A Query-Based Summarizer based on the Context

Divya Vidyadharan[1], Anju CR[2]

[1]Department of Computer Science and Engineering, KMCT College of Engineering, Kozhikode, India

Abstract: ***Summarization condenses a document or multiple documents into a smaller version by preserving the information content and its meaning. It is very difficult to summarize large documents manually. Many summarizers have been developed to capture the content of the document. Existing models use similarity between the sentences to extract the most salient features. The techniques do not depend on the context of the document. A context sensitive model based on association of the terms is introduced. The resulting index weights are used to calculate sentence similarity.***

Keywords: query summarizer, lexical association, term frequency, sentence similarity

1. Introduction

Document summarization is an information retrieval task which aims at extracting a condensed version of the original document. Readers will decide whether to read a complete document only after going through the summary. A summary of a document or a set of documents will give the overview of the content that is present in it. Even scientific people read a document only after reading the summary. Summarization has become an important tool.

The summary will provide the main details of the document. The main goal of a summary is to present the main idea in a document or a set of documents in a short and readable paragraph. Summaries can be produced either from a single document or many documents. The summaries produced from multiple documents are called multi document summarizer.

There are various techniques of summarization. Some of them rely on centroid based techniques, semantic analysis. Some of the techniques will be biased to a particular topic, as in the case of query based summarization. These techniques actually retrieve documents related to a particular query.

Topic summarization deals with the evolution of topics in addition to providing the informative sentences. The major issues for multi-document summarization are as follows: first of all, the information contained in different documents often overlaps with each other, therefore, its necessary to find an effective way to merge the documents while recognizing and removing redundancy. Another issue is identifying important difference between documents and covering the informative content as much as possible.

2. Related work

There are various works till date that summarizes the documents. Text summarization can be either “abstractive” or “extractive”. The abstraction- based method provide summary by sentence compression and reformulation. It allows the summarizers to increase the information without increase in summary length. Such models require complex linguistic methods. The extraction – based models, use various statistical features to identify to identify those sentences that convey the meaning of the document.

D.R. Radev, H. Jing, M. Styas, and D. Tam [1] proposed a summarizer called MEAD, which generates summaries using cluster centroids produced by topic detection and tracking system. Centroid based summarization model that used tf-idf score to identify the centroid It uses techniques based on sentence utility and subsumption, which we have applied to the evaluation of both single and multiple document summaries. A key feature of MEAD is its use of cluster centroids, which consist of words which are central not only to one article in a cluster, but to all the articles. The problem of this summarizer is that the summary will mostly consist of sentences that are in the centroid of the article.

Many of them used a combination of statistical and linguistic methods such as term frequency, sentence position, topic signature, lexical chains for computing the saliency score of the sentences. Summaries are also produced by computing the semantic similarity of the sentences. D. Wang, T. Li, S. Zhu, and C. Ding [2] focus on the similarity between the sentences. The framework is based on sentence level semantic analysis and symmetric non-negative matrix factorization.

BAYESUM [3] makes use of sentence extraction in query-focused summarization. BAYESUM leverages the common case in which multiple documents are relevant to a single query. Using these documents as reinforcement for query terms. The key requirement of BAYESUM is that multiple relevant documents are known for the query in question. BAYESUM is built on the concept of language models for information retrieval. A sentence appears in a document because it is relevant to some query, because it provides background information about the document. The model assumes that each word can be assigned a discrete, exact source.

SUMMARIST [4] creates a robust automated text summarization system, based on the equation: summarization=topic identification+ interpretation + generation. The task is to produce synopsis of any document(s) submitted to it. The input is selected and filtered to determine the most important, central, topics. The topic identification can done by methods based on position, cue phrases, word frequency and discourse segmentation. A collection of extracted concepts are fused into their one (or more) higher level unifying concept(s). This is the most difficult step of automated text summarization.

3. Problem definition

The main goal of this system is to provide the summary which is not a mere set of sentences but is based on the context. The summary that is produced by the summarizers till date continues a set of sentences that is picked from the document. This does not provide the context of the document which is a major disadvantage. A document when read by a human will understand the whole content of it and then summarize accordingly. It will clearly depict the content of the document. It is difficult to make a computer understand the content to create a summary. All the previous method do not depend on the context in the summary. A set of sentences picked randomly from the document will not be a good summary. All the previous method can be put into a single category i.e. context independent document indexing.

