A Semantic Free-text Summarization System Using Ontology Knowledge

Rakesh Verma
Computer Science Department
University of Houston
4800 Calhoun Rd., Houston, TX 77204
rverma@uh.edu

Ping Chen
Computer and Mathematical Sciences Department
University of Houston-Downtown
1 Main St., Houston, TX 77002
chenp@uhd.edu

Wei Lu
Biomedical Engineering Department
University of Texas-Austin
Austin, TX 78712

Abstract

As huge amounts of knowledge are created rapidly, effective information access becomes an important issue. Especially for critical domains, such as medical and financial areas, efficient retrieval of concise and relevant information is highly desired. In this paper we propose a new user query based text summarization technique that makes use of WordNet, a general knowledge source from Princeton University. Our summarization system is specially tuned to summarize medical documents by integrating Unified Medical Language System, a medical ontology knowledge source from National Library of Medicine. We participated in the Document Understanding Conference 2007 Main Task and ranked in the middle tier of 32 systems.

KEYWORDS: WordNet, UMLS, Text Summarization, Information Retrieval

1 Introduction

Information plays a key role in our society. As huge amounts of knowledge are created and available through WWW, how to efficiently and effectively distribute and access these valuable data becomes critical. A general Web search engine tries to serve as an information access agent. It retrieves and ranks information according to a user's query, and it already makes a huge impact on how we search and organize information. But current search engines only perform shallow string processing due to the lack of deep understanding of natural languages and human intelligence, and users usually have to go through pages before they find something useful or give up. It may not matter much if a user needs information about a pair of shoes, but it will be a serious problem for crucial tasks, such as in medical or financial domains.

A concise summary will improve productivity since not all documents come with an abstract or summary. Even if some documents do provide abstracts, these abstracts are written by authors to summarize the "main" ideas of an article. However, real world information retrieval often starts with a user's query, which is a set of keywords. These keywords may not match the main ideas of a document. In this case the author written abstracts will not be a good summary for a particular query. Hence, summaries that a user wants need to be generated on the fly based on his query keywords. It is impossible for static author-written abstracts to satisfy such dynamic requirements. To solve this problem we are building a general free text summarization system specially geared for the medical domain. This is the first year we participated in the Document Understanding Conference. Although the DUC 2007 corpus only includes news in general domain, our summarization system performed quite well.

1.1 Related work

Here are a few existing projects focusing on medical information retrieval, summarization, and management.

1. PERSIVAL

   PERSIVAL is designed to provide personalized access to a distributed patient care digital library. PERSIVAL supports search and summarization of online multimedia information from clinical records to both patients and healthcare providers [10]. PERSIVAL uses context to help the user formulate meaningful queries and extract important information from the clinical record. And it uses patient information to rerank articles, and uses segmentation and domain knowledge to summarize echocardiogram video.

2. HelpfulMed

   HelpfulMed [2] provides access to medical information on the Internet and in medical-related databases for professional and advanced users. Users can locate medical information by extracting noun phrases and determining relationships with other medical terminology through concept-based search support [7].

3. QCS

   QCS indexes documents, retrieves documents relevant to a query, clusters the subset of retrieved documents, and produces a single summary for each of the clusters [3]. QCS system has been evaluated with some news corpus. It is not specially designed for a medical domain, and no ontology knowledge is used in the system.

Although these systems improve the efficiency of medical information access by automatic collection and analysis of medical information, summarization and ranking through limited utilization of ontology, but they fail to take full advantage of existing rich ontology knowledge, such as Unified Medical Language System available from National Library of Medicine. Due to the huge numbers of terms and concepts used in medical domain, analysis of terms and their relationships is key to improve the medical information system performance as shown by [6].

In this paper, we discuss background information about WordNet and the medical ontology knowledge sources in Section 2. Our summarization system architecture and algorithm are presented in Section 3. We present the evaluation results from the DUC 2007 main task in Section 4. In Section 5 we conclude and discuss our future work.

2 Ontology knowledge

We use two main ontology knowledge sources in our summarization system, WordNet and UMLS. We will provide a brief overview.

2.1 WordNet

WordNet is a machine-readable lexical database for English widely used in computational linguistics community developed at Princeton University. The database consists of linked words, primarily nouns, verbs, adjectives and adverbs. These words are organized into synonym sets called synsets, and connected by three lexicon-semantic relations – hypernym, meronym and pertainym.

2.2 UMLS

Huge numbers of terms are used in medical domain. To interpret a medical document, understanding of these term and their relationships is very important. An ontology is a description of the concepts and relationships. High-quality ontology knowledge is the key to improve the quality of medical information retrieval and management. In this paper we use Unified Medical Language System (UMLS) from National Library of Medicine (NLM) as our main medical ontology knowledge base.

