Update Summarizer using MMR Approach

C Ravindranath Chowdary and P Sreenivasa Kumar
Department of Computer Science and Engineering,
IIT Madras, Chennai 600036,
India
{chowdary,psk}@cse.iitm.ac.in

Abstract: *A Huge amount of information is present on the WWW and lot is being added to it constantly. In this context, a query specific text summarization is one of the solutions to solve this problem. In this paper we apply MMR to accomplish the task of update summary generation.*

1. Introduction

A huge amount of information is being added to the World Wide Web (WWW) continuously. So, information overload has become a problem. Information Retrieval (IR) systems such as Google, Yahoo etc. address the problem of information overload by identifying documents relevant to the user's query, ranking them and presenting them as an ordered list. But the number of search results is very high and information pertaining to a query might be distributed across several sources. So it is a tedious task for a user to sift through the search results and find the information she needs. It would be very useful to have a system which could filter and aggregate information relevant to the user's query from various sources and present it as a digest or summary. This summary would help in getting an overall understanding of the query topic. The query biased summarization of general purpose articles available on web poses significant challenges like maintaining coherence, intelligibility and non-redundancy. Coherence determines the readability and information flow, while intelligibility/responsiveness is the property that determines if the summary satisfies user's needs or not.

2. Related Work

Summarization can be classified as abstractive and extractive. We focus on extraction based query specific summarization approach. Extraction based approaches use a scoring function to score each sentence in the document set. Several clustering based approaches

[3] were tried where similar sentences are clustered and a representative sentence of each cluster is chosen as a digest. MEAD [3] is a centroid based multi-document summarizer. It uses features like cluster centroids, position etc., to summarize documents. The documents are clustered together a priori by a topic detection system. Some other machine learning approaches other than clustering have also been tried out in [4, 5]. Recently, graph based models are being used to represent text. They use measures like degree centrality [6] and eigen vector centrality [7] to rank sentences. Most of these methods are inspired by PageRank. Highly ranked sentences are selected into the summary.

3. Proposed Model

We have submitted one run for update summary generation task. Our Id is 7. The task is to generate summaries on set A and set B for the given query. But the assumption for summarizing set B is that the user has already read the documents in set A.

Strategy adopted for summarizing set A: We have summarized the set A using the query specific summarizer QueSTS[1], that is developed by us earlier. But this system is not designed to generate summaries of fixed length. So, the quality of summaries suffered a loss when the summaries generated were truncated to 100 words.

Strategy adopted for summarizing set B: While generating summaries for set B, summary generated on set A is available with us. We followed MMR [2] approach to summarize set B. Each sentence in documents of set B is assigned scores based on the node score mechanism proposed in [1]. Now the score of each sentence is recomputed using MMR approach. The highest scored sentence is included into the summary. The former process is repeated till the summary of desired length is reached. Equation 1 is the MMR equation. λ is taken as 0.6. NodeScore() and Sim(,) are described in [1].

$$\lambda \ \text{NodeScore}(ni) - (1-\lambda)\text{Max}_j\{sim(ni,sj)\} \quad (1)$$

Here, ni is a sentence from set B and sj is the sentence from summary of set A. So, from Equation 1 it is clear that the sentence which contributes maximum will get higher score.

4. Conclusions

We initially had an intuition that by using MMR approach we can filter that information which is already present in the summaries that were generated on set A and then select that set of sentences from set B that are both informative and non repetitive. But MMR approach alone is not giving satisfactory results.

References

[1] M. Sravanthi, C. R. Chowdary, and P. Sreenivasa Kumar. "QueSTS: A Query Specific Text Summarization System" *In Proceedings of the 21st International FLAIRS Conference*, pages 219 - 224, Florida, USA, May 2008. AAAI Press.

[2] Carbonell, J.G., Goldstein, J. "The use of MMR, diversity-based reranking for reordering documents and producing summaries". *In SIGIR*, pages 335-336, Melbourne, Australia 1998, ACM.

[3] D R Radev, H Jing and M Budzikowska. "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies". *In NAACL-ANLP 2000 Workshop on Automatic Summarization*, 21–30. Morristown, NJ, USA: Association for Computational Linguistics.

[4] Chuang, W. T., and Yang, J. 2000. "Extracting sentence segments for text summarization: a machine learning approach". In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 152–159. New York, NY, USA: ACM Press.

[5] Fisher, S.; Roark, B.; Yang, J.; and Hersh, B. 2005. "OGI/OHSU Baseline Query-directed Multidocument Summarization System for DUC-2005". *Proceedings of the Document Understanding Conference (DUC)*.

[6] Salton, G.; Singhal, A.; Mitra, M.; and Buckley, C. 1997. "Automatic text structuring and summarization". *Inf. Process. Manage.* 33(2):193–207.

[7] Mihalcea, R., and Tarau, P. "TextRank: Bringing order into texts". *Proceedings of EMNLP 2004*, 404–411. Barcelona, Spain: Association for Computational Linguistics.