

Ontology-enriched Multi-Document Summarization in Disaster Management

Lei Li, Dingding Wang, Chao Shen, Tao Li
School of Computing and Information Sciences
Florida International University
Miami, FL 33199
{lli003, dwang003, cshen001, taoli}@cs.fiu.edu

ABSTRACT

In this poster, we propose a novel document summarization approach named *Ontology-enriched Multi-Document Summarization (OMS)* for utilizing background knowledge to improve summarization results. *OMS* first maps the sentences of input documents onto an ontology, then links the given query to a specific node in the ontology, and finally extracts the summary from the sentences in the subtree rooted at the query node. By using the domain-related ontology, *OMS* can better capture the semantic relevance between the query and the sentences, and thus lead to better summarization results. As a byproduct, the final summary generated by *OMS* can be represented as a tree showing the hierarchical relationships of the extracted sentences. Evaluation results on the collection of press releases by Miami-Dade County Department of Emergency Management during Hurricane Wilma in 2005 demonstrate the efficacy of *OMS*.

Categories and Subject Descriptors: H.3.3[Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Performance

Keywords: Ontology, Multi-Document Summarization, Disaster Management

1. INTRODUCTION

Ontology is a philosophy concept, dealing with questions about what entities exist or can be said to exist, and how such entities can be grouped within a hierarchy, and subdivided according to their similarities and differences. Ontology has been applied in many research areas in information retrieval, particularly, in text mining. For example, D. Sánchez et al use the ontology to compute semantic similarity [1], and I. Yoo et al utilize the ontology to improve document clustering [2]. However, relatively few research efforts have been reported on using the ontology for improving document summarization.

Generally, given a query, multi-document summarization is the process of generating a query-focused/relevant condensation (i.e., a generated summary) of the content of the entire input set. Existing summarization methods usually rank the sentences in the documents according to their scores calculated by a set of predefined features, such as term frequency-inverse sentence frequency (TF-ISF), sentence or term position, and number of keywords [3]. The above anal-

yses are difficult to capture the hidden semantic relationships between the sentences and queries. Ontology, with abundant concise concepts and rich domain-related information, can capture the hidden semantic information.

In this poster, we develop a novel method, *OMS*, to generate query-relevant summary from a collection of documents by making use of the ontology. In particular, *OMS* first links the sentences of documents being considered onto a domain-related ontology, then maps the given query to a specific node in the ontology, and finally extracts the summary from the sentences in the subtree rooted at the corresponding query node, by using FGB, a text summarization approach described in [4]. As a byproduct, the summaries we finally acquire can be represented as a tree showing the hierarchical relationships of the extracted sentences.

We apply *OMS* to disaster management for evaluation. For natural calamities, such as hurricanes and earthquakes, vast amount of related news and reports are generated through time for broadcasting and recording events. Experimental results on such disaster management demonstrate the efficacy of *OMS*.

2. FRAMEWORK OVERVIEW

Figure 1 shows the framework of *OMS*. First of all, a domain specific ontology hierarchy was created by domain experts to describe concepts appearing in disaster related document sets. Given a collection of documents related to disasters, we disassemble them into a set of sentences. Then we map these sentences onto the ontology hierarchy based on their semantic correlations, discarding some sentences not relevant to any concept in the ontology, and omitting from the ontology some concepts with less importance.

Up to this point, we obtain an ontology-sentence tree representation in which each node is linked by a set of relevant sentences. Given a query q , *OMS* links it to a specific node i in the ontology-sentence tree according to the semantic relationship between the query and the nodes, and extracts a sub-hierarchy rooted at i . Then the FGB model [4] is applied to summarize sentences linked to each node in this subtree. Finally, a summary tree that satisfies the query q is achieved.

Ontology Refinement: In the ontology hierarchy, a myriad of concepts relevant to disaster management documents are specified by domain experts. Given a subset of documents, some concepts may not be predominant or even not appear in this documents set. We need to refine the ontology hierarchy so that the ontology-sentence tree representation can better reflect the subject of the document set.

To do so, we first rank nodes at the same level in order of sentence counts, then ignore the nodes and their subtrees with sentence counts less than the average. By iteratively running the above procedure in a top-down manner, a thematic ontology-sentence tree representation is generated.