UMLS is designed to help a medical information system "understand" the meanings of the concepts and terms and their relationships in biomedicine and health domain [13]. The UMLS Knowledge Sources are multi-purpose, and it can be used to create, process, retrieve, integrate, and/or aggregate biomedical and health data and information. UMLS divides medical ontology knowledge into three sources: the Metathesaurus, the Semantic Network, and the SPECIALIST lexicon. SPECIALIST lexicon is designed to provide

the lexical information for the SPECIALIST Natural Language Processing System. In our current system we use the Metathesaurus and Semantic network since our focus is on semantic analysis of a medical document. Here is a brief overview about them.

The Metathesaurus is a multi-lingual vocabulary database that contains definitions of biomedical terms, their various names (such as synonyms and abbreviations), and the relationships among them.

The Semantic Network categorizes all concepts contained in the Metathesaurus into organisms, anatomical structures, biological function, chemicals, events, physical objects, and concepts or ideas. The Semantic Network also defines a set of relationships between these concepts. These relationships provide the structure for the network. The primary relationship is the "isa" link, which establishes the hierarchy of types within the Network. There is also a set of non-hierarchical relationships, such as, "physically related to", "spatially related to", "temporally related to", "functionally related to" and "conceptually related to". Here are a few examples,

- C0002871|CHD|C0002891|isa|MSH|MSH||

  Anemia, Neonatal (C0002891) has "CHILD REL" and "isa REL" to Anemia (C0002871)

- C0002871|RB|C0221016||MTH|MTH||

  Red blood cell disorder, NOS (C0221016) has "broader REL" to Anemia (C0002871)

- C0002871|RL|C0002886|mapped to|SNMI|SNMI||

  Anemia, Macrocytic (C0002886) has "like" relationship to Anemia (C0002871)

- C0002871|RO|C0002886|clinically associated with|CCPSS|CCPSS||

  Megaloblastic anemia due to folate deficiency, NOS (C0151482) has "clinically associated with" relationship to Anemia (C0002871)

We use two primary Metathesaurus relationship files, MRREL.RRF and MRCONSO.RRF. MRREL.RRF contains the "distance 1" hierarchical relationships, i.e., immediate parents, immediate child, and immediate sibling relationships, as well as other types of intra-source relationships. MRCONSO.RRF contains medical concept names, their identifiers and key characteristics.

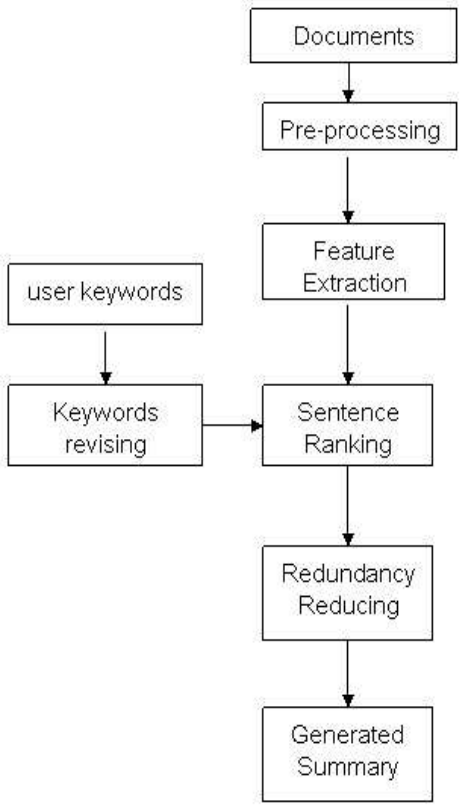

Figure 1. Free Text Summarization System Architecture

3 Summarization System Architecture

Text summarization has become an important area in text mining and generated a lot of research interest recently. There are two types of approaches [8]:

1. The knowledge-based approaches build a semantic representation for the summarization task, such as a set of logical forms [11], using ontology knowledge [5], or a template describing some key concept [1], etc.

2. The surface features-based approaches select summary material from the source based on po-

sition information, specific terms or cue phrases [4, 12].

Our summarization technique is knowledge-rich and user query-based. We represent the original document with a semantically connected concept network. We choose a subset of sentences from the original document as its summary. Our approach is totally term-based, i.e., we recognize and process only terms defined in Wordnet for general documents (UMLS for medical documents) and ignore all other words. Figure 1 shows the architecture for our summarization system. Here is the summarization procedure:

1. Revise the query with WordNet or UMLS ontology knowledge. We will add relevant keywords, delete redundant keywords. We return the revised query and let the user finalize it.

