Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management xxx (2007) xxx–xxx
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

The use of domain-specific concepts in biomedical text summarization

Lawrence H. Reeve [a], Hyoil Han [a,*], Ari D. Brooks [b]

[a] *Drexel University, College of Information Science and Technology, Philadelphia, PA, USA*
[b] *Drexel University, College of Medicine, Philadelphia, PA, USA*

Received 17 July 2006; received in revised form 4 January 2007; accepted 8 January 2007

Abstract

Text summarization is a method for data reduction. The use of text summarization enables users to reduce the amount of text that must be read while still assimilating the core information. The data reduction offered by text summarization is particularly useful in the biomedical domain, where physicians must continuously find clinical trial study information to incorporate into their patient treatment efforts. Such efforts are often hampered by the high-volume of publications. This paper presents two independent methods (*BioChain* and *FreqDist*) for identifying salient sentences in biomedical texts using concepts derived from domain-specific resources. Our semantic-based method (BioChain) is effective at identifying thematic sentences, while our frequency-distribution method (FreqDist) removes information redundancy. The two methods are then combined to form a hybrid method (*ChainFreq*). An evaluation of each method is performed using the ROUGE system to compare system-generated summaries against a set of manually-generated summaries. The BioChain and FreqDist methods outperform some common summarization systems, while the ChainFreq method improves upon the base approaches. Our work shows that the best performance is achieved when the two methods are combined. The paper also presents a brief physician's evaluation of three randomly-selected papers from an evaluation corpus to show that the author's abstract does not always reflect the entire contents of the full-text.

Keywords: Text summarization; Biomedicine; Concept chaining; Concept frequency

1. Introduction

Text summarization is a data reduction process. The use of text summarization allows a user to get a sense of the content of a full-text, or to know its information content, without reading all sentences within the full-text. Data reduction increases scale by (1) allowing users to find relevant full-text sources more quickly, and (2) assimilating only essential information from many texts with reduced effort. Text summarization is particularly useful in the biomedical domain, where oncologists must continuously find clinical trial

[*] Corresponding author.
E-mail addresses: lhr24@drexel.edu (L.H. Reeve), hyoil.han@acm.org (H. Han), ari.brooks@DrexelMed.edu (A.D. Brooks).

0306-4573/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2007.01.026

study information related to their specialty, evaluate the study for its strength, and then possibly incorporate the new study information into their patient treatment efforts (Brooks & Sulimanoff, 2002; Jaques, 2002). The US National Institutes of Health Clinical Trials database contains over 13,500 clinical trials (United States National Library of Medicine, 2005a), while PUBMED contains in excess of 16 million citations from over 4800 journals (United States National Library of Medicine, 2006a). Large and continuously-updated information sources such as these impede a physician's ability to improve their treatment efforts.

The contribution of this work is to present three novel extractive summarization methods (BioChain for concept chaining, FreqDist for frequency distribution, and a hybrid of the two, ChainFreq) using domain-specific concepts as a feature for identifying salient sentences in biomedical texts, and present an evaluation of them along with several other publicly-available summarizers. We show that our semantic-based summarizer (BioChainSumm) is effective at identifying thematic sentences, while our frequency-distribution summarizer (FreqDistSumm) removes information redundancy. The best performance is achieved when the two methods, BioChain and FreqDist are combined to form a third hybrid summarizer, ChainFreqSumm.

The paper is organized as follows. Section 2 provides related work and background on the need for biomedical text summarization and the methods used for concept annotation. Section 3 presents our three concept-based algorithms. Section 4 describes the evaluation methodology and Section 5 discusses the results of the evaluation and presents a physician's observations on the mismatch between abstracts and full-text of several selected papers. Section 6 provides concluding remarks and identifies future work.

2. Background and related work

2.1. Need for biomedical text summarization

Clinical trial studies and other scientific publications usually supply a summary of the paper in the form of an abstract produced by the author(s) of a study. We have identified at least five reasons for wanting to generate text summaries from a full-text source even in the presence of the author's abstract. (1) There exists no 'ideal' summary. An ideal summary is dependent on each user, including factors such as information need and domain background. An author's abstract is one view of an ideal summary, but users may want alternative summaries. (2) The abstract may be missing content from the full-text (Cohen & Hersh, 2005). (3) Customized summaries can be useful in question-answering systems where they provide personalized information. (4) The use of automatic or semi-automatic summary generation by commercial abstract services may allow them to scale the number of published texts they can evaluate. (5) The generation and evaluation of summaries allows for evaluation of sentence selection methods that may be useful in multi-document summarization.

