Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Weighted consensus multi-document summarization

Dingding Wang, Tao Li *

School of Computer Science, Florida International University, United States

ARTICLE INFO

Article history:
Received 9 July 2010
Received in revised form 27 June 2011
Accepted 6 July 2011
Available online 23 August 2011

Keywords:
Multi-document summarization
Weighted consensus

ABSTRACT

Multi-document summarization is a fundamental tool for document understanding and has received much attention recently. Given a collection of documents, a variety of summarization methods based on different strategies have been proposed to extract the most important sentences from the original documents. However, very few studies have been reported on aggregating different summarization methods to possibly generate better summary results. In this paper, we propose a weighted consensus summarization method to combine the results from single summarization systems. We evaluate and compare our proposed weighted consensus method with various baseline combination methods. Experimental results on DUC2002 and DUC2004 data sets demonstrate the performance improvement by aggregating multiple summarization systems, and our proposed weighted consensus summarization method outperforms other combination methods.

1. Introduction

Multi-document summarization aims to generate a compressed summary by extracting the major information in a collection of documents sharing the same or similar topics. With the explosive growing of the volume and complexity of document data (e.g., news, blogs, web pages) on the Internet, multi-document summarization provides a useful solution for understanding documents and reducing information overload. Thus, multi-document summarization has attracted much attention in recent years, and many applications have been developed. For example, summarized informative snippets in web search can help users in further browsing (Turpin, Tsegay, Hawking, & Williams, 2007), and short summaries for news groups in news services can facilitate users to better understand the news articles (Sampathsampath & Martinovic, 2002).

A variety of multi-document summarization methods have been developed in the literature. The most commonly used methods are centroid based, which usually rank sentences in the document collection according to their scores calculated by a set of predefined features, such as term frequency-inverse sentence frequency (TF-ISF) (Lin & Hovy, 2002; Radev, Jing, Stys, & Tam, 2004), sentence or term position (Lin & Hovy, 2002; Yih, Goodman, Vanderwende, & Suzuki, 2007), and number of keywords (Yih et al., 2007). Another type of methods use sentence graph representation and select sentences based on the votes from their neighbors using the ideas similar to PageRank (Erkan & Radev, 2004; Mihalcea & Tarau, 2005). In addition, latent semantic analysis (LSA) and non-negative matrix factorization (NMF) have also been used to produce the summaries by selecting semantically and probabilistically important sentences in the documents (Gong & Liu, 2001).

Different 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

* Corresponding author.
E-mail addresses: dwang003@cs.fiu.edu (D. Wang), taoli@cs.fiu.edu (T. Li).

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.07.003

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 (Wang & Li, 2010).

As a good ensemble requires the diversity of the individual members, in this paper, we first study the most widely used multi-document summarization systems based on a variety of strategies (e.g., the centroid-based method, the graph-based method, LSA, and NMF), and evaluate different baseline combination methods (e.g., average score, average rank, Borda count, median aggregation, round-robin scheme, correlation based weighting method, and graph based combination) for obtaining a consensus summarizer to improve the summarization performance. 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. Experiments on DUC2002 and DUC2004 data sets demonstrate the performance improvement using various consensus multi-document summarization methods, and our proposed weighted consensus scheme outperforms the other baseline combination methods.

The rest of this paper is organized as follows. Section 2 discusses the related work on multi-document summarization and consensus ranking methods. Our proposed weighted consensus summarization is studied in Section 3. The summarization methods and aggregation methods implemented in our experimental study are presented in Section 4. Experimental results are shown and discussed in Section 5. Finally Section 6 concludes.

2. Related work

2.1. Multi-document summarization

Multi-document summarization has been widely studied recently. In general, document summarization can be divided into extractive summarization and abstractive summarization. In extractive summarization the important sentences are selected from original documents based on their assigned scores. Abstractive summarization involves information fusion, sentence compression and reformulation (Knight & Marcu, 2002; Jing & McKeown, 2000). Although an abstractive summary could be more concise, it requires deep natural language processing techniques. Thus extractive summaries are more feasible and has become the standard in document summarization. In this paper we focus on extractive multi-document summarization. There are several most widely used extractive summarization methods as follows.

