ACHIEVING MULTI-DOCUMENT SUMMARIZATION BASED ON MULTIPLE-RANKING METHODOLOGY WITH THE HELP OF NEAREST NEIGHBORS IN CLUSTERS

K.PADMAPRIYA[1]

[1]Research Scholar, Sathyabama University, Chennai.
kpriyainbox@yahoo.co.in

Dr.S.Sridhar[2]

[2]Professor & Dean - Cognitive & Central Computing Facility, R.V College of Engineering, Bangalore.
drssridhar@yahoo.com

Abstract

The goal of multi-document summarization is to give a prejudiced summary on a particular topic. This paper describes how to perform summarization task by means of multiple-ranking of sentences from various documents. Our method utilizes the association among all the sentences in the documents and the association between the sentences and the particular query. For each sentence in the documents, ranking score has been computed to denote the richness of information in the sentences. We used greedy algorithm to force multiple range of consequences on each sentence. By selecting the sentences with highly prejudiced information richness and high information originality, the summary has been produced. We make use of both TREC and TDT documents for our experimental purpose and our results show that the proposed methodology outperforms well when compared with the existing methodologies.

Keywords: **multi-document summarization; multiple ranking; greedy algorithm**

1. Introduction

The goal of multi-document summarization is to deliver a summary with the majority of information from a set of documents on particular topic explicitly or implicitly. Given a description of a specific topic (user query), our query-based multi-document summarization should produce a summary from the documents which should answer the requirements of information stated in the query and illustrates about the query.

In recent years, automatic multi-document summarization attracts more attention in research and it shows the evidence of practicability in terms of document management and in search systems. Multi-document summary is used to provide the information stored in a cluster of documents and help the users to understand it effectively. For example, news like NewsBlaster and Google News have been developed to provide group of news articles into new topics and then provide a summary under each topic. The users can look at the summary if they have interested in that topic and it can be easily understood by them. Topic-focussed summary provides personalized services only if the user profiles are produced manually or automatically. By getting their users' interests, the above said news services will be personalized and both the news summary related to their user profile and the related news article will be delivered to the particular user.

The challenges of multi-document summarization on a given topic are – 1. The overlapping of information stored in various documents 2. The information given in the summary should be prejudiced to the given topic. Hence we require the best summarization methods to merge the information from various documents on a given topic. If it is possible the summary is able to contrast the difference among the information contained in the documents. As well as the information should be kept as novel. Now-a-days a series of workshops and conferences conducted on automatic text summarization (e.g. DUC, NTCIR) and topic sessions in ACL, SIGIR and have advanced technology that produce experimental online systems.

In this paper, we have studied multiple-ranking of sentences to query-based multi-document summarization. As a first step, we compute the multiple-ranking score for each sentence by means of applying multiple-ranking method to prejudice the information richness of the sentences. Then we use Greedy algorithm to force the consequences of sentences overlapping other sentences from various documents. The summary has

been produced by selecting the sentences with highest multi-ranking scores that are estimated with informative and novel and highly prejudiced to the given topic. In our multiple-ranking algorithm, the inter-document and the intra-document links among the sentences are differentiated with various weights. Our experimental results show that the proposed approach outperforms well than the top performing approaches used in DUC tasks and other baseline approaches.

2. Related work

Many approaches have been developed for multi-document summarization. But they are either abstractive summarization or extractive summarization. In abstractive summarization, they require combination of information, sentence compression and process of reframing. In extractive summarization, it assigns some scores to the sentences of the documents and mines the sentences with highest scores. In our approach, we use extractive summarization.

In [13], they used centroid-based method for extractive summarization. They called it as MED which scores sentences according to its features, cluster centroids and its position. In [9], they have implemented NeATS for positioning the sentences, topic signature, term frequency and term clustering to choose important contents. In [5], they used MMR to remove redundancy. In [7], they have proposed XDoX to identify the most important themes from the document set by means of passage clustering and compose a summary that describes these main themes. In [6], they have investigated about 5 various topic representations and delivered a method on topic themes. In [2, 11, 12], they have implemented graph-based methods to rank the sentences and the paragraphs alike PageRank and HITS to compute the importance of the sentences.

Most of the query-based document summarization methods integrate the information given on the query with the standard summarizers and mine the sentences mostly appropriate with users' need. In [14], a simple query based scorer has been developed for finding the resemblance between each sentence and the query is integrated into common summarizers to generate the query based summary. In [4, 1], the query and its entities are investigated for query based/ event focused multi-document summarization. In [8], the most important sentences are chosen for summarization based on basic elements' scores. In [3], thematic analysis has been done on the documents, and then compares these themes with the other document. Many research work have been published in DUC 2003 and DUC 2005.

