

Many are Better Than One: Improving Multi-Document Summarization via Weighted Consensus

Dingding Wang Tao Li
School of Computer Science
Florida International University
Miami, FL 33199
{dwang003,taoli}@cs.fiu.edu

ABSTRACT

Given a collection of documents, various multi-document summarization methods have been proposed to generate a short summary. However, few studies have been reported on aggregating different summarization methods to possibly generate better summarization results. We propose a weighted consensus summarization method to combine the results from single summarization systems. Experimental results on DUC2004 data sets demonstrate the performance improvement by aggregating multiple summarization systems, and our proposed weighted consensus summarization method outperforms other combination methods.

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

General Terms: Algorithms, Experimentation.

Keywords: Weighted consensus, summarization.

1. INTRODUCTION

Various multi-document summarization methods base on different strategies and usually produce diverse outputs. A natural question arises: can we perform ensemble or consensus summarization by combining different summarization methods to improve summarization performance? In general, the terms of “consensus methods” or “ensemble methods” are commonly reserved for the aggregation of a number of different (input) systems. Previous research has shown that ensemble methods, by combining multiple input systems, are a popular way to overcome instability and increase performance in many machine learning tasks, such as classification, clustering and ranking. The success of ensemble methods in other learning tasks provides the main motivation for applying ensemble methods in summarization. To the best of our knowledge, so far there are only limited attempts on using ensemble methods in multi-document summarization.

As a good ensemble requires the diversity of the individual members, here we study several widely used multi-document summarization systems based on a variety of strategies, and evaluate different baseline combination methods for obtaining a consensus summarizer to improve the summarization performance. Motivated from [5], we also propose a novel weighted consensus scheme to aggregate the results from

individual summarization methods, in which, the relative contribution of an individual summarizer to the consensus is determined by its agreement with other members of the summarization systems. Note that usually a high degree of agreements does not automatically imply the correctness since the systems could agree on a faulty answer. However, each of the summarization systems has shown its effectiveness individually, so the agreement measure can be used in the consensus summarization.

2. WEIGHTED CONSENSUS SUMMARIZATION (WCS)

2.1 Notations

Suppose there are K single summarization methods, each of which produces a ranking for the sentences containing in the document collection. Then we have K ranking lists $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_K\}$ and $\mathbf{r}_i \in \mathbb{R}^N, i = 1, \dots, K$, where N is the total number of sentences in the documents. The task is to find a weighted consensus ranking of the sentences \mathbf{r}^* with a set of weights $\{w_1, w_2, \dots, w_K\}$ assigning to each of the individual summarization methods.

2.2 Formulation

Our goal is to minimize the weighted distance between \mathbf{r}^* and all the \mathbf{r}_i . Let $\mathbf{w} = [w_1, w_2, \dots, w_K]^T \in \mathbb{R}^K$. The problem can be formulated as follows.

$$\begin{aligned} \arg \min_{\mathbf{w}} \quad & (1 - \lambda) \sum_{i=1}^K w_i \|\mathbf{r}^* - \mathbf{r}_i\|^2 + \lambda \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \sum_{i=1}^K w_i = 1; \quad w_i \geq 0 \quad \forall i, \end{aligned}$$

where $0 \leq \lambda \leq 1$ is the regularization parameter which specifies the tradeoff between the minimization of the weighted distance and the smoothness enforced by \mathbf{w} . In experiments, λ is set to 0.3 empirically. For simplicity, we use Euclidean distance to measure the discordance of the consensus ranking \mathbf{r}^* and each of individual sentence rankings \mathbf{r}_i .

We initialize $w_i = \frac{1}{K}$, and this optimization problem can be solved by iterating the following two steps:

Step 1: Solve for \mathbf{r}^* while fixing \mathbf{w} . The optimal solution is the weighted average $\mathbf{r}^* = \sum_i w_i \mathbf{r}_i$.

Step 2: Solve for \mathbf{w} while fixing \mathbf{r}^* . Let

$$\mathbf{d} = [\|\mathbf{r}^* - \mathbf{r}_1\|^2, \|\mathbf{r}^* - \mathbf{r}_2\|^2, \dots, \|\mathbf{r}^* - \mathbf{r}_K\|^2]^T \in \mathbb{R}^K.$$

Note that

$$(1 - \lambda) \sum_{i=1}^K w_i \|\mathbf{r}^* - \mathbf{r}_i\|^2 + \lambda \|\mathbf{w}\|^2 = (1 - \lambda) \mathbf{d}^\top \mathbf{w} + \lambda \mathbf{w}^\top \mathbf{w}$$

$$= \lambda \|\mathbf{w} - \frac{\lambda - 1}{2\lambda} \mathbf{d}\|^2 - \frac{(\lambda - 1)^2}{4\lambda} \|\mathbf{d}\|^2$$

For fixing \mathbf{r}^* , the optimization problem becomes

$$\arg \min_{\mathbf{w}} \|\mathbf{w} - \frac{\lambda - 1}{2\lambda} \mathbf{d}\|^2 \text{ s.t. } \sum_{i=1}^K w_i = 1; \quad w_i \geq 0, \quad \forall i$$

This is a quadratic function optimization problem with linear constraints with K variables. This is a problem of just about tens of variables (i.e., weights for each input summarization system) and thus can be computed quickly. It can also be solved by simply projecting vector $\frac{\lambda - 1}{2\lambda} \mathbf{d}$ onto $(K - 1)$ -simplex. With step 1 and 2, we iteratively update \mathbf{w} and \mathbf{r}^* until convergence. Then we sort \mathbf{r}^* in ascending order to get the consensus ranking.