- *Centroid-based methods*: This type of methods ranks sentences by computing their salience using a set of features. For example, MEAD (Radev et al., 2004) is a typical centroid-based algorithm which extracts sentences according to three parameters, i.e. centroid value, positional value, and first-sentence overlap. The centroid value of a sentence is computed as the average cosine similarity between the sentences and the rest of the sentences in the document collection. The positional value is computed as follows: the leading sentence is assigned score 1 and the score decreases by $1/n$ for each sentence, where n is the number of sentences in these documents. The overlap value is computed as the cosine similarity between a sentence and the first sentence in the same document. Then the three values are linearly combined with equal weights.
- *Graph-based methods*: This type of methods constructs a sentence graph, in which each node is a sentence in the document collection, and if the similarity between a pair of sentences is above a threshold or the sentences belong to the same document, there is an edge between the pair of sentences. The sentences are selected to form the summaries by voting from their neighbors. Erkan and Radev (2004) propose an algorithm called LexPageRank to compute the sentence importance based on the concept of eigenvector centrality (prestige) which has been successfully used in Google PageRank. Other graph-based summarization have been proposed in Mihalcea and Tarau (2005) and Wan and Yang (2008).
- *Latent semantic analysis (LSA)*: Gong and Liu (2001) propose a method using latent semantic analysis (LSA) to select highly ranked sentences for summarization. The method first creates a term–sentence matrix, where each column represents the weighted term-frequency vector of a sentence in the set of documents. Then singular value decomposition (SVD) is used on the matrix to derive the latent semantic structure. The sentences with the greatest combined weights across all the important topics are included in the summary.
- *Non-negative matrix factorization (NMF)*: This type of methods conducts NMF on the sentence–term matrix to extract sentences with the highest probability in each topic. NMF can also be viewed as a clustering method, which has many nice properties and advantages (Ding, He, & Simon, 2005). Intuitively, this method clusters these sentences and chooses the most representative ones from each cluster to form the summary.
- *Other methods*: Other methods include CRF-based summarization (Shen, Sun, Li, Yang, & Chen, 2007), and hidden Markov model (HMM) based method (Conroy & O'Leary, 2001). Some query-based summarization systems are also proposed (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999; Wan & Yang, 2007). For example, Language Computer Corporation (LCC) (LCC, xxxx), a DUC participant, that proposes a system combining the question-answering and summarization system and using k-nearest neighbor clustering based on cosine similarity for the sentence selection.

Although various summarization approaches have been developed in literature, few efforts have been reported on aggregating document summarization methods. One work related to ensemble summarization is described in Thapar, Mohamed, and Rajasekaran (2006), where a graph-based meta-summarization approach by comparing the document graph of individual summary with the centric graph for all the summary from different summarization systems is proposed. In our work, we systematically evaluate different baseline combination methods for ensemble summarization and propose a novel weighted consensus scheme to aggregate the results from individual summarization methods.

2.2. Rank aggregation

Since different summarization systems rank the sentences in the document collection using various strategies, the results from each system can be viewed as a ranking of the sentences. The problem of combining multiple ranking results into a consensus ranking is known as rank aggregation (Aslam & Montague, 2001; Erp & Schomaker, 2000; Manmatha, Rath, & Feng, 2001).

Most rank aggregation approaches implicitly conduct majority voting to create the final rank. For example, the simplest approaches can average the scores or ranks from individual systems. The round-robin scheme (Sidney, 1976) can also be applied by selecting the first entity from the first ranker, the first entity from the second ranker, and so on. Borda count (Erp & Schomaker, 2000) sorts the entities based on their positions and counts the number of points the entities get from each voter. There are two types of rank aggregation: unsupervised and supervised. Most of the unsupervised rank aggregation approaches count the entities ranked below them in all the ranking lists. Median rank aggregation (Erp & Schomaker, 2000) sorts the entities based on the medians of their ranks in all the ranking lists. One fundamental problem of these methods is that they treat all the rankings equally. However, different systems have different accuracies and should be treat differently. Supervised rank aggregation usually determines the weights of each ranking list by learning an aggregation function using labeled data (Liu, Liu, Qin, Ma, & Li, 2007; Lillis, Toolan, Collier, & Dunnion, 2006). Although supervised aggregation can achieve higher accuracy, in practice the labeled data are not always available. Recently, Klementiev, Roth, and Small (2007) propose a framework called ULARA to learn the weights of the ranking lists online without supervision by optimizing the weighted Borda count. They also develop an EM-based algorithm which uses each ranking as an observation to estimate the parameters for combining the ranking lists (Klementiev & Roth, 2008).

2.3. Consensus document summarization

In this paper, we propose a weighted consensus scheme to aggregate diverse summaries from different summarization systems and compare the results with both individual methods and other base aggregation methods.

3. Weighted consensus summarization (WCS)

3.1. 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 $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K\}$ and $\mathbf{r}_i \in \mathbb{R}^N,\ i = 1, \ldots, K$, 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 $\{w_1, w_2, \ldots, w_K\}$ assigning to each of the individual summarization methods.

3.2. Optimization-based weighted consensus summarization

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

$$arg\ min_{\mathbf{w},\mathbf{r}^*} \quad (1-\lambda)\sum_{i=1}^{K} w_i \|\mathbf{r}^* - \mathbf{r}_i\|^2 + \lambda\|\mathbf{w}\|^2 \tag{1}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} w_i = 1; \quad w_i \geqslant 0 \quad \forall i,$$

where $0 \leqslant \lambda \leqslant 1$ is the regularization parameter which specifies the tradeoff between the minimization of the weighted distance and the smoothness enforced by $\mathbf{w}$. In our experiments, λ is set to 0.3 empirically.[1] For simplicity, we use Euclidean distance to measure the discordance of the consensus ranking $\mathbf{r}^*$ and each of individual sentence rankings $\mathbf{r}_i$. Hence $w_i\|\mathbf{r}^* - \mathbf{r}_i\|^2$ measures the weighted distance between summarizer i and $\mathbf{r}^*$ and the first term of Eq. (1) is used to minimize the weighted distance from each individual summarizer to $\mathbf{r}^*$. The second term of Eq. (1) is a regularization term to enforce the smoothness of the weights.