3. Multi-document summarization based on the multiple-ranking methodology

3.1. Overview

The multiple Ranking methodology consists of two steps: 1. The multiple-ranking score has been calculated for each sentence, which shows the prejudiced information richness of the sentences. 2. The various penalty is enforced on each sentence and 3. Overall ranking score has been computed for each sentence to reproduce the prejudice information richness and the novelty of the sentence. At last, the overall ranking scores determine the sentences to be selected for summarization.

Prejudiced information richness: given a collection of sentences S= $\{s_i | 1 \le i \le n\}$ and a Query Q, the prejudiced information richness of a sentence s_i is used to mention its information degree with respect to both the sentence collection and the Query i.e. the information richness of a sentence s_i prejudiced towards Q.

Novelty of the information: Given a sentence collection in the summary R= $\{s_i | 1 \le i \le m\}$, the novelty of information in needed to compute the novelty degree of information stored in the sentence si, among the other sentences in the set R.

3.2. The multiple-ranking process

We use the baseline algorithm to rank the data points given by Zhou et al., in [17]. The ranking has been done as 1. Points stored in the same structure like cluster, may have the similar ranking scores. 2. The neighboring points may have the similar ranking scores. By having above said two assumptions, multiple ranking is defined as: a weighted network is created on the data by assigning positive rank score to all relevant points and assigning zero to the remaining points that are to be ranked. All points then share their ranking scores with its nearest neighbours thru the weighted network. This process is repeated till the global firm state is obtained and as a result all points will get their ultimate ranking scores.

Given a set of data points $S= \{s_0, s_1, ..., s_n\} \subset R^m$, where the first point s_0 is the description of the query and other n points are the sentences in the documents. Since query sentence is normally short, we call it as pseudo sentence and processed like other sentences in the documents.

Let the ranking function $f: S \rightarrow R$ assigns a ranking score value fi to each point in s_i ($0 \le i \le n$). We can treat f as a vector $f= [f_0, f_1, ..., f_n]^T$. And we have one more vector $p= [p_0, p_1, ... , p_n]^T$ in which $p_0=1$ because s_0 is the pseudosentence and $p_i=0 (1 \le i \le n)$ for all the remaining sentences in the documents.

Algorithm1: The multiple ranking algorithm

Step 1: Using the standard cosine measure, find the pair-wise similarity values between the sentences.

Step 2: Calculate the weight related with the term t using the formula $wf_t * iswf_t$, where wf_t is the frequency related with term t in the sentence and $iswf_t$ is the inverse sentence frequency of the term t, i.e., $1+\log (N/n_t)$, where N is the total number of sentences and n_t is the number of sentences having the term t. Given two sentences s_i and s_j, the cosine similarity is mentioned as sim (s_i, s_j), calculated as the inner product of the corresponding term vectors.

Step 3: if the similarity of any two points exceed 0, connect them with an edge.

Step 4: we produce a affinity matrix A by A_{ij}=sim (s_i, s_j), if an edge is linking s_i and s_j. And assume A_{ii}=0 to avoid loops in the graph.

Step 5: Symmetrically normalize A by $SN=D^{-1/2}AD^{-1/2}$ where D is the diagonal matrix (i, i) with number of elements equivalent to the sum of the i-th row of A.

Step 6: repeat f (t+1) =α SN f (t) + (1-α) p till convergence, where α is a parameter in (0, 1).

Step 7: Let f_i* mention the sequence limit {f_i (t)}. Rank every point in s_i by means of its ranking score f_i* and display it in descending order.

In the above algorithm, first a connected network is created and the network is weighted. In this algorithm, step5 is essential to prove the convergence of the algorithm. In step6, all points share their ranking score to its neighbours thru the weighted network till it reaches global firm state. Then the points are ranked. The parameter α specifies qualified contributions to the ranking score from neighbours and the initial ranking scores. Remember that self-reinforcement is evaded because the affinity matrix's elements are set to zero.

The theorem in [17] guarantees the sequence {f (t)} converges with

$$f^*=\beta (I - \alpha SN)^{-1} p \qquad (1)$$

where β=1-α. The algorithm is preferable in large scale problems due to its computational efficiency. While computing the difference between the scores in two successive iterations for any points is coming under a given threshold value, the convergence of this algorithm is achieved.