2.2. Biomedical domain concepts and automated concept annotation

The main idea of this work is to use domain-specific (biomedical) concepts to identify important sentences within a biomedical text which can be extracted to form a summary. To achieve text summarization through the use of concepts, three resources are required: (a) a list of domain-specific concepts, (b) one or more synonymous phrases which occur in text and are associated with each domain-specific concept, and (c) an automated method for identifying concepts within a text. Concept annotation of each paper in the evaluation corpus was performed using the UMLS MetaMap Transfer application (United States National Library of Medicine, 2005b). Summary generation using the discovered concepts then takes place in two stages: (1) biomedical concept annotation of the source text, and (2) summary generation from the concept-annotated text using the discovered concepts.

In the biomedical domain, the National Library of Medicine (http://www.nlm.nih.gov/) provides resources for identifying concepts and their relationships under the framework of the Unified Medical Language System (UMLS) (United States National Library of Medicine, 2005c). UMLS contains many sub-components, but we use only three: Metathesaurus, Semantic Network, and MetaMap Transfer. The UMLS Metathesaurus, derived from over 100 different biomedical vocabulary sources, contains concepts and real-world instances of the concepts, including a concept name and its synonyms (United States National Library of Medicine, 2006b). The UMLS Semantic Network organizes the Metathesaurus concepts into semantic types (United States National Library of Medicine, 2004).

For automated concept annotation, the MetaMap Transfer (MMTx) (United States National Library of Medicine, 2005b) application maps free-form biomedical text to Metathesaurus concepts. MMTx performs text-to-concept mapping by first identifying noun phrases in each sentence, generating term variants of the phrase, finding candidate concepts from the generated phrase variants, and scoring each candidate concept. The highest scoring concept and its semantic type are then returned. It is possible for a noun phrase to map to more than one concept. In this case, no disambiguation is performed, and MMTx returns multiple concepts and their semantic types.

2.3. Related work

We mainly investigate previous text summarization work related to lexical chaining and frequency methodologies because our work largely involves these two approaches.

Lexical chaining is a method for determining lexical cohesion among terms in a text (Barzilay & Elhadad, 1997), and has been used for many years for text summarization. Lexical cohesion is a property of text that causes a discourse segment to "hang together" as a unit (Morris & Hirst, 1991). 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 (Morris & Hirst, 1991). 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 for text summarization were first introduced by Morris and Hirst (1991). 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 (1997) 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 algorithm runs in exponential time. A linear time algorithm was later defined and implemented by Silber and McCoy (2002). A more recent implementation focuses on improving word sense disambiguation based on the idea of one sense per discourse (Galley & McKeown, 2003). All of these implementations use WordNet (Fellbaum, 1998) as the knowledge source for identifying semantic relationships between terms. A computational model for semantic relationships between terms was developed by Fellbaum (1998). To our knowledge, no biomedical text summarization used lexical chaining with UMLS.

Term frequency was first used in extractive text summarization in the late 1950s (Luhn, 1958). A follow-up study of an analysis of five term frequency methods showed high agreement in sentence selection among the methods (Rath, Resnick, & Savage, 1961). Subsequent research using frequency methods focused on the use of frequency as one feature among many for identifying important sentences, such as cue phrases (Edmundson, 1999; Pollock & Zamora, 1975). Summarization using larger units of text has also been researched. The LAKE system uses keyphrases for summarization (D'Avanzo, Magnini, & Vallin, 2004). The SUMMARIST system (Hovy & Lin, 1999) uses WordNet (Fellbaum, 1998) concept counting not for identifying salient sentences, but for generalizing concepts for topic interpretation (e.g., {pear, apple} → fruit). Most recently, the SumBasic algorithm uses term frequency as part of a context-sensitive approach to identifying important sentences while reducing information redundancy (Nenkova & Vanderwende, 2005). The use of frequency as a feature in locating important areas of a text has been proven useful in the literature (Edmundson, 1999; Luhn, 1958; Pollock & Zamora, 1975; Rath et al., 1961). This is most likely due to reiteration, where authors state important information in several different ways, in order to reinforce main points (Sparck Jones, 1999).

3. Summarization methods

In extractive text summarization, the task is to identify sentences in a source text which are relevant to the user while simultaneously reducing information redundancy. Sentences are scored based on a set of features.

The top-*n* highest scoring sentences in a text are then extracted, using *n* as an upper bound, and presented to the user in their order of appearance in the original source text.

3.1. Concept chaining (BioChain)

The BioChain method (Reeve, Han, & Brooks, 2006) applies the concepts and methods of lexical chaining to biomedical text using concepts rather than terms. Previous lexical chaining approaches use linkages among word instances to identify semantically-related terms, and the resulting linkages are used to identify the themes of a text. In the BioChain method, automatically identified concepts in the source text (see Section 2.2) are chained based on their UMLS semantic type(s). Each concept chain contains a list of concepts belonging to a particular UMLS semantic type. If a concept belongs to multiple UMLS semantic types (i.e., multiple concept chains), the concept appears in multiple concept chains. Fig. 1 shows the concept-chaining summarization process. The source text is first processed by the UMLS Metamap Transfer application to identify noun phrases, which are then automatically mapped into UMLS concepts as well as UMLS semantic types. The summarizer (BioChainSumm) then takes all discovered concepts and constructs chains of concepts which are related by their UMLS semantic type.