[1] Experiments are conducted in Section 5.3.3 to demonstrate the parameter tuning.

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

$$\mathbf{r}^* = \sum_i w_i \mathbf{r}_i. \tag{2}$$

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

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

Note that

$$(1-\lambda)\sum_{i=1}^{K} w_i \|\mathbf{r}^* - \mathbf{r}_i\|^2 + \lambda\|\mathbf{w}\|^2 = (1-\lambda)\mathbf{d}^\top p\mathbf{w} + \lambda\mathbf{w}^\top p\mathbf{w} = \lambda\|\mathbf{w} - \frac{\lambda-1}{2\lambda}\mathbf{d}\|^2 - \frac{(\lambda-1)^2}{4\lambda}\|\mathbf{d}\|^2.$$

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

$$\begin{aligned} &arg\ min_{\mathbf{w}} \quad \|\mathbf{w} - \frac{\lambda-1}{2\lambda}\mathbf{d}\|^2 \\ &\text{s.t.} \quad \sum_{i=1}^{K} w_i = 1; \quad w_i \geqslant 0, \quad \forall i. \end{aligned} \tag{3}$$

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. An efficient projection algorithm can be referred to Duchi, Shalev-Shwartz, Singer, and Chandra (2008).

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.

Convergence: It is straightforward to show the convergence of the above two-step alternating optimization procedure. For simplicity, denote Eq. (1) as D and denote the initial $\mathbf{w}$ and $\mathbf{r}^*$ as $\mathbf{w}_0$ and $\mathbf{r}^*_0$. Starting from the initialization, we repeat the two-step procedure: $\mathbf{r}^*_i = arg\ minD(\mathbf{w}_{i-1}, \mathbf{r}^*_{i-1})$ and $\mathbf{w}_i = arg\ minD(\mathbf{w}_{i-1}, \mathbf{r}^*_i)$, where i is used to denote the ith iteration. Hence $D(\mathbf{w}_i, \mathbf{r}^*_i) \leqslant D(\mathbf{w}_{i-1}, \mathbf{r}^*_i) \leqslant D(\mathbf{w}_{i-1}, \mathbf{r}^*_{i-1})$. So D decreases strictly with each iteration and is always positive. Thus, convergence can often be established for this procedure.

4. Implemented systems

In the section, we describe four typical multi-document summarization methods and eight aggregation methods implemented in our experimental study.

4.1. Individual summarization methods

- *Centroid*: similar to MEAD algorithm proposed in Radev et al. (2004) using centroid value, positional value, and first-sentence overlap as features.
- *LexPageRank*: a graph-based summarization method recommending sentences by the voting of their neighbors (Erkan & Radev, 2004).
- *LSA*: conducts latent semantic analysis on terms by sentences matrix as proposed in Gong and Liu (2001).
- *NMF*: performs NMF on terms by sentences matrix and ranks the sentences by their weighted scores (Lee & Seung, 2001).

Each of the individual summarization methods ranks the sentences based on different criteria. For the methods which partially rank sentences for each cluster/topic (e.g. NMF), we integrate the rankings using the following scheme: we first order the clusters/topics according to their importance (we just simply order the clusters/topics based on their sizes empirically). Then we pick the top-ranked sentence in the most important cluster/topic and then the top-ranked sentence in the second important cluster. After all the top-ranked sentences in all the clusters have been selected, we start to select the second-ranked sentences in the most important cluster, and so on until all the sentences are ranked. This scheme is consistent with the strategies in most cluster-based summarization methods on selecting sentences with different summary lengths.

4.2. Aggregation methods

Here we list the aggregations methods used in our experimental study.

- *Average score (Ave_Score)*: normalizes the raw scores from different summarization systems between 0 and 1, and then averages the scores as follows:

$$Score(S_i) = \frac{\sum_{k=1}^{K} Score\_k(S_i)}{K},$$

where K is the number of summarization systems, and $Score\_k(S_i)$ is the normalized individual score by the kth system. Finally the sentences are re-ranked based on their average scores.

- *Average rank (Ave_Rank)*: averages the individual rankings from different summarizers as follows:

$$Rank(S_i) = \frac{\sum_{k=1}^{K} Rank\_k(S_i)}{K},$$

where $Rank\_k(S_i)$ is the ranking by the kth system. Then the sentences are sorted by their average ranking.