4. System Overview

The overview of the system is shown in figure 1. In the proposed system, the input is the URL of the web page which needs to be summarized. The contents of the web page are extracted to a text file before the remaining processes continue.

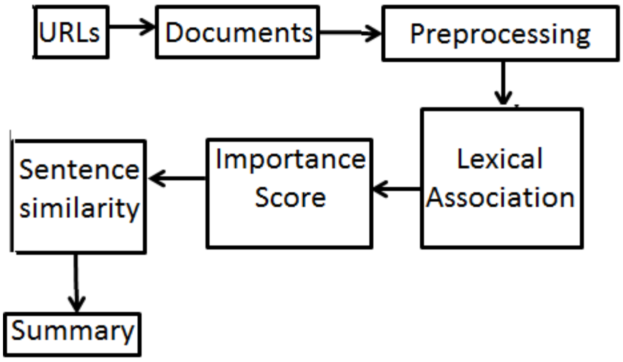

Figure 1: System overview

4.1 Text Extraction

The query for this system is the URL of a web page. The content of the page is extracted using an html parser and stored to a text file. The source code of the web page is obtained which contains html tags. The html tags of the page are removed before further processing. Further steps of summarization are done onto this text file to get the summary.

4.2 Pre-processing

Pre-processing is one of the important task done in summarizing. The input document will contain various stop words which will occur in abundance. They do not have any significant meaning as it does not define the relation of the document. There will also be different forms of the same words. They are removed so that it won't be a problem during indexing them. Removal of stop words, stemming, noun and verbs finding, etc. are some of the techniques that are done.

4.3 Lexical association

Lexical association shows how a term in the document is associated tor related to the document. The term frequency and inverse document frequency are used to find the term association. The number of times a term occurred in a sentence is called as term frequency. It is represented as "tf". The number of times a term occurred in the whole document is called as 'Document frequency'. The logarithmic inverse of it will give the inverse document frequency. Together the if-idf will give the weight of the words. Tf-idf will help to recognize which words are more associated to the document.

4.4 Importance score

The individual weights of the words are calculated to find out the most important sentence. The sum of the weight of all the terms in a sentence will give the score of the sentence.

4.5 Sentence Similarity

To include the sentences in the summary, the similarity among the sentences are calculated. One of the major similarity techniques, cosine similarity is used for this purpose. The sentences with high score are chosen to calculate the similarity of the sentence.

5. Evaluation and results

The system is evaluated on about 140 webpages available online. These 140 pages included topics on same themes and also different themes.

5.1 ROUGE evaluation

Summaries are evaluated using ROUGE [5].which is the most widely used toolkit for the evaluation of the system generated summaries. ROUGE is a recall-oriented summary evaluation metric that is widely used for the evaluation of summarization techniques. It measures the summarization performance by calculating the number of overlapping n-grams between an evaluated summary and a set of reference summaries. There will be a set of reference summaries: which is human generated summaries. It has been shown that the results of comparisons based on ROUGE-1 and ROUGE-2 (i.e., unigram- and bigram-overlap) . Therefore, we use ROUGE-1 and ROUGE-2 to evaluate the consistency of manual summaries derived by the compared methods.

The Table 1 shows the evaluation results of our system. The average score for all the topics tested with our system is displayed in the system with ROUGE.

Table 1 The average recall, average precision and average F-score generated by ROUGE package for the system.

	Avg_R	*Avg_P*	*Avg_F*
Rouge-1	0.4752	0.85644	0.61125
Rouge-2	0.43911	0.76289	0.45963

5.2 Performance evaluation

For the evaluation of the performance, the execution time of the system is calculated for different size of the input. The execution time of the system increases as the size of the input increase. Figure 2 shows the increase in the execution time against the different input.