Once all concepts in the text have been chained, the strongest concept chains are identified through scoring. The original lexical chaining paper defines three strong chain features: reiteration, density, and length (Morris & Hirst, 1991). Reiteration is repetition of concepts throughout a text. Density suggests concepts closer together are more likely to be related. Length is the number of concept instances within a concept chain. We score concept chains by multiplying the frequency of the most frequent concept in the concept chain by the number of distinct concepts in the concept chain (Barzilay & Elhadad, 1997; Doran, Stokes, Dunnion, & Carthy, 2004). Once all concept chains are scored, the strongest concept chains are identified. Lexical chaining uses two standard deviations above the mean of all chain scores to identify strong chains (Barzilay & Elhadad, 1997), and we follow that method. Strong concepts within each strong concept chain are then identified using two different methods: (1) using the most frequent concept within each concept chain (ties result in mul-

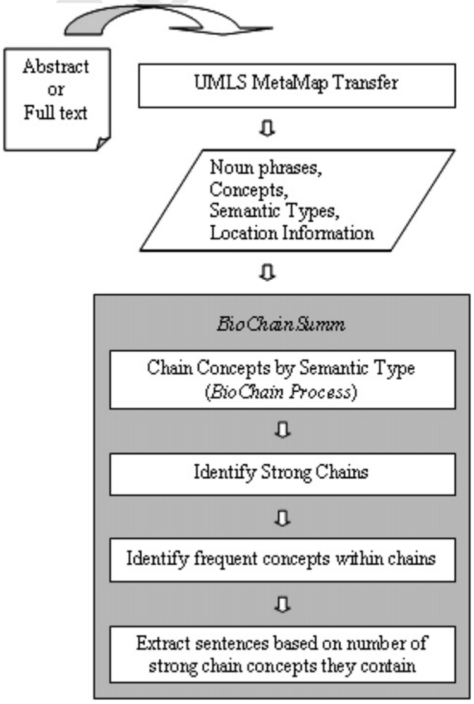

Fig. 1. BioChainSumm summarization process.

Please cite this article in press as: Reeve, L. H. et al., The use of domain-specific concepts in biomedical ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.026

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

Fig. 2. A sample concept chain for UMLS Semantic Type #81, quantitative concept.

tiple strong concepts for a concept chain) (called MostFrequentStrongChainConcept), and (2) using all concepts within a concept chain (called AllStrongChainConcepts). Sentences are scored based on the number of strong concepts they contain. A subset of the original source text sentences is then extracted based on their score and then presented in their original presentation order to form a summary. The size of the generated summary is user-controlled.

Fig. 2 shows an example concept chain which links together concepts found within a biomedical source text. In this example, the concept chain is for UMLS semantic type #81, which has a description of *Quantitative Concept*. The list of concept chain member phrases are phrases automatically found by the MetaMap Transfer application. Each phrase also shows the UMLS concept which maps to it, as well as the sentence and section (paragraph) in which it occurs. This information is also automatically generated by the MetaMap Transfer application.

3.2. Frequency distribution (FreqDist)

When using frequency as the only feature for identifying salient sentences, unit items (e.g., word, concept, and phrase) are counted and then each sentence is given a score based on the frequency count of each unit item in the sentence. A key issue in generating summaries is reducing redundancy. Each new sentence in the summary should add new information. Using the highest frequency terms will likely result in the same information repeatedly being selected. In the SumBasic context sensitive approach (Nenkova & Vanderwende, 2005), a term probability is first determined, and then each term's probability is reduced as it is used to select sentences. This is also related to the idea of Maximal Marginal Relevance (MMR), where marginal relevance is finding relevant sentences which contain minimal similarity to previously selected sentences (Carbonell & Goldstein, 1998). We use a context sensitive approach to scoring sentences based on a frequency distribution model rather than term probability. The rationale of our approach is that the frequency distribution of terms or concepts in the source text and the generated summary should be as similar as possible.