- *Median rank (Med_Rank)*: instead of using average rank, median rank is also often used to aggregate ranking lists.
- *Round robin (RR)*: picks the first sentence from the first ranking list, and then the first sentence from the second list. After all the first sentences are selected, the second sentence in the first list is selected, and so on until the summary length is reached.
- *Borda count (BC)*: each sentence gets 1 point for each last place ranking received, and 2 points for each next-to-last place ranking, and all the way up to N points for each first place ranking, where N is the number of sentences in total. Then the sentences are ranked based on their points obtained.
- *Correlation-based weighting* (CW): uses Kendall's Tau correlation (Abdi, 2007) to measure the agreement between two sentence rankings. The average correlation between a ranking list and all the other lists is computed as the weight of the system. The Kendall's Tau correlation is defined below:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)},$$

where N is the number of sentences, N_c is the number of concordant pairs of sentences, and N_d is the number of discordant pairs of sentences. A pair of sentences (e.g., S_i and S_j) is concordant if the order of the two sentences are the same in the two ranking lists (e.g., $Rank\_m$ and $Rank\_n$), i.e. if $Rank\_m\ (S_i) > Rank\_n(S_j)$, then $Rank\_n(S_i) > Rank\_n(S_j)$. So, a high value means the consistency of the two ranking lists. Finally, the weights are normalized between 0 and 1.

- *ULARA*: ULARA is an unsupervised rank aggregation method proposed in Klementiev et al. (2007). The weights of the summarization systems are calculated by optimizing the weighted Borda count, which aims to find a consensus ranking with the minimum average Spearman's distance (Spearman, 1904) to all the individual ranking lists. An online algorithm is derived using iterative gradient descent (Klementiev et al., 2007).
- *Graph-based combination (graph)*: constructs a sentence graph for each of the summary produced by the summarization systems using cosine similarity, where each node is a sentence and there is an edge if the similarity between a pair of sentences is above a predefined threshold. Then a consensus summary is generated by selecting sentences most similar to all the sentence graphs. The basic idea is similar to the work proposed in Thapar et al. (2006), however, we use cosine similarity to generate the sentence graph without natural language processing so that we can compare this method with other combination methods fairly.
- *Weighted consensus summarization (WCS)*: our proposed weighted consensus document summarization algorithm as described in Section 3.

In the experiments, we examine the summarization performance of the implemented individual and combination systems, and compare our proposed weighted consensus summarization algorithm with the other combination methods.

5. Experiments

In this section, we conduct experiments on DUC benchmark data to compare and evaluate individual and consensus summarization performance.

5.1. Data set

To evaluate the summarization results empirically, we use the DUC2002 and DUC2004 data sets, both of which are open benchmark data sets from Document Understanding Conference (DUC) for generic automatic summarization evaluation. Table 1 gives a brief description of the data sets, in which data source indicates where the documents are obtained. For example, DUC2002 data come from the Text REtrieval Conference (TREC), and DUC2004 data are from the Topic Detection and Tracking (TDT) research.

5.2. Evaluation methods

We use ROUGE (Lin & Hovy, 2003) toolkit (version 1.5.5) to measure the summarization performance, which is widely applied by DUC for performance evaluation. It measures the quality of a summary by counting the unit overlaps between

Table 1
Description of the data sets for multi-document summarization.

	DUC2002	DUC2004
Number of document collections	59	50
Number of documents in each collection	~10	10
Data source	TREC	TDT
Summary length	200 words	665 bytes

the candidate summary and a set of reference summaries. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-SU. ROUGE-N is an n-gram recall computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S\in \text{ref}}\sum_{\text{gram}_n\in S}\text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S\in \text{ref}}\sum_{\text{gram}_n\in S}\text{Count}(\text{gram}_n)}, \tag{4}$$

where n is the length of the n-gram, and ref stands for the reference summaries. $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of n-grams co-occurring in a candidate summary and the reference summaries, and Count (gram_n) is the number of n-grams in the reference summaries. ROUGE-L uses the longest common subsequence (LCS) statistics, while ROUGE-W is based on weighted LCS and ROUGE-SU is based on skip-bigram plus unigram. Each of these evaluation methods in ROUGE can generate three scores (recall, precision and F-measure). As we have similar conclusions in terms of any of the three scores, for simplicity, in this paper, we only report the average F-measure scores generated by ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W and ROUGE-SU to compare the implemented systems.

5.3. Experimental results

5.3.1. Overall performance

First of all, we compare the implemented consensus summarization methods with individual summarization systems to examine the effectiveness of consensus methods for summarization performance improvement. We also compare the performance of our proposed WCS method with other implemented combination methods.

Tables 2 and 3 show the ROUGE scores of different individual and combination methods using DUC2002 and DUC2004 data sets respectively. The bolded results highlight the best results in this set of experiments. And $*$ indicates that the improvement over the best individual summarizer LexPageRank is statistically significant (significance is measured by t-test statistics). From the results, we have the following observations.

1. Most of the combination summarization systems outperform all the individual systems except the round robin combination. The poor performance of the round robin combination may come from the inaccuracy or overlap of the very top ranking sentences of the single summarization results. The results demonstrate that in general consensus methods can improve the summarization performance.
2. Our proposed weighted consensus summarization (WCS) method outperforms other combination methods. For simple average combination schemes (such as Ave_Score, Ave_Rank, Med_Rank, RR, BC, and CW), they treat each individual summarization system equally. However, individual summarization methods may have different performance results on different data sets, thus introducing weights to form weighted combination is necessary. From the results we observe that the weighted combination methods are more effective than average combination methods. Among different weighted combination methods (e.g. CW, ULARA, and WCS), our WCS method optimizes the weighted distance between the consensus sentence ranking to individual rankings and updates the weights and consensus ranking iteratively, which is clo-

Table 2
Overall performance comparison on DUC2002 data using ROUGE evaluation methods.

