding Conference, 2004.