Our FreqDist method (Reeve, Han, Nagori et al., 2006) uses a frequency distribution approach with two stages: Initialization and Summary Generation. In the Initialization stage, the unit items (terms, concepts, etc.) of the source text are counted to form a frequency distribution model of the text, and a pool of sentences from the source text is created. A summary frequency distribution model is then created from the unit items found in the source text, and their frequency counts are initialized to zero. In the Summary Generation stage, new sentences are selected to be added to the summary. Identifying the next sentence to be added to the summary is accomplished by finding the sentence which most closely aligns the frequency distribution of the summary to

Please cite this article in press as: Reeve, L. H. et al., The use of domain-specific concepts in biomedical ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.026

```
Initialization:
// Note: '-model' means  'frequency distribution model'
INITIALIZE source-model to unit-items in source-text;
INITIALIZE summary-model,
             candidate-model from source-model;
            set all frequency values of both models to 0;
INITIALIZE sentence-pool to source-text sentences;

Summary Generation:
REPEAT
  INITIALIZE sentence-pool scores to 0;
  INITIALIZE best-score to 0;
  INITIALIZE best-sentence to first sentence in pool;
  INITIALIZE summary-output to empty sentence list;

  FOR each sentence-entry in sentence-pool
    INITIALIZE candidate-model from summary-model;
    ADD sentence unit-item frequencies to candidate-model;
    SET sentence-entry.score =
            similarity(source-model, candidate-model);

    IF sentence-entry.score > best-score
      SET best-score to sentence-entry.score;
      SET best-sentence to sentence-entry;
    ENDIF
  ENDFOR

  ADD unit-items from best- sentence to summary-model;
  ADD best-sentence to summary-output;
  REMOVE best-sentence from sentence-pool;
UNTIL desired summary size reached or
          sentence-pool exhausted;
RETURN summary-output as a final summary;
```

Fig. 3. FreqDistSumm summarization algorithm.

the frequency distribution of the original source text. For each sentence in the sentence pool, a candidate summary is first initialized to the summary generated so far, and then the sentence is added to the candidate summary. The candidate summary frequency distribution is then compared for similarity to the original source text frequency distribution. This similarity score (see next paragraph) is assigned to the sentence. That is, the similarity calculation step is applied to each sentence in the sentence pool one-by-one. After all sentences from the sentence pool have been evaluated for their contribution to the candidate summary, the highest scoring sentence is added to the summary and removed from the sentence pool. This process of identifying the next sentence to be added to the summary is iterative, and repeats until the desired length of the summary is reached. Fig. 3 shows the complete FreqDist method implemented in the FreqDistSumm summarizer.

We compared five similarity functions to find which type of function worked best to evaluate a candidate summary's frequency distribution to the original source text frequency distribution (Reeve, Han, Nagori et al., 2006). Each frequency distribution (candidate summary and original source text) is modeled as a vector of unit items. Similarity functions are then applied to the two vectors. The five similarity functions used are: (1) Cosine similarity (Baeza-Yates & Ribeiro-Neto, 1999); (2) Dice's coefficient (Dice, 1945); (3) Euclidean distance; (4) vector subtraction (Subhash, 1996); and (5) vector model comparison considering only unit item frequency (Lee, Chuang, & Seamons, 1997). We found that Dice's coefficient, which looks at the number of common terms between the two vectors, performed the best (Reeve, Han, Nagori et al., 2006).

3.3. Hybrid method (ChainFreq)

The BioChain and FreqDist methods use different approaches to identify relevant sentences for building an extractive summary. A problem not addressed in the current BioChainSumm summarizer (Reeve, Han,

Please cite this article in press as: Reeve, L. H. et al., The use of domain-specific concepts in biomedical ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.026

```
Initialization:
  INITIALIZE source-sentences to source-text sentences;
  INITIALIZE important-sentences to NULL;

Summary Generation:
  important-sentences = BioChain(source-sentences);

  important-sentences = FreqDist(source-text, important-sentences,
    summary size);

RETURN  important-sentences as final summary;
```

Fig. 4. Hybrid summarization method *ChainFreqSumm* using the *BioChain* method to identify sentences, and the *FreqDist* method to remove redundancy.

Nagori et al., 2006) is reducing information redundancy. Sentences containing the strongest concepts in the text are extracted without a complimentary method for reducing redundancy from sentences already selected. To overcome this limitation, we propose combining the BioChain and FreqDist methods to form a hybrid method, called ChainFreq. The hybrid ChainFreq method first uses the BioChain method to identify candidate sentences containing strong concepts. The candidate sentences (*Sc*) and their corresponding concepts (*Cc*) are then passed to the FreqDist method, which produces a set of summary sentences from the candidate sentences. That is, only *Sc* are used as a pool of sentences in the FreqDist method in Section 3.2. A summary frequency distribution model is then created from the *Cc*, and their frequency counts are initialized to zero. The FreqDist method then selects sentences containing concepts in the same distribution as the original source text with respect to *Cc* which reduces redundancy to the same proportion it exists in the source text.

