Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Document Summarization Based on Data Reconstruction

Zhanying He and **Chun Chen** and **Jiajun Bu** and **Can Wang** and **Lijun Zhang**
Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science,
Zhejiang University, Hangzhou 310027, China.
{hezhanying, chenc, bjj, wcan, zljzju}@zju.edu.cn

Deng Cai and **Xiaofei He**
State Key Lab of CAD&CG, College of Computer Science,
Zhejiang University, Hangzhou 310058, China.
{dengcai, xiaofeihe}@cad.zju.edu.cn

Abstract

Document summarization is of great value to many real world applications, such as snippets generation for search results and news headlines generation. Traditionally, document summarization is implemented by extracting sentences that cover the main topics of a document with a minimum redundancy. In this paper, we take a different perspective from data reconstruction and propose a novel framework named *Document Summarization based on Data Reconstruction* (DSDR). Specifically, our approach generates a summary which consist of those sentences that can best reconstruct the original document. To model the relationship among sentences, we introduce two objective functions: (1) linear reconstruction, which approximates the document by linear combinations of the selected sentences; (2) nonnegative linear reconstruction, which allows only additive, not subtractive, linear combinations. In this framework, the reconstruction error becomes a natural criterion for measuring the quality of the summary. For each objective function, we develop an efficient algorithm to solve the corresponding optimization problem. Extensive experiments on summarization benchmark data sets DUC 2006 and DUC 2007 demonstrate the effectiveness of our proposed approach.

Introduction

With the explosive growth of the Internet, people are overwhelmed by a large number of accessible documents. Summarization can represent the document with a short piece of text covering the main topics, and help users sift through the Internet, catch the most relevant document, and filter out redundant information. So document summarization has become one of the most important research topics in the natural language processing and information retrieval communities.

In recent years, automatic summarization has been applied broadly in varied domains. For example, search engines can provide users with snippets as the previews of the document contents (Turpin et al. 2007; Huang, Liu, and Chen 2008; Cai et al. 2004; He et al. 2007). News sites usually describe hot news topics in concise headlines to facilitate browsing. Both the snippets and headlines are specific forms of document summary in practical applications.

Most of the existing generic summarization approaches use a ranking model to select sentences from a candidate set (Brin and Page 1998; Kleinberg 1999; Wan and Yang 2007). These methods suffer from a severe problem that top ranked sentences usually share much redundant information. Although there are some methods (Conroy and O'leary 2001; Park et al. 2007; Shen et al. 2007) trying to reduce the redundancy, selecting sentences which have both good coverage and minimum redundancy is a non-trivial task.

In this paper, we propose a novel summarization method from the perspective of data reconstruction. As far as we know, our approach is the first to treat the document summarization as a data reconstruction problem. We argue that a good summary should consist of those sentences that can best reconstruct the original document. Therefore, the reconstruction error becomes a natural criterion for measuring the quality of summary. We propose a novel framework called *Document Summarization based on Data Reconstruction* (DSDR) which finds the summary sentences by minimizing the reconstruction error. DSDR firstly learns a reconstruction function for each candidate sentence of an input document and then obtains the error formula by that function. Finally it obtains an optimal summary by minimizing the reconstruction error. From the geometric interpretation, DSDR tends to select sentences that span the intrinsic subspace of candidate sentence space so that it is able to cover the core information of the document.

To model the relationship among sentences, we discuss two kinds of reconstruction. The first one is linear reconstruction, which approximates the document by linear combinations of the selected sentences. Optimizing the corresponding objective function is achieved through a greedy method which extracts sentences sequentially. The second one is *non-negative* linear reconstruction, which allows only additive, not subtractive, combinations among the selected sentences. Previous studies have shown that there is psychological and physiological evidence for parts-based representation in the human brain (Palmer 1977; Wachsmuth, Oram, and Perrett 1994; Cai et al. 2011). Naturally, a document summary should consist of the parts of sentences. With the nonnegative constraints, our method leads to parts-based reconstruction so that no redundant information needs to be subtracted from the combination. We formulate the nonnegative linear reconstruction as a convex optimization problem and design a multiplicative updating algorithm which guarantees converging monotonically to a global minima.

Extensive experiments on summarization benchmark data sets DUC 2006 and DUC 2007 demonstrate the effectiveness of our proposed approach.

Related work

Recently, lots of extractive document summarization methods have been proposed. Most of them involve assigning salient scores to sentences of the original document and composing the result summary of the top sentences with the highest scores. The computation rules of salient scores can be categorized into three groups (Hu, Sun, and Lim 2008): feature based measurements, lexical chain based measurements and graph based measurements. In (Wang et al. 2008), the semantic relations of terms in the same semantic role are discovered by using the WordNet (Miller 1995). A tree pattern expression for extracting information from syntactically parsed text is proposed in (Choi 2011). Algorithms like PageRank (Brin and Page 1998) and HITS (Kleinberg 1999) are used in the sentence score propagation based on the graph constructed based on the similarity between sentences. Wan and Yang (2007) show that graph based measurements can also improve the single-document summarization by integrating multiple documents of the same topic.