3. EXPERIMENTS

In the experiments, we use four typical multi-document summarization methods as individual summarizers and compare our WCS method with other eight aggregation methods. The four individual summarization methods are: (a) Centroid [7], (b) LexPageRank [1], (c) LSA [2], and (d) NMF [4]. And the baseline aggregation methods are: (1) average score (Ave_Score), which normalizes and averages the raw scores from different summarization systems; (2) average rank (Ave_Rank), which averages individual rankings; (3) median aggregation (Med_Rank); (4) Round Robin (RR); (5) Borda Count (BC); (6) correlation-based weighting (CW), which weights individual systems by their average Kendall’s Tau correlation between the ranking list they generated and all the other lists; (7) ULTRA [3], which aims to find a consensus ranking with the minimum average Spearman’s distance [8] to all the individual ranking lists; (8) graph-based combination (Graph), the basic idea of which is similar to the work proposed in [9], however, we use cosine similarity so that we can compare this method with other combination methods fairly. We conduct experiments on DUC benchmark data for generic multi-document summarization and use ROUGE [6] toolkit (version 1.5.5) to measure the summarization performance.

3.1 Overall Summarization Performance

Table 1 show Rouge-1, Rouge-2, and Rouge-SU scores of different individual and combination methods using DUC2004 data sets (intuitively, the higher the scores, the better the performance). From the results, we observe that (1) Most of the combination summarization systems outperform all the individual systems except the round robin combination. The results demonstrate that in general consensus methods can improve the summarization performance. (2) Weighted combinations (e.g., CW, ULTRA, and WCS) outperform average combination methods which treat each individual system equally. (3) Our WCS method outperforms other weighted combination methods because WCS optimizes the weighted distance between the consensus sentence ranking to individual rankings and updates the weights and consensus ranking iteratively, which is closer to the nature of consensus summarization than other approximation based weighted methods.

Systems	R-1	R-2	R-SU
DUCBest	0.382	0.092	0.132
Centroid	0.367	0.073	0.125
LexPageRank	0.378	0.085	0.130
LSA	0.341	0.065	0.119
NMF	0.367	0.072	0.129
Ave_Score	0.388	0.089	0.132
Ave_Rank	0.385	0.087	0.131
Med_Rank	0.385	0.087	0.131
RR	0.364	0.072	0.126
BC	0.378	0.085	0.129
CW	0.378	0.085	0.131
ULTRA	0.392	0.090	0.133
Graph	0.379	0.086	0.132
WCS	0.398	0.096	0.135

Table 1: Overall performance comparison on DUC2004. Remark: DUCBest shows the best results from DUC 2004 competition.

3.2 Diversity of Individual Summarizers

In this set of experiments, we further examine if the four individual summarization methods are complementary to each other. We use our WCS method to aggregate any three of the four summarization methods and compare the results with the aggregation utilizing all the four methods. Table 2 shows the comparison results. From the results, we observe that adding any of the four individual methods improves the summarization performance. This is because these individual summarization methods are diverse and their performance is data dependant.

Systems	R-1	R-2	R-SU
Centroid+LexPageRank+LSA	0.383	0.088	0.132
Centroid+LexPageRank+NMF	0.385	0.090	0.133
Centroid+LSA+NMF	0.376	0.082	0.131
LexPageRank+LSA+NMF	0.382	0.087	0.132
All	0.398	0.096	0.135

Table 2: WCS results on DUC2004.

Acknowledgements: The work is partially supported by an FIU Dissertation Year Fellowship and NSF grants IIS-0546280 and DMS-0915110.

4. REFERENCES

- [1] G. Erkan and D. Radev. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, 2004.
- [2] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR*, 2001.
- [3] A. Klementiev, D. Roth, and K. Small. An unsupervised learning algorithm for rank aggregation. In *ECML*, 2007.
- [4] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, pages 788–791, 1999.
- [5] T. Li and C. Ding. Weighted consensus clustering. *SIAM Data Mining*, 2008.
- [6] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NLT-NAACL*, 2003.
- [7] D. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, pages 919–938, 2004.
- [8] C. Spearman. The proof and measurement of association between two things. *Amer. J. Psychol.*, 1904.
- [9] V. Thapar, A. A. Mohamed, and S. Rajasekaran. A consensus text summarizer based on meta-search algorithms. In *Proceedings of 2006 IEEE International Symposium on Signal Processing and Information Technology*, 2006.