Fig. 4 shows how the two summarization methods, BioChain and FreqDist, are combined to form the new hybrid summarizer, ChainFreqSumm. First, all source sentences with their corresponding concept annotations are collected and passed to the BioChain method. The BioChain method takes advantage of domain-specific knowledge, specifically UMLS semantic types, to find sentences which are important in the domain. There is no limit on the number of sentences generated by the BioChain method. The subset of source-text sentences identified by the BioChain method are then passed to the FreqDist method. The FreqDist method then finds a further subset of sentences whose concept distribution best aligns with the concept distribution of the source text. A user-defined summary size limits the number of sentences output at this stage. Both the BioChain method and the FreqDist method work together to (a) find the important sentences according to the domain (using the BioChain method), and (b) reduce redundancy by further reducing the number of important sentences based on how well their concept distribution aligns with the source text's concept distribution (using the FreqDist method), which has the effect of reducing redundancy.

4. Evaluation

The purpose of the evaluation is to evaluate the usefulness of concept frequency as a singular feature for identifying salient sentences for extractive text summarization. The evaluation was done by first asking three domain experts to manually generate extractive summaries from 24 biomedical texts (see Section 4.1). A series of automated summarizers (see Section 4.5) then generated summaries of the biomedical texts. The output of each summarizer is automatically compared using an automated tool called ROUGE (Lin, 2005) (see Section 5.3). The results are given in Section 5. The rest of this section details the evaluation implementation.

4.1. Corpus

A corpus of 24 biomedical texts was generated from a citation database of oncology clinical trial papers. The database contains approximately 1200 papers physicians feel are important to the field (Brooks & Sulimanoff, 2002). Of the 1200 papers cited, 24 were randomly selected. The PDF versions of these papers were then obtained and converted to plain-text format. The papers were manually processed to remove graphics, tables, figures, captions, citation references, and the bibliography section. The resulting text was further split

Please cite this article in press as: Reeve, L. H. et al., The use of domain-specific concepts in biomedical ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.026

into an abstract text and a full-text source text (without the abstract). The number of papers chosen (24) was based on the minimum requirements of the ROUGE summary evaluation tool (Lin, 2004) as well as the resources available to complete the manual processing of each paper.

4.2. Rouge

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) tool (version 1.5.5) (Lin & Hovy, 2003) developed by the Information Science Institute at the University of Southern California was used. ROUGE is an automated tool which compares a generated summary from an automated system with one or more ideal summaries. The ideal summaries are called models. ROUGE uses N-grams to determine the overlap between a summary and the models. ROUGE was used in the 2004 and 2005 Document Understanding Conferences (DUC) (National Institute of Standards and Technology (NIST), 2005) as the evaluation tool. We used parameters from the DUC 2005 conference. ROUGE-2 and ROUGE-SU4 recall scores are used to measure each summarizer. ROUGE-2 evaluates bigram co-occurrence while ROUGE-SU4 evaluates "skip bigrams," which are pairs of words (in sentence order) having intervening word gaps no larger than four words.

4.3. Model summaries

To compare summaries generated automatically from systems, we used four models (i.e., four ideal summaries) for each of the 24 papers. The models represent different versions of ideal summaries. The first model is the abstract of the paper (author's summary). In addition, three models from three different domain experts were generated. The domain experts are medical students in their final year. Each was given the task of performing extractive text summarization by selecting 20% of the sentences within a paper which formed the best summary for that paper. Selecting a summary size was problematic. The news summarization domain typically selects a size of less than five sentences, which represents about 20% of the size of a typical news story (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999). It has been generally thought that a summary should be no shorter than 15% and no longer than 35% of the source text (Hovy, 2005).

4.4. Additional summarizers

In this evaluation, eight additional extractive summarizers were randomly selected based on the type of summarization method and availability. There are roughly four categories of summarizers selected: baseline, frequency-based, multiple feature, and redundancy-sensitive, and we select two summarizers in each category. The two baseline summarizers are Baseline-Lead, which sequentially selects the first 20% of sentences in the source text, and Baseline-Random, which randomly selects 20% of the sentences in the source text. The frequency-based summarizers are AutoSummarize in Microsoft Word (Microsoft Coporation, 2002) and Open Text Summarizer (OTS) (Rotem, 2003). AutoSummarize is a feature of the Microsoft Word (Microsoft Coporation, 2002) word processing software, and although exact details of the algorithm are not documented, online help for the product indicates sentences using frequently-used words are given a higher score than sentences containing low frequency words. OTS is an open source project where stemming can also be performed to eliminate word variations. The two summarizers using multiple features to identify sentences are SweSum (Dalianis, 2000) and MEAD (Radev et al., 2004). SweSum is a multi-lingual summarizer for Swedish and English text using features such as sentence position and numerical data identification. MEAD is a single and multi-document summarizer using features such as position of sentence within the text, overlap of each sentence with the first sentence, sentence length, and a centroid method based on a cluster of related documents. Finally, the two summarizers which reduce information redundancy are Lemur Maximal Marginal Relevance (MMR) (The Lemur Project, 2006) and SumBasic (Nenkova & Vanderwende, 2005). Lemur MMR iteratively selects sentences having a high query similarity to an automatically-generated query, and which are also maximally dissimilar to sentences already included in the summary. SumBasic uses a probability distribution of terms in the text, and reduces term probability as sentences containing the terms are selected. SumBasic was also adapted to use concepts as the input source, rather than terms.

