Survey on Multi-Document Summarization in Disaster Management based on Ontology

Pranjali Avinash Yadav-Deshmukh[1], **R. Ambekar**[2]

[1]ME Student, Department of Computer Engineering, Sinhagad Institute of Technology, University of Pune, Maharashtra, India

[2]Department of Computer Engineering, Sinhagad Institute of Technology, University of Pune, Maharashtra, India

Abstract: ***For semantic representation of textual information, domain ontology will provide very useful framework. In case of solving multi-document summarization problems in the domain of disaster management, the feasibility of using the ontology is tried to explore. An empirical study of different approaches is provided where for summarization tasks, ontology is used.***

Keywords: Summarization, Ontology, Disaster Management, Multidocument Summarization, CRF

1. Introduction

A natural disaster like earthquakes, hurricanes causes tremendous loss of life and property and physical destruction. Effective information gathering methods are required to analyze and study the trends of the disasters and minimize the consequent loss for future situation,

Specifically, text documents can be used to record the large number of news and reports that are related to the disaster. For analysis, condensed information about the detailed disaster event description is expected by domain experts. Condensed information like public service's operational status, the evolutionary tendency of the disaster reconstruction process of the homestead. For this, a representative scenario is provided. It contains the information which is frequently investigated by a disaster analyst .

During disaster, local government or local emergency releases more than thousands of reports. These reports contains information most events relevant to the disaster and the time span will be days to months, depending on how severe the disaster is. The information will be presented in a format of newswire having large number of routine reporting on many aspects of the disaster. As large number of information is available, so it is very difficult to immediately find either of the most vital information or the most relevant information to the specified query. For such situation multi document summarization techniques is most useful to extract the important information from large number of reports.

Domain experts provide domain ontology related to disaster management. It describes the various concepts and corresponding relation of those concepts. Such ontology possesses lots of conceptual information regarding to the document set, which may b advantageous to users to summarize the documents. A most natural question is how we can use the ontology to get high-quality summaries, i.e., representing topics with non-redundant sentences.

2. Generic Summarization

In generic summarization, each sentence is associated with a saliency score. After that the sentences are ranked according to the saliency score. According to ranking, sentences are top ranked and selected the summary based on the ranking result. To analyze the information contained in a document set and to extract highly salient sentences into the summary based on syntactic or statistical features [1]–[5], unsupervised and supervised methods are proposed.

However, most of the existing methods, the conceptual information in the sentence level is ignored. Users more readable results for summaries can be provided by the conceptual information. To address multi-document summarization [6], [7], some researchers use the explicit concepts within sentences like using Wikipedia. For domain-specific document summarization tasks such techniques cannot be directly applied as Wikipedia contains so many concepts not related to a specific domain e.g., using Wikipedia. However, such techniques cannot be directly applied to domain-specific document summarization tasks, since Wikipedia contains too many concepts not relevant to a specific domain. In our previous work [8], we explored the possibility of using domain-specific ontology for multi-document summarization; however, no detailed semantic relationship is considered.

3. Multi-Document Summarization

Multi-document summarization main objective is to create a compressed summary by retaining the main characteristics of the original set of documents. Statistics and machine learning techniques to extract sentences from documents are used by many approaches. A new multi-document summarization framework based on sentence-level semantic analysis and symmetric non-negative matrix factorization is stated in. Using semantic analysis firstly sentence-sentence similarities are calculated and the similarity matrix is constructed. To group sentences into clusters, symmetric matrix factorization, which has shown to be equivalent to normalized spectral clustering is used. Then from each group to form the summary, the most informative sentences are selected.

The major issues for multi-document summarization are as follows [9]: firstly, the information present in various documents often overlaps with each other, therefore, it is needed to find an effective way to merge the documents while recognizing and removing redundancy . In English to exclude repetition, we tend to use different word to describe the same person, the same topic as a story goes on. Thus simple word-matching types of similarity such as cosine can not faithfully capture the content similarity. Also the sparseness of words between similar concepts leads the similarity metric to become uneven. Other issue is identifying main difference between documents and covering the informative content as much as possible [10]. Current document summarization methods usually includes natural language processing and machine learning techniques [11, 12, 13], like classification, clustering, conditional random fields (CRF) , etc.

4. Query-Focused Summarization

In query-focused summarization, the information related to a given topic or query should be incorporated into summaries. The sentences suiting the users declared information need should be extracted. To incorporate the query information, various methods for generic summarization can be extended to contain query information. A robust summarization system developed within the GATE architecture is proposed in Saggionet al.[14]. It uses the robust components for semantic tagging and co-reference resolution provided by GATE.

Weiet al. [15] stated the query influence into the mutual reinforcement chain to deal with the need for query-oriented multi-document summarization. Wan et al. [16] made use of both relationships among sentences and relationships between the given query and the sentences by manifold ranking. Probability models have also been proposed with different assumptions on the generation process of the documents and the queries [17], [18].