In our framework, the edges (links) between the sentences in the documents are categorized as 1. Inter document link and 2. Intra document link. If si and sj are taken from different documents, then the link is said to be inter document link whereas if si and sj are taken from the same document, then it is said to be inter document link. The links between the query and the other sentences are inter document links. But both inter document links and intra document links may have unequal contributions. Hence distinct weights are given to the intra document and inter document links respectively. In the next step, the affinity matrix will be decomposed as

$$A=A_{intra} + A_{Inter} \qquad (2)$$

Where A_{intra} is the affinity matrix contains only the intra document links and A_{inter} is the affinity matrix contains only the inter document links. The entries of both the inter-document and intra document links are set to 0. We can differentiate these two links as

$$\tilde{A}=\lambda_1 A_{intra} + \lambda_2 A_{inter} \qquad (3)$$

Let λ_1, λ_2 [0, 1]. Suppose $\lambda_1<\lambda_2$, the importance of inter document links are more than intra document links and vice versa. If $\lambda_1=\lambda_2=1$, then Eqn.3 reduces to Eqn.2. Hence $\tilde{A}$ is normalized into $\widetilde{SN}$ and the sixth step will be $f(t+1)=\alpha \widetilde{SN} f(t)+(1-\alpha)p$.

3.3. Various penalty imposition

The affinity matrix A is normalized by $\overline{SN} = D^{-1} A$ to compose the sum of each row equal to 1. According to $\overline{SN}$, the greedy algorithm is applied to enforce the various penalty and calculate the overall ranking scores, reflecting both the prejudiced information richness and novelty of the information in the sentences.

Algorithm 2: Various Penalty imposition algorithm

Step 1: initialize two sets X=∅ and Y= {s_i | i=1, 2,...,n} And the overall ranking score of each sentence is initialized as Rank_score(s_i) = f_i*, i=1,2, ... , n.

Step 2: Based on present overall ranking scores, sort the sentences of Y in descending order.

Step 3: if s_i is the maximum ranked sentence which has been sorted first in the ranked list, move sentence s_i from Y to X. Then the various penalty is enforced to the overall ranking score of each sentence linked with s_i in Y as

For each sentence sj Y,

Rank_score (s_j) =Rank_score (s_j) -$\omega \overline{SN}$ f_i*

Where ω>0 is the penalty degree factor. The greater penalty denoted as the larger ω is enforced to the overall ranking score. If ω=0, then there is no various penalty.

Step 4: Repeat the steps 3 and 4 until Y=∅.

In the above algorithm, step3 is needed to avoid the less informative sentence. After obtaining the overall ranking scores for all sentences, the sentences with highest ranking scores are selected to give the summary.

4. Experimental Evaluation

4.1 Data sets

Query based multiple document summarization can be done on task 2 and 3 of DUC 2003 and one task of DUC 2005, where each task contains standard data set have document clusters and reference summaries. We used task 2 of DUC 2003 for training and parameter tuning and the other tasks for testing. The summarization will be categorized as 1. Summaries focused by events using task 2 of DUC2003 2. Summaries focused by viewpoints using task 3 of DUC2003 and 3. Summaries focused by DUC topics using the task of DUC 2005. Since there are no substantial differences among them, we treat all representations uniformly. As a pre-processing step, we removed all the sentences given in quotation marks.

Table 1. Summary of data sets

	DUC 2003	**DUC 2003**	**DUC 2005**
Task	Task 2	Task 3	The only task
Data source	TDT	TREC	TREC
Number of Clusters	30	30	50
Summary length	100 words	100 words	250 words

4.2. Experimental results:

We used ROUGE [from 10] for evaluation purpose, which was adopted by DUC. The proposed methodology was compared with top three systems (selected from the performing systems on each task) and two baseline systems (lead baseline and coverage baseline). The lead baseline will take the sentences one by one from the last document (assume it is in chronological order) and the coverage baseline will take the sentences one by one from the first document to the last document.

Table 2. System Comparison using Task 3 of DUC 2003

System	**ROUGE-1**	**ROUGE-2**	**ROUGE-W**
Multiple Ranking	0.37332	0.07677	0.11869
Similar Rank1	0.36088	0.07229	0.11540
Similar Rank2	0.35001	0.07305	0.10969
Lead Baseline	0.34542	0.07283	0.11155
Coverage Baseline	0.31986	0.05831	0.10016
ID13	0.31809	0.04981	0.09887
ID16	0.30290	0.05968	0.09678
ID17	0.28200	0.04468	0.09077

Apart from these two baseline systems, we have implemented two more systems – SimilarRank1 and SimilarRank2. The SimilarRank1 finds the similarity between the query description and the each sentence in the documents. Then greedy algorithm is applied to enforce the various penalty on each sentence with initial overall ranking score. The sentences which have maximum overall ranking scores are selected to give the summary. SimilarRank2 ranks the sentences in the documents by their similarity value with the query description.