5. Results

5.1. Summarization method performance

The ROUGE results for all summarizers is shown in Tables 1 and 2. The following sections describe the ROUGE performance of the various summarization methods.

5.1.1. BioChainSumm summarizer

For the BioChainSumm summarizer, the most frequent strong chain concept (MostFrequentStrongChainConcept) scoring method outperforms the use of all strong chain concepts (AllStrongChainConcepts) according to both ROUGE-2 and ROUGE-SU4. The use of AllStrongChainConcepts does not have the effect of finding the most important sentences, but instead a broad coverage of the text, since all concepts within a strong chain are used to score sentences. While at first the broad coverage may seem desirable for summarizing a complete text, the result is that this technique does not find the most important sentences, which is what is desired in a summary. By using the MostFrequentStrongChainConcept, the effect is to use the UMLS semantic type to find the general idea of a text, and then the main concept within the chain to further refine the idea. Both versions of BioChainSumm outperform the two baseline summarizers Baseline-Lead and Baseline-Random, as

Table 1
ROUGE-2 scores

Summarizer	ROUGE-2 score
FreqDistSumm-Term-Dice	0.12653
ChainFreqSumm-AllStrongChainConcepts-Dice	0.12216
FreqDistSumm-Concept-Dice	0.12070
SumBasic-Term	0.11673
SumBasic-Concept	0.10940
Lemur-MMR	0.10708
ChainFreqSumm-MostFrequentStrongChainConcept-Dice	0.10652
BioChainSumm-MostFrequentStrongChainConcept	0.10419
BioChainSumm-AllStrongChainConcepts	0.09708
Mead	0.09254
Baseline-Random	0.08001
MSWord	0.07977
SweSum	0.07513
OTS	0.07474
Baseline-Lead	0.07076

Table 2
ROUGE-SU4 scores

Summarizer	ROUGE-SU4 score
ChainFreqSumm-AllStrongChainConcepts-Dice	0.22303
FreqDistSumm-Term-Dice	0.22176
FreqDistSumm-Concept-Dice	0.21997
SumBasic-Term	0.21112
ChainFreqSumm-MostFrequentStrongChainConcept-Dice	0.20158
SumBasic-Concept	0.20034
Lemur-MMR	0.19874
BioChainSumm-MostFrequentStrongChainConcept	0.19173
BioChainSumm-AllStrongChainConcepts	0.18557
Mead	0.17629
Baseline-Random	0.16396
MSWord	0.15171
SweSum	0.15115
OTS	0.14919
Baseline-Lead	0.13953

Please cite this article in press as: Reeve, L. H. et al., The use of domain-specific concepts in biomedical ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.026

well as the generic summarizers AutoSummarize, OTS, SweSum, and MEAD. This indicates the use of domain-specific concepts for selecting sentence is an improvement over the use of terms which lack domain-specific knowledge. However, neither version of BioChainSumm outperforms SumBasic, Lemur (MMR), FreqDistSumm, nor the hybrid ChainFreqSumm. All of these summarizers have a redundancy-removal component as part of their design. This shows that while BioChainSumm is effective at identifying important sentences within a domain, it falls short of removing redundancy, which is an important part of text summarization.

5.1.2. FreqDistSumm summarizer

The term and concept versions of FreqDistSumm outperform all other summarizers except for the hybrid ChainFreqSumm (BioChain method plus the FreqDist method) (see Section 5.1.3). In both ROUGE-2 and ROUGE-SU4, the term version of FreqDistSumm outperforms the concept version. Interestingly, this is also true for the SumBasic summarizer. We believe this is due to the fact that concepts map multiple instances of an expression to a single concept, and in single document summarization there is not enough variation in language to allow concepts to outperform terms. For example, in two articles discussing lung cancer, one article may use 'lung cancer' repeatedly, whereas the second article uses 'pulmonary carcinoma.' Using concepts, these two instances will be merged. When using terms, the different instances will remain unique. The FreqDist method keeps redundancy in check by only allowing it to occur in the same degree as the source text. Reiteration is a technique often used by authors to emphasize important points. The FreqDistSumm summarizer attempts to mirror the source-text in a reduced form, and so the reiteration will also be expressed in the generated summary. This is a different approach than BioChainSumm, which finds all important sentences using domain-defined criteria, but does not then reduce any redundancy in the important sentences. In contrast, FreqDistSumm will find a subset of source-text sentences using domain-defined criteria, but then eliminates some of the sentences from the subset if they emphasize main points more than the source text does.