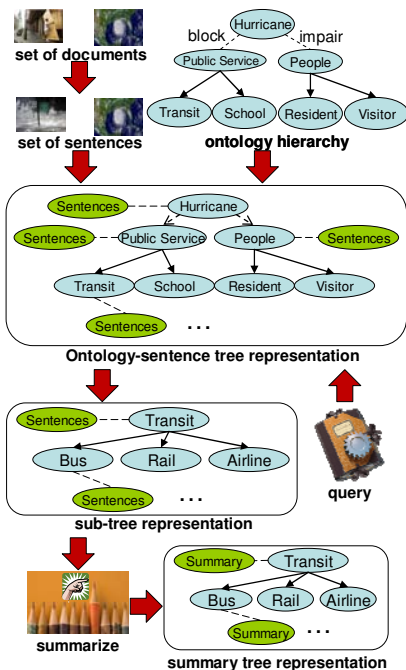


Figure 1: OMS Framework

3. EXPERIMENTS

3.1 Real World Data

The document set used in our experiments is a collection of press releases from Miami-Dade County Department of Emergency Management and Homeland Security during Hurricane Wilma from Oct. 19, 2005 to Nov. 4, 2005. It contains approximately 1,700 documents, about half of which contain similar contents. We randomly select 100 documents from this document set as our experiment data.

Data Preprocessing: For the sake of the theme embodiment of ontology-sentence tree representation, some sentences are removed from the whole sentences set in the procedure of sentence mapping, since they have no semantic relationship with any ontology concept node. For example, “For those outside of Miami Dade County can call (305) 468-5900 to reach the Answer Center.” describes the phone number of the Answer Center; however, this kind of sentences repeatedly appears in most disaster related documents we are considering, and they have no specific meaning for summarization.

3.2 An Illustrative Case Study

In order to illustrate the interpretability of our proposed method, we provide an example of queries and the corresponding result generated by OMS. Figure 2 demonstrates this case study.

Given the query “get all the information related to *transit* in Miami-Dade County after Hurricane Wilma passed”, the result is represented as a summary tree in Figure 2, in which

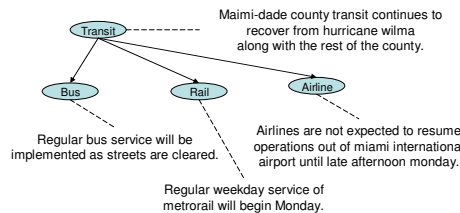


Figure 2: Summary tree related to *transit*

the topics in eclipses are domain concepts, and the sentences linked to them are summaries correlating to concepts. We observe that (1) every summary in the summary tree concretely reflects the status of relevant concept; (2) summaries generated by OMS exhibit apparent semantic hierarchy.

3.3 Performance Evaluation

For performance evaluation, 42 queries related to some specific concepts in the ontology are designed by domain experts for experimentation. In order to evaluate the quality of the generated summaries by OMS and other methods, we use human generated summaries as references. For hurricane data, we hire 5 human labelers to manually create summaries based on the selected document set and the given queries. Then we run OMS on such document set; meanwhile, we apply FGB [4] to automatically summarize query-relevant documents without using ontology. The summarization results are evaluated by ROUGE, a document summarization evaluation tool described in [5]. The experiment results are shown in Table 1. The results indicate that summarization efficacy is significantly improved by adopting ontology.

Measure	Using Ontology			Without Ontology		
	AVG-R	AVG-P	AVG-F	AVG-R	AVG-P	AVG-F
ROUGE-1	0.83236	0.75748	0.78326	0.56601	0.47562	0.50340
ROUGE-2	0.77324	0.72221	0.74223	0.43841	0.40197	0.41570
ROUGE-L	0.82032	0.75032	0.77512	0.53661	0.46032	0.48460
ROUGE-S	0.76362	0.68136	0.70562	0.43231	0.37311	0.39071
ROUGE-SU	0.77299	0.68715	0.71202	0.44875	0.38025	0.39913

Table 1: Summarization results comparison between OMS and FBG without using ontology. Remark: We use ROUGE-N, ROUGE-L, ROUGE-S and ROUGE-SU, and compare F-scores of the two different methods.

Acknowledgements

The work is partially supported by an FIU Presidential Fellowship and NSF grants IIS-0546280 and HRD-0833093.

4. REFERENCES

- [1] D. Sánchez, M. Batet, A. Valls, and K. Gibert. Ontology-driven web-based semantic similarity. *Intelligent Information Systems*, October 2009.
- [2] I. Yoo and X. Hu. Clustering large collection of biomedical literature based on ontology-enriched bipartite graph representation and mutual refinement strategy. PAKDD, 2006.
- [3] D. Jurafsky and J.H. Martin. *Speech and Language Processing*. Pearson, second edition, 2008.
- [4] D. Wang, S. Zhu, T. Li, Y. Chi, Y. Gong. Integrating Clustering and Multi-Document Summarization to Improve Document Understanding. CIKM, 2008
- [5] C. Lin. Rouge: A package for automatic evaluation of summaries. Post-Conference Workshop of ACL, 2004.