Systems	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
DUC best	0.49869	0.25229	0.46803	0.20071	0.28406
Centroid	0.45379	0.19181	0.43237	0.17971	0.23629
LexPageRank	0.47963	0.22949	0.44332	0.18978	0.26198
LSA	0.43078	0.15022	0.40507	0.15220	0.20226
NMF	0.44587	0.16280	0.41513	0.16072	0.21687
Ave_Score	0.48589*	0.23752*	0.45590	0.19167*	0.26835*
Ave_Rank	0.48315*	0.23569*	0.45283	0.19069	0.26523*
Med_Rank	0.48302*	0.23524*	0.45155	0.19051	0.26571*
RR	0.46717	0.19506	0.43293	0.17839	0.25023
BC	0.48016	0.23281*	0.44823	0.19024	0.26308*
CW	0.48024	0.23232	0.44935	0.19008	0.26319*
ULARA	0.49037*	0.24628*	0.46091*	0. 19788*	0.27610*
Graph	0.48196*	0.23419	0.45116	0.19227*	0.26858*
WCS	0.49334*	0.24837*	0.46283*	0.19976*	0.27891*

Table 3
Overall performance comparison on DUC2004 data using ROUGE evaluation methods.

Systems	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
DUC best	0.38224	0.09216	0.38687	0.13325	0.13233
Centroid	0.36728	0.07379	0.36182	0.12439	0.12511
LexPageRank	0.37842	0.08572	0.37531	0.13121	0.13097
LSA	0.34145	0.06538	0.34973	0.12042	0.11946
NMF	0.36747	0.07261	0.36749	0.12961	0.12918
Ave_Score	0.38826*	0.08953*	0.37981*	0.13203	0.13215*
Ave_Rank	0.38522	0.08741*	0.37759	0.13187*	0.13112
Med_Rank	0.38538*	0.08733	0.37762	0.13162*	0.13128
RR	0.36463	0.07273	0.36449	0.12531	0.13637*
BC	0.37857	0.08562	0.37587	0.12928	0.12986
CW	0.37885*	0.08586	0.37613	0.12937	0.13024
ULARA	0.39217*	0.09027*	0.38797*	0.13712*	0.13311*
Graph	0.37921	0.08674*	0.37622	0.13118	0.13154
WCS	0.39872*	0.09611*	0.38928*	0.13866*	0.13532*

ser to the nature of consensus summarization than other approximation based weighted methods such as CW and ULARA and avoids trivial solutions. In addition, the performance of WCS is also better than the graph-based combination because the graph-based method only considers the subset of sentences selected by individual summarization systems.

3. We also list the results of the best team in the DUC competition, and notice that although the performance of each single summarization method is not as good as the best team, many of the consensus summarization solutions outperform the best team of the DUC participants (especially on DUC2004 data set). Note that the good performance of the best team in DUC benefits from their preprocessing on the data using deep natural language analysis which is not applied in our implemented systems.

To better demonstrate the results, Figs. 1 and 2 visually illustrate the comparison. Note that we subtract the ROUGE scores of the best single summarization method from all the combination methods in these figures, thus the difference can be observed more clearly. We show ROUGE-1 results in these figures.

5.3.2. Diversity of individual summarization methods

In this paper, we use four individual summarization methods as the baselines. These individual summarization methods are selected as the representatives of the most widely used types of summarization methods, and they are fundamentally different in both algorithm design and implementation, which makes them diverse and complimentary with each other. The centroid-based summarization usually includes the sentences of the highest similarities with all the other sentences in the documents into the summary, which is good since these sentences deliver the majority of information contained in