5. Query Expansion

The process of augmenting the user’s query with additional terms in order to improve search results is Query expansion. For instance, when user is ready to search “panther" by some search engine, we can expand such query by adding synonyms of “panther” to the query, such as “jaguar,” “cougar,” etc. In the field of document summarization query expansion is very popular. Because here the quality of the generated summary can be improved. For example, Daume and Marcu [19] states a justified query expansion technique in the language modeling for IR framework. However, it does not consider the semantic relatedness between the sentences and the query string**.**

6. Disaster Management Domain

A. Domain Description

Hurricanes, earthquakes, and other natural disasters cause immense physical destruction, loss of life and property all over the world. To enhance efficient coordination and collaboration among public safety organizations by enabling the interoperable sharing of emergency alerts and incident related data between disparate systems is the main purpose of the disaster management. One of the disaster management systems aims at to analyze news and reports related to the disaster to provide concise and recapitulative information for domain experts.

B. Domain-Specific Ontology

In disaster management domain [20], ontology is many times provided by domain experts. Such ontology provides answers for the questions concerning what entities exist in disaster management, and how such entities can be related within a hierarchy and subdivided according to similarities and differences among them. The ontology described in [20] is related to the domain of hurricane management, involving 109 concepts and 326 concept relations. Ontology is gathered from the disaster management project at Florida International University [21] (http://www.bizrecovery.org). The ontology is created for the purpose of research included in this project, and is provided by the domain experts from the State Emergency Operations Center (EOC) 1 of Florida.

7. Summarization Approaches

[20]To address the summarization issues in the domain of hurricane management, we first map most sentences in the document set onto the domain ontology, and then take advantage of the intrinsic properties of the ontology to represent each sentence. In this section, we explore the effect of the ontology in multi-document summarization tasks from two directions: generic summarization and query-focused summarization.

Following are approaches for Summarization

1. Sentence Mapping
2. Sentence Representation
3. Generic Summarization
4. Query-Focused Summarization

8. Conclusion

The survey took general idea about disaster management. It studied summarization. Then survey took an account of generic summarization, multi-document summarization, and query focus summarization. Query expansion is also studied. Various summarization approaches are also discussed. Such detailed study will help for further research in area of multi-document summarization in disaster management.

References

[1] H. Hsu, C. Tsai, M. Chiang, and C. Yang, “Topic generation for web document summarization,” inProc. IEEE SMC, 2008, pp. 3702–3707.

[2] X. Yong-dong, W. Xiao-long, L. Tao, and X. Zhi-ming, “Multi-document summarization based on rhetorical structure: Sentence extraction and evaluation,” inProc. IEEE SMC, 2008, pp. 3034–3039.

[3] D. Wang, T. Li, S. Zhu, and C. Ding, “Multi-document summarization via sentence-level semantic analysis

and symmetric matrix factorization," in Proc. SIGIR, 2008, pp. 307–314.

[4] G. Erkan and D. Radev, "Lexpagerank: Prestige in multi-document text summarization," inProc. EMNLP, vol. 4, 2004, pp. 365–371.

[5] X. Wan and J. Yang, "Multi-document summarization using clusterbased link analysis," inProc. SIGIR, 2008, pp. 299–306.

[6] V. Nastase, "Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation," in Proc. EMNLP, 2008, pp. 763–772.

[7] C. Lee, Z. Jian, and L. Huang, "A fuzzy ontology and its application to news summarization," IEEE Trans. Syst., Man, Cybern., B Cybern., vol. 35, no. 5, pp. 859–880, Oct. 2005.

[8] L. Li, D. Wang, C. Shen, and T. Li, "Ontology-enriched multi-document summarization in disaster management," in Proc. SIGIR, 2011, pp. 819–820.

[9] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi-document summarization. In Proceedings of IJCAI 2007.

[10] D. Radev, E. Hovy, and K. Mckeown. Introduction to the special issue on summarization. Computational Linguistics, pages 399–408, 2002.

[11] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In Proceedings of IJCAI 2007.

[12] M. Amini and P. Gallinari. The use of unlabeled data to improve supervised learning for text summarization. In Prodeedings of SIGIR 2002.

[13] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In Prodeedings of SIGIR 2005.

[14] H. Saggion, K. Bontcheva, and H. Cunningham, "Robust generic and query-based summarisation," inProc. ECAL, 2003, pp. 235–238.

[15] F. Wei, W. Li, Q. Lu, and Y. He, "Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization," in Proc. SIGIR, 2008, pp. 283–290.

[16] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topicfocused multi-document summarization," in Proc. IJCAI, 2007, pp. 2903–2908.

[17] J. Tang, L. Yao, and D. Chen, "Multi-topic based query-oriented summarization," inProc. SDM, 2009.

[18] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in Proc. HLT-NAACL, 2009, pp. 362–370.

[19] H. Daum´ e and D. Marcu, "Bayesian query-focused summarization," in Proc. ACL, vol. 44, no. 1. 2006, p. 305.

[20] Lei Li and Tao Li," An Empirical Study of Ontology-Based Multi-Document Summarization in Disaster Management" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL. 44, NO. 2, FEBRUARY 2014

[21] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, S. Chen, and V. Hristidis, "Using data mining techniques to address critical information exchange needs in disaster affected public-private networks," inProc. SIGKDD, 2010, pp. 125–134.