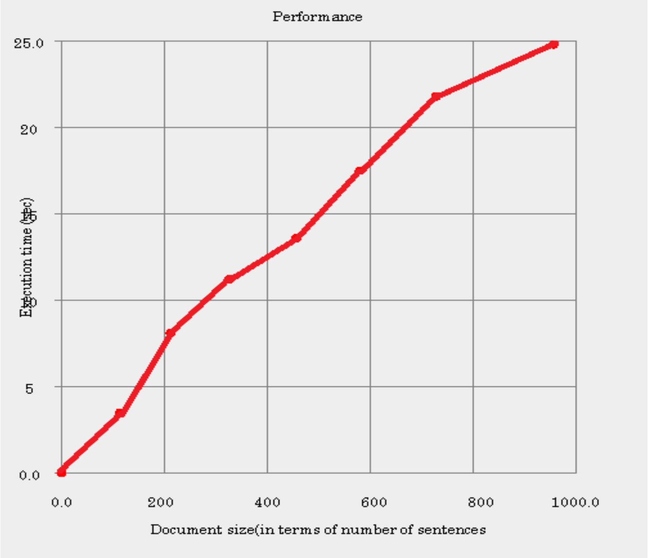

Figure 2: Performance analysis

6. Conclusion and future scope

The system summarizers web pages whose URLs are given to it. It makes use of term frequency and sentence score to find out the sentences which are associated with the document. Further improvement would be to provide summary based on the topic given to it. Also, the system is evaluated on a very small dataset. To use this system on a very large dataset will require additional improvements to be done.

References

[1] Radev, D.R., Jing, H., & Budzikowska, M. (2000). Centroid - based summarization of multiple documents: sentence extraction, utility based evaluation, and user studies

[2] D. Wang, T. Li, S. Zhu, and C. Ding, Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization", Proc.31st Ann.Intl ACM SIGIR Conf. Research in Information Retrieval, pp. 307- 314, http://doi.acm.org/10.1145/1390334.1390387, 2008.

[3] H. Daume III and D. Marcu, Bayesian Query-Focused Summa-rization, Proc. 21st Intl Conf. Computational Linguistics, pp. 305-312, http://dx.doi.org/10.3115/1220175.1220214, 2006.

[4] E. Hovy and C.-Y. Lin, Automated Text Summarization and the Summarist System, Proc. Workshop Held at Baltimore, Maryland (TIPSTER 98), pp. 197-214, http://dx.doi.org/10.3115/1119089. 1119121, 1998.

[5] Chin-Yew Lin. (2004) "ROUGE : A Package for Automatic Evaluation of Summaries", In theProceedings of the Workshop on the Text Summarization Branches Out (WAS 2004), Barcelona, Spain: ACL.

[6] P. Goyal, L. Behera, and T. McGinnity, "Query Representation Through Lexical Assoc. for Information Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 24, no. 12, pp. 2260-2273, Dec. 2011

[7] L.L. Bando, F. Scholer, and A. Turpin, "Constructing Query- Biased Summaries: A Comparison of Human and System Generated Snippets," Proc. Third Symp. Information Interaction in Context, pp. 195-204, http://doi.acm.org/10.1145/1840784.1840813, 2010.

[8] J.G. Conrad, J.L. Leidner, F. Schilder, and R. Kondadadi, "Query- Based Opinion Summarization for Legal Blog Entries," Proc. 12th Int'l Conf. Artificial Intelligence and Law, pp. 167-176, http://doi.acm.org/10.1145/1568234.1568253, 2009.

[9] X. Wan and J. Xiao, "Exploiting Neighborhood Knowledge for

[10] Single Document Summarization and Keyphrase Extraction,"ACM Trans. Information Systems, vol. 28, pp. 8:1-8:34, http://doi.acm.org/10.1145/1740592.1740596, June 2010.

Author Profile

Divya Vidyadharan is a student pursuing MTech in Computer Science & Engineering at KMCT College of Engineering, Calicut university, Kerala. She has done BTech. in computer science & engineering from Calicut university, Kerala in 2012.

Anju CR is Assistant Professor in Computer Science & Engineering at KMCT College of Engineering, Calicut university, Kerala. She has done MTech. in computer science & engineering from KMCT College of Engineering, Calicut university, Kerala in 2014. She has done BTech. in computer science & engineering from SNS college of Technology, Anna university, Tamil Nadu in 2007.