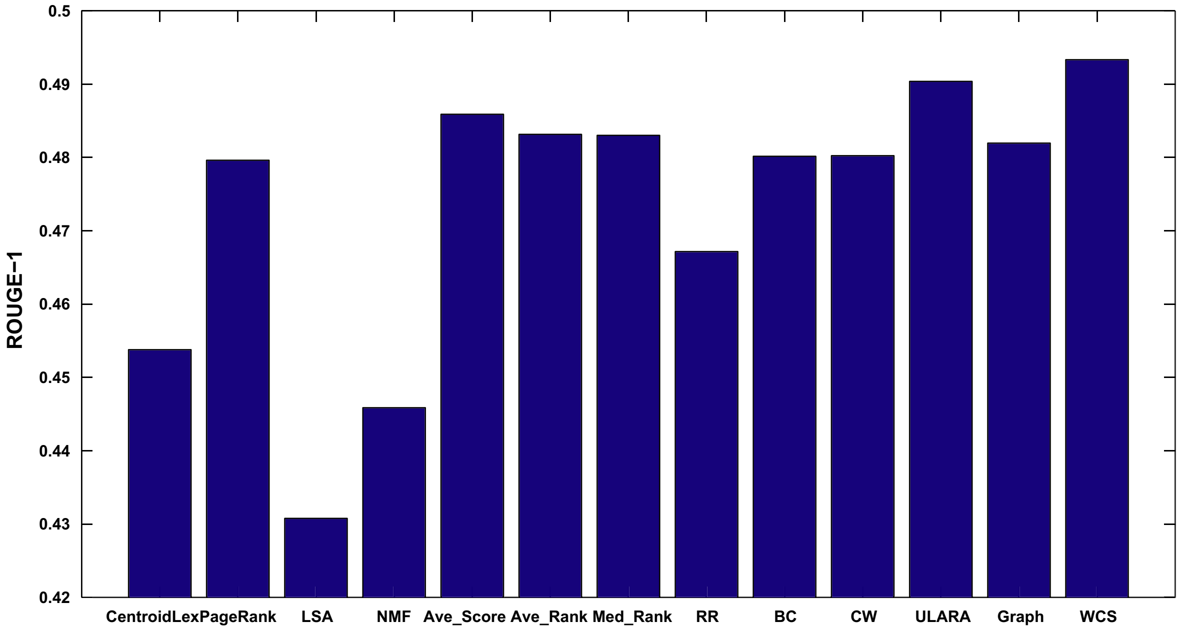

Fig. 1. Overall summarization performance on DUC2002 data using ROUGE-1.

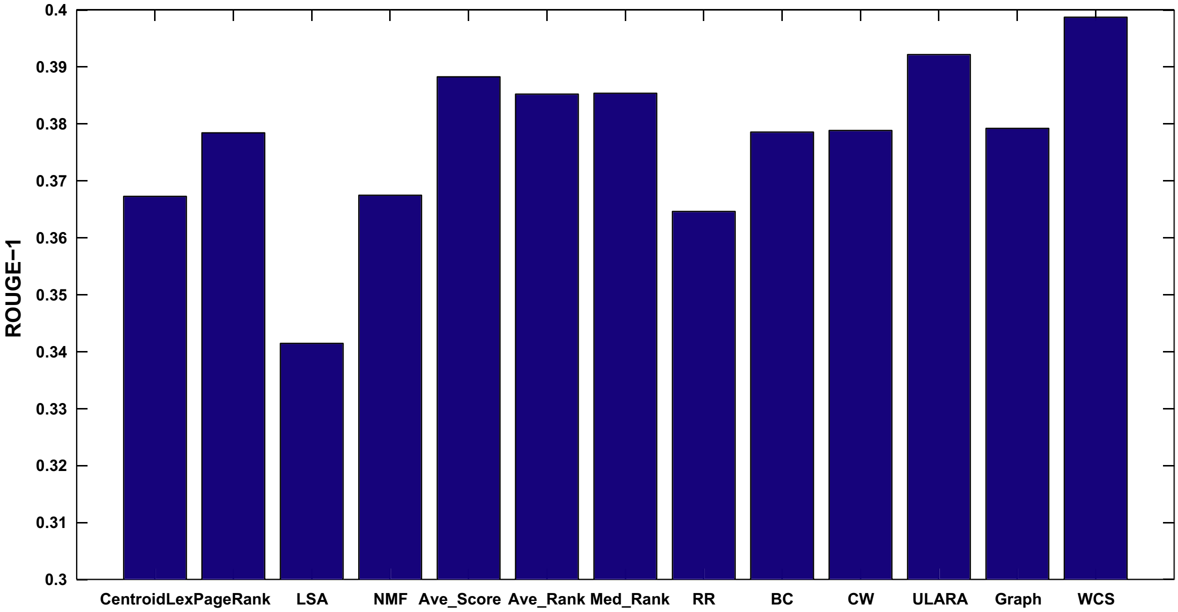

Fig. 2. Overall summarization performance on DUC2004 data using ROUGE-1.

Table 4
WCS results on DUC2002 data using ROUGE evaluation methods. The abbreviations are: C + P + L (Centroid + LexPageRank + LSA); C + P + N (Centroid + LexPageRank + NMF); C + L + N (Centroid + LSA + NMF); and P + L + N (LexPageRank + LSA + NMF).

Systems	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
C + P + L	0.48673	0.23451	0.45075	0.19327	0.26815
C + P+N	0.48930	0.23676	0.45383	0.19582	0.27036
C + L+N	0.48009	0.23208	0.44809	0.18966	0.25296
P + L+N	0.48529	0.23422	0.44937	0.18966	0.25296
All	**0.49334**	**0.24837**	**0.46283**	**0.19976**	**0.27891**

the documents, however the redundancy needs to be further removed and the subtopics in the documents are hard to detect. The graph-based methods such as LexPageRank apply graph analysis and take the influence of other sentences into consideration, which provides a better view of the relationships embedded in the sentences. LSA and NMF are both factorization based techniques which extract the semantic structure and hidden topics in the documents and select the sentences representing each topic as the summary. However, with nonnegative constrains, NMF provides more natural interpretations of document data. 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. Tables 4 and 5 show the comparison results. The bolded results represent the best results in this set of experiments. Figs. 3 and 4 demonstrate the same results for better illustration.