5.1.3. ChainFreqSumm summarizer

As can be seen from the Baseline-Random summarizer, randomly picking sentences performs well. This is an indication that biomedical texts contain a large amount of redundancy. As can also be seen from the BioChainSumm evaluation (see Section 5.1.1), redundancy can decrease summarizer performance. The hybrid ChainFreqSumm summarizer (BioChain method plus the FreqDist method) is an attempt to find a subset of the most important sentences using domain-specific criteria, and then remove redundancy from the subset. The ChainFreqSumm summarizer performs best when all concepts in the strong chains are used, which is the opposite of what occurs when the BioChain method is used alone. This is most likely because using all strong concepts results in a larger pool of sentences for the FreqDist method to select from. Using the ROUGE-SU4 metric, the hybrid ChainFreqSumm summarizer is the best performer, but is slightly outperformed by the FreqDistSumm term method when the ROUGE-2 metric is used. The result in combining the two approaches is that the use of concept approaches for finding salient sentences is improved over the individual methods of FreqDist and BioChain. We believe that a summarizer which (a) first identifies a subset of important sentences based on domain-specific criteria, and (b) then prunes the subset by removing redundancy leads to an effective domain-specific summarizer.

5.2. Physician evaluation

We provided a practicing oncologist our corpus of evaluation papers and asked him to select several papers and give his observations on the abstract and the full-text. He chose three papers based on their chaining performance in the original BioChainSumm evaluation (Reeve, Han, Nagori et al., 2006). The intent is to subjectively evaluate the usefulness of abstracts of a full-text in the practice of oncology treatment. The following observations were made about each paper:

Paper #1 (Riethmuller et al., 1998):

- The abstract and the full-text had different information regarding the number of patients.
- Information about the study design was not complete in the abstract, causing a misleading conclusion.

Paper #2 (Perry et al., 1987):

- The abstract gives conclusions without providing context of the results, such as follow-up time and response criteria.

Paper #3 (Thomas et al., 1998):

- The abstract makes a statement that is not made in the source text.

The physician concludes that the abstracts are a good starting point, but that they may miss critical information necessary for evaluating the results or overstate the conclusion. The physician also believes some bias is inherent in author-generated abstracts, and that additional ways of producing summaries are needed to address different needs of medical practitioners.

We followed up the physician's evaluation to see if our FreqDistSumm-generated summaries addressed the shortcomings of the abstracts. For Paper #1, complete study design information was included. In Paper #2, the summary included additional context information, such as patient eligibility and stratification and randomization. In Paper #3, the full-text did not contain the information in the abstract and so could not possibly be extracted. Based on this short evaluation, we note that automatically generated summaries addressed abstract incompleteness in two out of three cases.

6. Conclusion

We presented three novel semantic-based methods for extractive text summarization. The first method (BioChain) chains together semantically-related concepts, and then extracts sentences having concepts in the strongest of the chains. The second method (FreqDist) uses a frequency-distribution approach, where a summary is gradually constructed by adding new source sentences so that the summary and source text have similar concept frequency distributions. For single document summarization, we show that these two concept-based approaches are competitive with existing term-based approaches. In addition, we combine the two approaches (BioChain and FreqDist) to improve performance above all existing summarizers. The use of concepts can be more useful than terms for generating personalized summaries and multi-document summarization. An envisioned system allows a user to select domain-specific concepts important to the user, and then have the summarizer generate a summary where those concepts are more highly weighted than the concepts appearing in the source text.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (1st ed.). Harlow, England: Addison-Wesley.

Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the intelligent scalable text summarization workshop (ISTS'97)* (pp. 10–18). ACL, Madrid, Spain.

Brooks, A. D., & Sulimanoff, I. (2002). Evidence-based oncology project. *Surgical Oncology Clinics of North America*, 3–10.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 335–336). Melbourne, Australia. Available from http://doi.acm.org/10.1145/290941.291025.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6*(1), 57–71.

D'Avanzo, E., Magnini, B., & Vallin, A. (2004). Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004. In *Proceedings of the 2004 document understanding conference*, Boston, USA.

Dalianis, H. (2000). *SweSum – A text summarizer for Swedish No. TRITA-NA-P001(5)*. Stockholm, Sweden: NADA, KTH.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297–302.

Doran, W. P., Stokes, N. S., Dunnion, J., & Carthy, J. (2004). 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* (Vol. 2945, pp. 627–635). Seoul, South Korea.

Edmundson, H. P. (1999). New methods in automatic extracting. In I. Mani & M. T. Maybury (Eds.) (pp. 23–42). Cambridge, MA: MIT Press.