Table 2 and 3 explain about the system comparison results on the two tasks DUC 2003 and DUC 2005. In the tables, ID4 - ID17 are the system Ids of the top performing systems, whose details are given in DUC publications. The parameters of the multiple ranking are declared as ω=8, λ_1=0.3, λ_2=1 and $\propto$ = 0.6.

Table 3. System Comparison using DUC 2005

System	ROUGE-1	ROUGE-2	ROUGE-W
Multiple Ranking	0.38434	0.07317	0.10226
Similar Rank1	0.37396	0.06843	0.09867
Similar Rank2	0.37383	0.07243	0.09860
Lead Baseline	0.37354	0.06835	0.09948
Coverage Baseline	0.36900	0.07163	0.09752
ID4	0.35753	0.06890	0.09595
ID15	0.34569	0.05910	0.09105
ID17	0.30472	0.04765	0.08085

From the above tables, we can easily find out that our proposed method outperforms well when compared with other approaches. The main reasons lead to this achievement is

- Multiple ranking process – it makes use of inter-relationships between the sentences by sharing its rank scores to its neighbours.
- Various penalty enforcement – if our method does not impose various penalty on sentences, the ROUGE-1 scores will be decreased.
- Differentiation between inter document / intra document links - if our method does not Differentiate between inter document and intra document links, the ROUGE-1 scores will be decreased reasonably.

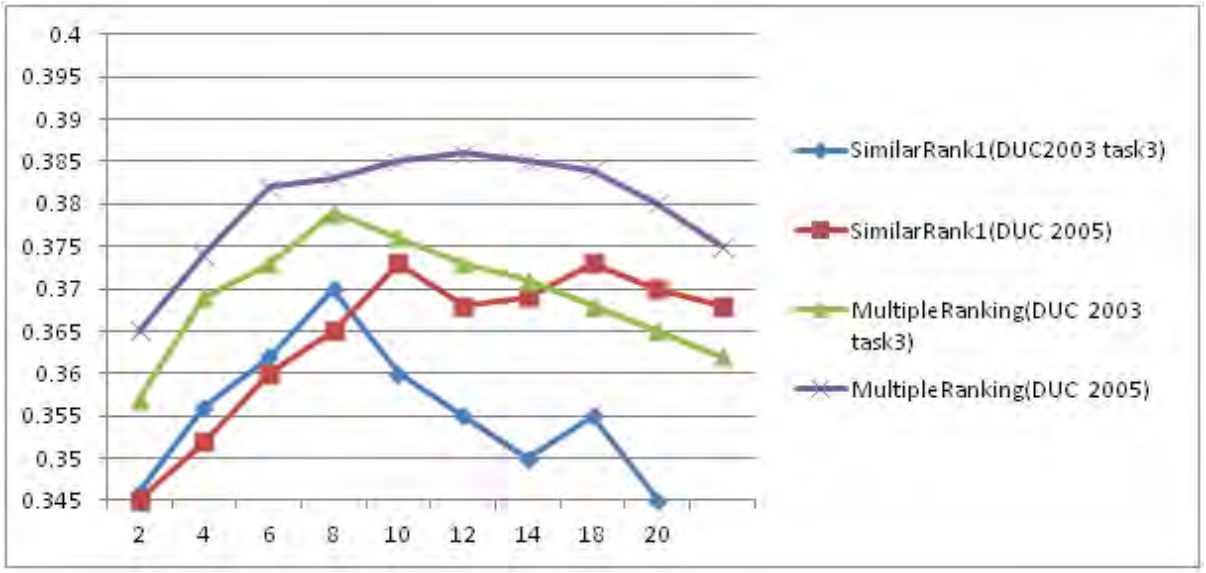

Fig.1 ROUGE 1 vs. ω

Figure1 shows the influence of the penalty factor ω in the multiple ranking methodology and the baseline approach when λ_1: λ_2 = 0.3: 1 and α = 0.6. We can see that multiple ranking methodology is better than SimilarRank1. It checks the relationships between the sentences of the documents are useful in the summarization task.

Figure 2 shows the influence of the inter document / intra document link by differentiating weight λ_1: λ_2 when ω = 8 and α = 0.6. it is noticed that when more importance is given to intra document links, the performances decrease gradually. If inter document links are not considered, it will become worst case. Still the performances are very well while giving importance to inter document links than intra document links for summarization purpose.