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, i.e., some methods may work well on certain data. Thus the four methods are complementary to each other, and combining them do improve the overall summarization performance.

Table 5
WCS results on DUC2004 data using ROUGE evaluation methods. The abbreviations are: C + P + L (Centroid + LexPageRank + LSA); C + P + N (Centroid + LexPageRank + NMF); C + L + N (Centroid + LSA + NMF); and P + L + N (LexPageRank + LSA + NMF).

Systems	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
C + P + L	0.38337	0.08852	0.37928	0.13387	0.13226
C + P + N	0.38542	0.09006	0.38156	0.13503	0.13305
C + L + N	0.37621	0.08257	0.37531	0.13012	0.13113
P + L + N	0.38219	0.08763	0.37911	0.13354	0.13212
All	**0.39872**	**0.09611**	**0.38928**	**0.13866**	**0.13532**

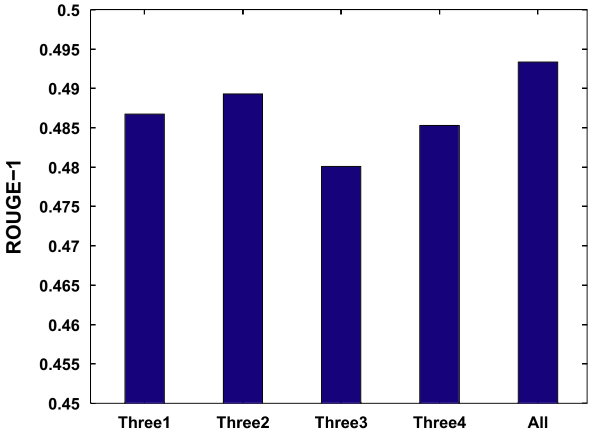

Fig. 3. WCS results on DUC2002 data. Remark: "Three1" represents Centroid + LexPageRank + LSA; "Three2" represents Centroid + LexPageRank + NMF; "Three3" represents Centroid + LSA + NMF; "Three4" represents LexPageRank + LSA + NMF.

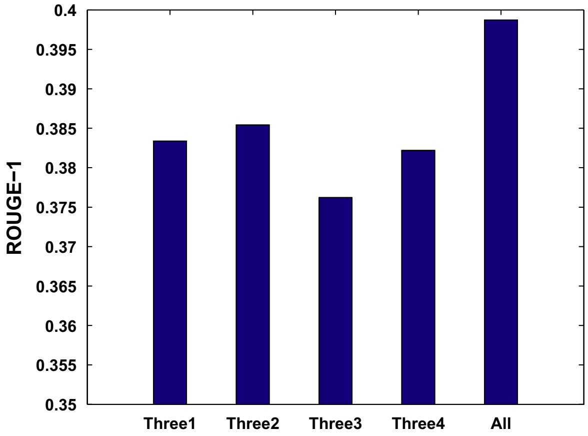

Fig. 4. WCS results on DUC2004 data.

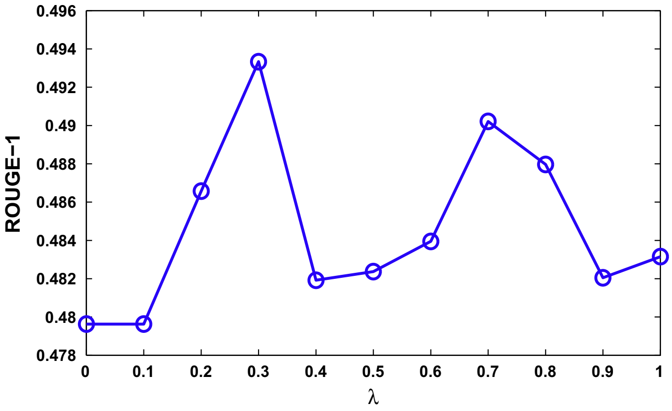

Fig. 5. ROUGE-1 results of WCS parameter tuning using DUC2002 data.

5.3.3. Parameter tuning

In Figs. 5 and 6, we gradually tune the parameter λ in our WCS method to adjust the weights between the weighed sentence ranking distance and the smoothness. When $\lambda = 0$, there will be a trivial solution to select the single method which is

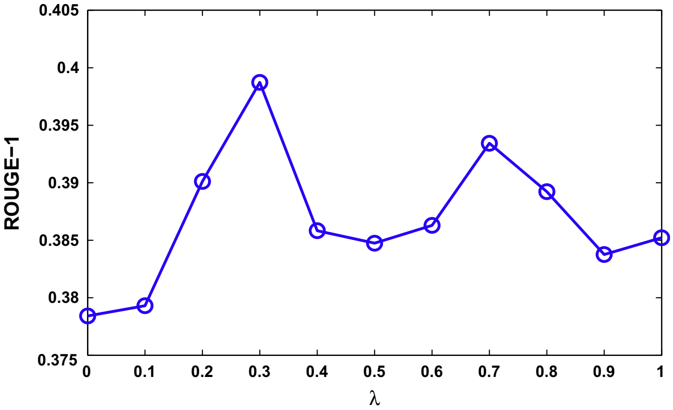

Fig. 6. ROUGE-1 results of WCS parameter tuning using DUC2004 data.