Fellbaum, C. (1998). *WORDNET: An electronic lexical database*. Cambridge, MA: The MIT Press.

Galley, M., & McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. In *Proceedings of the 18th international joint conference on artificial intelligence* (pp. 1486–1488). Acapulco, Mexico.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 121–128). Berkeley, California, United States. Availbale from http://doi.acm.org/10.1145/312624.312665.

Hovy, E. H. (2005). Automated text summarization. In R. Mitkov (Ed.), *The oxford handbook of computational linguistics* (pp. 583–598). Oxford: Oxford University Press.

Hovy, E., & Lin, C. (1999). Automated text summarization in SUMMARIST. In I. Mani & M. T. Maybury (Eds.), *Advances in automatic text summarization* (pp. 81–94). Cambridge, MA: MIT Press.

Jaques, D. P. (Ed.). (2002). *Surgical oncology clinics of North America: Prospective randomized clinical trials in oncology* (1st ed.). Philadelphia, PA, USA: W.B. Saunders Company.

Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *Software, IEEE, 14*(2), 67–75.

Lin, C. (2004). Looking for a few good metrics: Automatic summarization evaluation – How many samples are enough? In *Proceedings of the NTCIR workshop 4*, Tokyo, Japan.

Lin, C., & Hovy, E. H. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of 2003 language technology conference (HLT-NAACL 2003)* (Vol. 1(1), pp. 71–78). Edmonton, Canada.

Lin, C. (2005). Recall-oriented understudy for gisting evaluation (ROUGE). Retrieved August 20, 2005, from http://www.isi.edu/~cyl/ROUGE/.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159–165.

Microsoft Coporation. (2002). *Microsoft word 2002*. Redmond, Washington, USA.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics, 17*(1), 21–43.

National Institute of Standards and Technology (NIST). (2005). Document understanding conferences. Retrieved August 20, 2005, from http://www-nlpir.nist.gov/projects/duc/.

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

Perry, M. C., Kardinal, C. G., Korzun, A. H., Ginsberg, S. J., Raich, P. C., Holland, J. F., et al. (1987). Chemohormonal therapy in advanced carcinoma of the breast: Cancer and Leukemia Group B protocol 8081. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 5*(10), 1534–1545.

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

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., & Drabek, E. et al. (2004). MEAD – A platform for multidocument multilingual text summarization. In *Proceedings of language resources and evaluation 2004 (LREC 2004)*. Lisbon, Portugal.

Rath, G. J., Resnick, A., & Savage, R. (1961). The formation of abstracts by the selection of sentences [Electronic version]. *American Documentation, 2*(12), 139–208.

Reeve, L., Han, H., & Brooks, A. D. (2006). BioChain: Using lexical chaining methods for biomedical text summarization. In *Proceedings of the 21st annual ACM symposium on applied computing, bioinformatics track* (pp. 180–184). Dijon, France.

Reeve, L., Han, H., Nagori, S. V., Yang, J., Schwimmer, T., & Brooks, A. D. (2006). Concept frequency distribution in biomedical text summarization. In *Proceedings of the ACM 15th conference on information and knowledge management (CIKM'06)*. Arlington, VA, USA.

Riethmuller, G., Holz, E., Schlimok, G., Schmiegel, W., Raab, R., Hoffken, K., et al. (1998). Monoclonal antibody therapy for resected Dukes' C colorectal cancer: Seven-year outcome of a multicenter randomized trial. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 16*(5), 1788–1794.

Rotem, N. (2003). Open text summarizer (OTS). Retrieved July 3, 2006, from http://libots.sourceforge.net.

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

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

Subhash, S. (1996). *Applied multivariate techniques* (1st ed.). USA: John Wiley and Sons.

The Lemur Project. (2006). Lemur Language Modeling Toolkit. Retrieved July 3, 2006, from http://www.lemurproject.org/.

Thomas, G., Dembo, A., Ackerman, I., Franssen, E., Balogh, J., Fyles, A., et al. (1998). A randomized trial of standard versus partially hyperfractionated radiation with or without concurrent 5-fluorouracil in locally advanced cervical cancer. *Gynecologic oncology, 69*(2), 137–145.

United States National Library of Medicine. (2006a). PubMed. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

United States National Library of Medicine. (2006b). UMLS Metathesaurus fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html.

United States National Library of Medicine. (2005a). ClinicalTrials.gov. http://www.clinicaltrials.gov/.

United States National Library of Medicine. (2005b). MetaMap transfer. http://mmtx.nlm.nih.gov/.

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

United States National Library of Medicine. (2004). UMLS semantic network fact sheet. http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html.

Please cite this article in press as: Reeve, L. H. et al., The use of domain-specific concepts in biomedical ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.026