Concept Frequency in Biomedical Text Summarization

1st Author
1st author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
1st author's email address

2nd Author
2nd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
2nd E-mail

3rd Author
3rd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
3rd E-mail

ABSTRACT

Extractive text summarization is the process of identifying a subset of sentences within a source text document that capture the main ideas discussed in a text. Our contribution is to propose *concept frequency* as a method to identify important sentences within a text. Concepts are identified using domain-specific resources. Sentences having the most identified concepts within a text are then extracted and used to produce a summary of the original source text. The frequency counting process is adapted from existing frequency approaches, which have focused on counting other units, such as terms and phrases. Domain-specific resources are used to identify the concepts contained in the source text. An evaluation of this approach using biomedical texts is currently underway, and we will compare it to our existing work in biomedical text summarization using lexical chaining, as well as to other available summarization systems. Our goal is to provide the life science community with a means to overcome the information overload problems through the use of extractive text summarization.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language Parsing and Understanding, Text analysis.

General Terms

Algorithms.

Keywords

Text summarization, concept frequency, biomedicine.

1. INTRODUCTION

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 [3], [4]. 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, we are summarizing 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 frequency as a feature for identifying salient sentences.

The rest of the paper is organized as follows. Section 2 describes the concept frequency approach. Section 3 identifies related work. Section 4 describes our evaluation plans. Section 5 summarizes our efforts in this area.

2. APPROACH

Our domain is biomedical text, specifically oncology clinical trial result papers. The Unified Medical Language System (UMLS) Metathesaurus [5] and UMLS Semantic Network [6] are used as semantic resources. The summary generation takes place in two stages: 1) biomedical concept annotation of the source text, and 2) summary generation from the annotated text. Existing tools and methods exist for annotating biomedical texts with concepts, and for performing text summarization using terms as basic units. We adapt the summarization methods to use concepts as basic units instead. To our knowledge, combining domain-specific concept annotations with existing text summarization methods based on frequency is a novel approach to text summarization.

Concept annotation is performed using the UMLS MetaMap Transfer tool [7] to perform text-to-concept mapping. MetaMap Transfer takes a plain-text source document and processes it through a series of processing steps. The document 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 marks 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 due to ambiguity.

JCDL'06, June 11-15, 2006, Chapel Hill, North Carolina, USA.
Copyright 2006 ACM 1-58113-000-0/00/0004…$5.00.

Once the text has been annotated with biomedical concepts, the summarization stage is executed using the SumBasic algorithm [8]. SumBasic is adapted to identify candidate sentences based on the frequency of biomedical concepts within a sentence. SumBasic was chosen because it is a simple frequency approach which also incorporates a component for ensuring coverage of weaker concepts within a text. There are four steps in the algorithm. The first is to determine the probability distribution of all concepts found within a source text. This is done by computing for each concept the number of times the concept appeared in the text and dividing it by the total number of concepts found in the text. Once the probability distribution of concepts within the text is determined, subsequent steps use the concept probability distribution to iteratively identify sentences that will be placed in the summary. The second step is to score each sentence by summing the probabilities of all concepts within a sentence. The third step determines the sentence to be extracted by finding the highest-scoring sentence. The fourth step then reduces the probability of each concept appearing in future extracted sentences by multiplying each probability of each concept in the last extracted sentence by itself. This step helps to ensure the most frequent concepts do not overwhelm less frequent concepts, with the intention to provide more thorough coverage of all concepts within the text. Steps 2–4 are repeated until the desired number of sentences has been generated.

3. RELATED WORK

Frequency-based approaches have a long history in the literature. Term frequency was first used in the late 1950's [9]. Shortly after, an analysis of five term frequency methods showed high agreement in sentence selection among the methods [10]. The LAKE system uses keyphrases for summarization [11]. Our work is most closely related to the SUMMARIST system [12], which performs concept counting using concepts derived from resources such as WordNet [13].

4. EVALUATION

We have implemented the concept frequency approach in our text summarization system and are preparing biomedical clinical trial source documents for evaluation using ROUGE [14]. The input to ROUGE is the concept frequency summary and a model summary. The model summary is currently the abstract of the clinical trial document. Our plan is to execute ROUGE using the DUC2005 parameters and evaluate the ROUGE-2 and SU-4 recall scores. We are interested in comparing the concept frequency approach to the concept chaining approach used in BioChain [15], and also against general purpose summarizers, such as SweSum [16]. Our aim is to measure the effectiveness of concept frequency as a feature in text summarization.

5. CONCLUSION

We have presented a method for summarizing biomedical texts using biomedical concepts from domain-specific resources, rather than smaller linguistic units that is widely discussed in the literature. The summarization occurs in two stages: 1) concept annotation, and 2) concept frequency analysis using the concept annotation. An evaluation is currently underway to determine the effectiveness of using concept frequency for biomedical text summarization. Our initial results show that using the frequency of concepts is a promising feature for identifying important sentences within a text.

6. REFERENCES

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

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

[3] A.D. Brooks and I. Sulimanoff, "Evidence-Based Oncology Project," in *Surgical Oncology Clinics of North America,* vol. 11, 2002, pp. 3-10.

[4] D.P. Jaques Ed., *Surgical Oncology Clinics of North America: Prospective Randomized Clinical Trials in Oncology,* W.B. Saunders Company, 2002.

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

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

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

[8] A. Nenkova and L. Vanderwende, "The Impact of Frequency on Summarization," Microsoft Research., Redmond, Washington, Tech. Rep. MSR-TR-2005-101, 2005.

[9] H.P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development,* vol. 2, pp. 159-165, 1958.

[10] G.J. Rath, A. Resnick and R. Savage, "The Formation of Abstracts by the Selection of Sentences," *American Documentation,* vol. 2, pp. 139-208, 1961.

[11] E. D'Avanzo, B. Magnini and A. Vallin, "Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004," in Proceedings of the 2004 Document Understanding Conference, 2004,

[12] E. Hovy and C. Lin, "Automated Text Summarization in SUMMARIST," in *Advances in Automatic Text Summarization,* I. Mani and M.T. Maybury Eds. Cambridge, MA: MIT Press, 1999, pp. 81-94.

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

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

[15] 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, 2006.

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