the closest to the other single methods on average. When $\lambda = 1$, the solution will equal to the average ranking. Here, λ is adjusted from 0 to 1 in every 0.1 interval. We observe that when λ is 0.3, the performance is the best, and the weights assigned to the four methods are 0.2085, 0.5025, 0.0982, and 0.1908, respectively.

6. Conclusion

In this paper, we study four most widely used multi-document summarization systems (i.e. the centroid-based method, the graph-based method, LSA, and NMF) and propose a weighted consensus summarization method to combine the results from single summarization systems. We evaluate and compare our proposed weighted consensus method with various combination methods (e.g. average score, average rank, Borda count, median aggregation, round-robin scheme, correlation-based weighting method, and graph-based combination), and experimental results on DUC2002 and DUC2004 data sets demonstrate the performance improvement by aggregating multiple summarization systems, and our proposed weighted consensus summarization method outperforms other combination methods.

Acknowledgement

The work is partially supported by US NSF Grants IIS-0546280 and DMS-0915110.

References

http://www-nlpir.nist.gov/projects/duc/pubs/.
Abdi, H. (2007). The Kendall rank correlation coefficientN. J. Salkind (Ed.). *Encyclopedia of Measurement and Statistics*.
Aslam, J. A., & Montague, M. (2001). Models for metasearch. In *Proceedings of SIGIR* (pp. 276–284).
Conroy, J., & O'Leary, D. (2001). Text summarization via hidden markov models. In *Proceedings of SIGIR*.
Ding, C., He, X., & Simon, H. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of Siam data mining*.
Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the l_1-ball for learning in high dimensions. In W. W. Cohen, A. McCallum, & S. T. Roweis (Eds.), *ICML, ACM international conference proceeding series* (Vol. 307, pp. 272–279). ACM.
Erkan, G., & Radev, D. (2004). Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*.
Erp, M. V., & Schomaker, L. (2000). Variants of the Borda count method for combining ranked classifier hypotheses. In *Proceedings of the 7th international workshop on frontiers in handwriting recognition*.
Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of SIGIR*.
Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of SIGIR*.
Jing, H., & McKeown, K. (2000). Cut and paste based text summarization. In *Proceedings of NAACL*.
Klementiev, A., Roth, D., & Small, K. (2007). An unsupervised learning algorithm for rank aggregation. In *Proceedings of the 2007 European conference on machine learning (ECML)*.
Klementiev, A., Roth, D., & Small, K. (2008). Unsupervised rank aggregation with distance-based models. In *Proceedings of ICML* (pp. 472–479).
Knight, K., & Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 91–107.
Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *NIPS*.
Lillis, D., Toolan, F., Collier, R., & Dunnion, J. (2006). Probfuse: A probabilistic approach to data fusion. In *Proceedings of SIGIR* (pp. 139–146). ACM.
Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using *n*-gram co-occurrence statistics. In *Proceedings of NLT-NAACL*.
Lin, C.-Y., & Hovy, E. (2002). From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of ACL*.
Liu, Y.-T., Liu, T.-Y., Qin, T., Ma, Z.-M., & Li, H. (2007). Supervised rank aggregation. In *Proceedings of WWW* (pp. 481–490).
Manmatha, R., Rath, T., & Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. In *Proceedings of SIGIR* (pp. 267–275).
Mihalcea, R., & Tarau, P. (2005). A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP*.
Radev, D., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, 919–938.
Sampathsampath, G., & Martinovic, M. (2002). *A Multilevel Text Processing Model of Newsgroup Dynamics*.
Shen, D., Sun, J.-T., Li, H., Yang, Q., & Chen, Z. (2007). Document summarization using conditional random fields. In *Proceedings of IJCAI*.
Sidney, C. (1976). *The Art of Legging-Maxline International*.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*.

Thapar, V., Mohamed, A. A., & Rajasekaran, S. (2006). A consensus text summarizer based on meta-search algorithms. In *Proceedings of 2006 IEEE international symposium on signal processing and information technology*.

Turpin, A., Tsegay, Y., Hawking, D., & Williams, H. (2007). Fast generation of result snippets in web search. In *Proceedings of SIGIR*.

Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the thirty-first annual international SIGIR conference*.

Wan, X., Yang, J., & Xiao, J. (2007). Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI*.

Wang, D., & Li, T. (2010). Many are Better than One: Improving Multi-document Summarization via Weighted Consensus. In *Proceedings of SIGIR* (pp. 809–810).

Yih, W.-T., Goodman, J., Vanderwende, L., & Suzuki, H. (2007). Multi-document summarization by maximizing informative content-words. In *Proceedings of IJCAI*.