2. Calculate distance of each sentence in the document to the finalized query. Distance function used will be metrics (satisfying d(x,x) = 0, symmetry, and triangle inequality). If the distance is smaller than a threshold, the sentence will be a candidate to be included in the summary.

3. Calculate pair-wise distances among the candidate sentences (metrics can reduce the number of computations required). Then, divide candidate sentences into groups based on a threshold and select highest-ranked one from each group.

4 Evaluation Result

In this section we discuss the DUC 2007 evaluation results. This is the first year we participated in the Document Understanding Conference main task, and our system is ranked in the middle tier of overall participating systems (shown in the Table 1).

After analyzing the evaluation results on each news summary, we found the following problems with our system:

1. Insufficient redundancy reduction. Although we have performed a redundancy reduction step, it is not sufficient since the repetitive coverage of the same information from multiple documents.

2. Lack of syntax analysis. Due to the lack of syntax analysis, we rely solely on the quality of the original documents. Grammatically incorrect sentences in the original documents hurt the quality of our summary.

3. Query analysis. We need a more sophisticated way to analyze the original query.

5 Conclusion and future work

In this paper we presented our on-going work on user query-based summarization system and our experience of participating in the DUC 2007 (main task). Ontology knowledge is proven to be an effective way to go beyond the mere keyword-based information retrieval methods. With our experiment, we feel that ontology knowledge can be further utilized in other fields of broad information management and knowledge discovery process. Our future work includes:

1. Use thresholds for selecting sentences in the summary using statistical data of sentences in the abstract when it is available.

2. Utilize some natural language processing techniques in our method, such as parsing and syntax analysis.

3. Index and organize generated summaries for future access and reuse.

4. Integrate our summarization component into a broad medical information retrieval system, which may include document clustering, ranking and other components.

6 Acknowledgments

This work was supported in part by NSF grants CCF-0306475, DUE-0313880 and OCI-0453498. Wei Lu's work was performed while participating in the Summer 2006 REU program at the University of Houston under the guidance of the first author. We thank the National Library of Medicine for providing the UMLS DVD used in our system implementation.

Topics	Score	Rank
Content	2.533	19
Linguistic Quality	3.34	12
Basic Elements	0.043	24

Table 1. Average Evaluation Results

References

[1] R. Barzilay, N. Elhadad, and K. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 17:3 8, 2002

[2] H. Chen, A. Lally, B. Zhu and M. Chau, HelpfulMed: Intelligent Searching for Medical Information over the Internet, Journal of the American Society for Information Science and Technology (JASIST), pages 683-694, Volume 54, Issue 7, May 2003.

[3] D. Dunlavy, J. Conroy, and D. O'Leary, QCS: A Tool for Querying, Clustering, and Summarizing Documents, Proceedings of the HLT-NAACL Conference, Edmonton, Canada, 2003.

[4] N. Elhadad, K. McKeown. Towards generating patient specific summaries of medical articles. Workshop on Automatic Summarization, NAACL, Pittsburgh, USA, 2001

[5] M. Fiszman, T. Rindflesch, H. Kilicoglu. Abstraction Summarization for Managing the Biomedical Research Literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics. 2004.

[6] M. Fiszman, T. Rindflesch, H. Kilicoglu. Summarization of an Online Medical Encyclopedia. MedInfo. 2004 Sept.;2004: 506-510.

[7] G. Leroy, H. Chen, Meeting Medical Terminology Needs - the ontology-enhanced medical concept mapper. IEEE Transactions on Information Technology in Biomedicine 5(4): pp. 261-270, 2001.

[8] M. Maybury, I. Mani, Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, 1999

[9] Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., and Sundheim, B. 1999. The TIPSTER SUMMAC Text Summarization Evaluation. In Proceedings of the Ninth Conference on European Chapter of the Association For Computational Linguistics (Bergen, Norway, June 08 - 12, 1999). European Chapter Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 77-85.

[10] K. McKeown, S Chang, J. Cimino, S. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Johnson, D. Jordan, J. Klavans, A. Kushniruk, V. Patel, and S. Teufel. PERSIVAL: A System for Personalized Search and Summarization over Multimedia Healthcare Information. ACM/IEEE Joint Conference On Digital Libraries, Roanoke, VA, 2001, 331–340.

[11] T. Mori, M. Nozawa, and Y. Asada. Multi-Answer-Focused Multi-Document Summarization Using a Question-Answering Engine. In Proceedings of the 20th International Conference on Computational Linguistics, pp. 439–445, 2004.

[12] Y. Sun, S. Park, Generation of Non-redundant Summary Based on Sum of Similarity and Semantic Analysis, Asia Information Retrieval Symposium, Beijing, China, 2004

[13] Unified Medical Language System, available at www.nlm.nih.gov/research/umls/