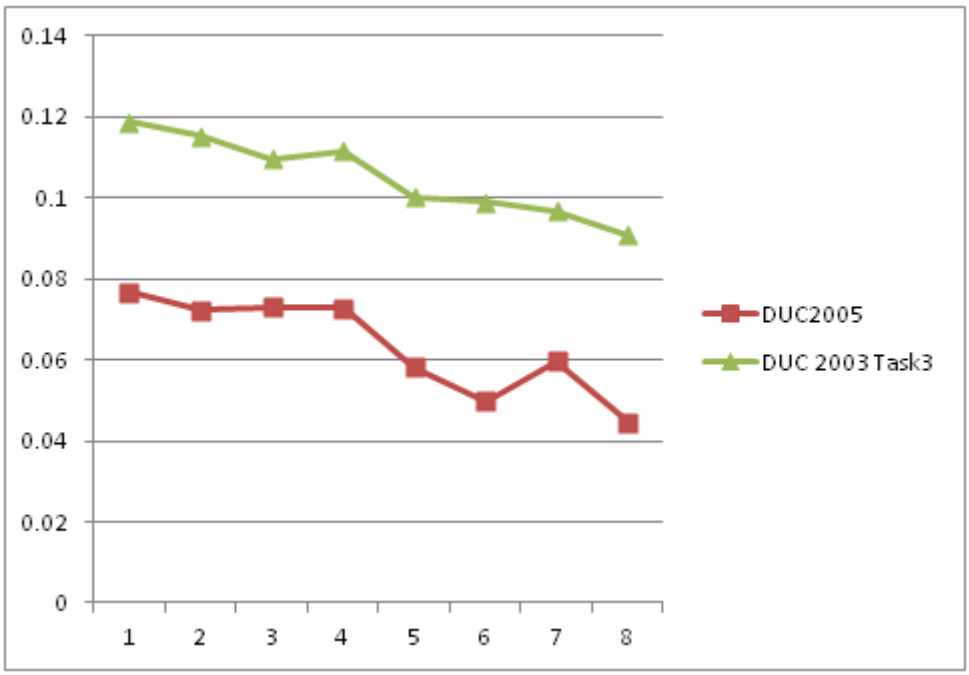

Fig. 2 ROUGE 1 vs. λ1: λ2

5. Conclusion and Future work

In this paper, we proposed query-based multiple ranking for multi-document summarization. We make use of the relationships among the sentences and the association between the query description and the sentences. Our experimental results prove that the proposed methodology performs well than the others.

In future we will apply machine learning methodology to estimate the parameters automatically. Finally, and most importantly, we are interested in applying this algorithm to wide-range of real-world problems.

References

[1] J. M. Conroy and J. D. Schlesinger.2005. CLASSY query-based multi-document summarization. In Proceedings of DUC'2005.
[2] G. Erkan and D. Radev. LexPageRank: prestige in multi-document text summarization. In Proceedings of EMNLP' 04.
[3] A. Farzindar, F. Rozon and G. Lapalme. 2005. CATS a topic-oriented multi-document summarization system at DUC 2005. In Proceedings of the 2005 Document Understanding Workshop (DUC2005).
[4] J. Ge, X. Huang and L. Wu. Approaches to event-focused summarization based on named entities and query words. In Proceedings of the 2003 Document Understanding Workshop (DUC2003).
[5] J. Goldstein, M. Kantrowitz, V. Mittal and J. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Proceedings of ACM SIGIR-1999.
[6] S. Harabagiu and F. Lacatusu. Topic themes for multi-document summarization. In Proceedings of SIGIR'2005.
[7] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise and X. Zhang. Cross-document summarization by concept classification. In Proceedings of SIGIR'2002.
[8] E. Hovy, C.-Y. Lin and L. Zhou. 2005. A BE-based multi-document summarizer with query interpretation.In Proceedings of DUC2005.
[9] C.-Y. Lin and E. H. Hovy. From Single to Multi-document Summarization: A Prototype System and its Evaluation. In Proceedings of ACL'2002.
[10] C.-Y. Lin and E.H. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of HLT-NAACL 2003.
[11] I. Mani and E. Bloedorn. Summarizing Similarities and Differences Among Related Documents.Information Retrieval, 1(1), 2000.
[12] R. Mihalcea and P. Tarau. A language independent algorithm for single and multiple document summarization. In Proceedings of IJCNLP'2005.
[13] D. R. Radev, H. Y. Jing, M. Stys and D. Tam. Centroid-based summarization of multiple documents. Information Processing and Management, 40: 919-938, 2004.
[14] H. Saggion, K. Bontcheva and H. Cunningham. Robust generic and query-based summarization. In Proceedings of EACL'2003.
[15] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In Proceedings of SIGIR'2005.
[16] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. SchÖlkopf. Learning with local and global consistency. In Proceedings of NIPS'2003.
[17] D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. SchÖlkopf. Ranking on data manifolds.