Biomedical Text Summarization from Semantically-Related Concepts

Lawrence Reeve, Hyoil Han

Drexel University, College of Information Science and Technology
3141 Chestnut Street
Philadelphia, PA 19104
lhr24@drexel.edu
hhan@ischool.drexel.edu

Abstract

Users of biomedical texts are faced with an ever-increasing amount of information that must be mastered. Text summarization can be used as a data-reduction method so that users can evaluate biomedical texts for relevancy and quality. Two contributions are described: 1) proposing the use of concepts to identify key ideas within a text; and 2) merging of information retrieval and natural-language processing approaches to provide online annotation. Chaining of concepts based on their semantic-relatedness is a promising method for identifying the subject(s) of a text. Other uses of concepts, such as concept counting, are also being evaluated. The performance of such systems is critical to the realization of concept-based summarization. Concept annotation using a combination of information retrieval and natural language processing approaches may yield high-performance and reasonable accuracy. 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. The publication record of this research completed to date as well as additional details about the work can be found at http://www.pages.drexel.edu/~lhr24/.

Motivation

Physicians and biomedical researchers need to master an ever increasing body of knowledge. They have the task of finding relevant texts, and then reading existing summaries in the form of paper abstracts to determine if the information contained in the text is relevant and of good quality. Abstracts do not always contain all of the information needed to make a relevancy/quality decision, and so users are forced to examine the full source text. We propose the use of concepts from domain-specific knowledge sources to identify important information in text. In addition, existing annotation methods in the biomedical domain suffer from drawbacks in performance or provided information. We also propose to merge the two main approaches, information retrieval and natural language processing, to remove the limitations of each.

Text Summarization

Text summarization can be used to reduce the amount of information users must process. Text summarization is a data reduction technique which distills the most salient sentences within a source text into a smaller text. Users can then use the smaller text to make decisions about the quality and relevancy of the full source text. It is also possible that the summary provides enough information so the user does not have to read the entire full source text.

A key problem in text summarization is identifying what a text is about; that is, what are the main concepts the author discusses. Many approaches have been tried for machine generation of summary text. Perhaps the earliest is the use of term frequency, where the most frequent terms identify the most important parts of the text. Other approaches use keyphrases (D'Avanzo, Magnini, & Vallin, 2004), and cue phrases (Edmundson, 1999). Our research focuses on using concepts defined within a domain resource to identify important parts of a text. In particular, we propose *concept chaining* to link semantically-related *concepts* within biomedical text together, using methods from lexical chaining. In addition, other methods such as concept frequency are also being evaluated. The work done to date on concept chaining is promising (Reeve, Han, & Brooks, 2006). Although the use of concepts is generalizable to almost any domain, the knowledge sources currently used to implement the concept approach are contained in the Unified Medical Language System (UMLS) (United States National Library of Medicine, 2005b).

Concept Annotation

One way to provide meaning to biomedical documents is by creating ontologies, and then linking information within each document to specifications contained in the ontology using a markup language (Berners-Lee, Hendler, & Lassila, 2001). Ontologies are conceptualizations of a domain that typically are represented using domain vocabulary (Chandrasekaran, Josephson, & Benjamins, 1999). Semantic annotation is the process of mapping

instance data to an ontology. Annotations are what provide the link between information stored within a document and the ontology (Berners-Lee et al., 2001). Annotation can be done manually, but suffers from problems of effort required, motivation of annotators, complexity of ontologies, and similar issues (Bayerl, Lüngen, Gut, & Paul, 2003). However, building a completely automatic annotation system is an open research problem. Instead, semi-automatic systems, rather than completely automatic systems, are used because it is not yet possible to automatically identify and classify all entities within source documents with complete accuracy (Popov et al., 2003).

In the biomedical domain, the National Library of Medicine provides the UMLS Metathesaurus and Semantic Network, which can be used as the source of concepts. The Metathesaurus contains concepts and real-world instances of the concepts, including a concept name and its synonyms, lexical variants, and translations. (United States National Library of Medicine, 2004). The Semantic Network provides a categorization of all concepts in the UMLS Metathesaurus, as well as relationships between concepts in the Metathesaurus. For annotating biomedical text, two approaches utilizing the UMLS resources stand out: Metamap Transfer and IndexFinder.

The MetaMap Transfer application (United States National Library of Medicine, 2005a) implements text-to-concept mapping through a natural language processing approach. MetaMap Transfer identifies phrases, generates variants of them, scores each, and then determines a final mapping of the phrase to one or more concepts. While MetaMap Transfer is thorough in its approach, it cannot be used in online applications due to performance.

IndexFinder (Zou, Chu, Morioka, Leazer, & Kangarloo, 2003) uses an information retrieval approach, treating a document as an unordered list of terms and finding combinations of terms that can be mapped to UMLS concepts. IndexFinder uses several in-memory data structures to perform fast concept mapping. Variant generation of phrases is not performed. IndexFinder is very fast, but does not annotate the location of concepts within a text. This is fine for indexing applications, but is not useful for identifying concepts within particular text units, such as sentences, which is required for text summarization. IndexFinder tells you *what*, not *where*, concepts are in the text.

We propose to merge the NLP approach of MetaMap Transfer and the indexing performance of IndexFinder to construct an approach which will annotate segments of text for online applications. The resulting annotator will initially be targeted at biomedical texts, but will be generalizable to nearly all domains.

Conclusion

We proposed to merge the NLP and IR approaches to perform text summarization for online applications. Identifying concepts and semantically-relating them is useful for identifying important parts of texts. Work is being done to perform online semantic annotation of texts. Single document, and eventually multi-document, text summarization can be performed through the utilization of semantically-annotated texts. Once semantic annotation and text summarization are linked, practitioners and researchers alike will realize reduced information overload.

References

Bayerl, P. S., Lüngen, H., Gut, U., & Paul, K. I. (2003). Methodology for Reliable Schema Development and Evaluation of Manual Annotations. *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003),* Florida, USA.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34-43.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies and why do we need them? *IEEE Intelligent Systems, 14*(1), 20-26.

D'Avanzo, E., Magnini, B., & Vallin, A. (2004). Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004. *Proceedings of the 2004 Document Understanding Conference*, Boston, USA.

Edmundson, H. P. (1999). New Methods in Automatic Extracting. In I. Mani, & M. T. Maybury (Eds.), (pp. 23-42). Cambridge, MA: MIT Press.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). KIM - Semantic Annotation Platform. *2nd International Semantic Web Conference (ISWC2003), 2870* 834-849.

Reeve, L., Han, H., & Brooks, A. D. (2006). BioChain: Using Lexical Chaining Methods for Biomedical Text Summarization. *Proceedings of the 21st Annual ACM Symposium on Applied Computing, Bioinformatics track*, Dijon, France.

United States National Library of Medicine. (2005a). *MetaMap Transfer*.http://mmtx.nlm.nih.gov/

United States National Library of Medicine. (2005b). *Unified Medical Language System (UMLS)*.http://www.nlm.nih.gov/research/umls/

United States National Library of Medicine. (2004). *UMLS Metathesaurus Fact Sheet*.http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Zou, Q., Chu, W. W., Morioka, C., Leazer, G. H., & Kangarloo, H. (2003). IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. *Proceedings of the AMIA Annual Symposium*, 763-767.