Most of these scoring-based methods have to incorporate with the adjustment of word weights which is one of the most important factors that influence the summarization performance (Nenkova, Vanderwende, and McKeown 2006). So much work has been studied on how to extract sentences without saliency scores. Inspired by the latent semantic indexing (LSA), the singular value decomposition (SVD) is used to select highly ranked sentences for generic document summarization (Gong and Liu 2001). Harabagiu and Lacatusu (2005) analyze five different topic representations and propose a novel topic representation based on topic themes. Wang *et al.* (2008) use the symmetric non-negative matrix factorization (SNMF) to cluster sentences into groups and select sentences from each group for summarization.

The Proposed Framework

Most of the existing summarization methods aim to obtain the summary which covers the core information of the document. In this paper, we study the summarization from a data reconstruction perspective. We believe that a good summary should contain those sentences that can be used to reconstruct the document as well as possible, namely, minimizing the reconstruction error.

In this section, we describe the details of our proposed framework *Document Summarization based on Data Reconstruction* (DSDR) which minimizes the reconstruction error for summarization. The algorithm procedure of DSDR is as follows:

- After stemming and stop-word elimination, we decompose the document into individual sentences and create a weighted term-frequency vector for every sentence. All the sentences form the **candidate set**.
- For any sentence in the document, DSDR selects the related sentences from the candidate set to reconstruct the given sentence by learning a reconstruction function for the sentence.
- For the entire document (or, a set of documents), DSDR aims to find an optimal set of representative sentences to approximate the entire document (or, the set of documents), by minimizing the reconstruction error.

We denote the candidate sentence set as $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]^T$ where $\mathbf{v}_i \in \mathbb{R}^d$ is a weighted term-frequency vector for sentence i. Here notice that, we use V to represent both the matrix and the candidate set $\{\mathbf{v}_i\}$. Suppose there are totally d terms and n sentences in the document, we will have a matrix V in the size of $n \times d$. We denote the summary sentence set as $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^T$ with $m < n$ and $X \subset V$.

Given a sentence $\mathbf{v}_i \in V$, DSDR attempts to represent it with a reconstruction function $f_i(X)$ given the selected sentence set X. Denoting the parameters of f_i as $\mathbf{a}_i$, we obtain the reconstruction error of $\mathbf{v}_i$ as:

$$L(\mathbf{v}_i, f_i(X; \mathbf{a}_i)) = \|\mathbf{v}_i - f_i(X; \mathbf{a}_i)\|^2, \tag{1}$$

where $\|\cdot\|$ is the L_2-norm.

By minimizing the sum of reconstruction errors over all the sentences in the document, DSDR picks the optimal set of representative sentences. The objective function of DSDR can be formally defined as:

$$\min_{X, \mathbf{a}_i} \sum_{i=1}^{n} \|\mathbf{v}_i - f_i(X; \mathbf{a}_i)\|^2. \tag{2}$$

In the following, we will discuss two types of the reconstruction function $f_i(X; \mathbf{a}_i)$, namely, linear reconstruction and nonnegative linear reconstruction.

Linear Reconstruction

First we define the reconstruction functions $f_i(X)$ as a linear function:

$$f_i(X; \mathbf{a}_i) = \sum_{j=1}^{m} \mathbf{x}_j a_{ij}, \quad X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^T. \tag{3}$$

Then a candidate sentence $\mathbf{v}_i$ can be approximately represented as:

$$\mathbf{v}_i \approx \sum_{j=1}^{m} \mathbf{x}_j a_{ij}, \quad 1 \le i \le n.$$

Now, the reconstruction error of the document can be obtained as:

$$\sum_{i=1}^{n} \|\mathbf{v}_i - X^T \mathbf{a}_i\|^2$$

The solution from minimizing the above equation often exhibits high variance and results in high generalization error especially when the dimension of sentence vectors is smaller than the number of sentences. The variance can be reduced by shrinking the coefficients $\mathbf{a}_i$, if we impose a penalty on its size. Inspired by ridge regression (Hoerl and Kennard 1970), we penalize the coefficients of linear reconstruction error in DSDR as follows:

$$\begin{array}{ll}\min\limits_{X,A} & \sum_{i=1}^{n} \|\mathbf{v}_i - X^T\mathbf{a}_i\|^2 + \lambda\|\mathbf{a}_i\|^2 \\ \text{s.t.} & X \subset V, |X| = m \\ & A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]^T \in \mathbb{R}^{n\times m}.\end{array} \quad (4)$$

The set $\{\mathbf{x}_i\}$ includes the selected representative sentences from the original candidate sentence set V and will be used as the document summary finally. λ is the regularization parameter controlling the amount of shrinkage.

The optimization problem in Eq. (4) faces two combinatorial challenges: (1) Evaluating the best reconstruction error of one candidate sentence $\mathbf{v}_i$, we would find the optimal X with size of m out of exponentially many options. (2) The optimal set for $\mathbf{v}_i$ is usually not optimal for $\mathbf{v}_j$. So to reconstruct all the candidate sentences, we would have to search over an exponential number of possible sets to determine the unique optimal X. Actually, a similar problem that selects $m < n$ basic vectors from n candidates to approximate a single vector in the least squares criterion has been proved to be NP hard (Natarajan 1995).

The optimization problem in Eq. (4) is equivalent to the following problem (Yu, Bi, and Tresp 2006):

$$\begin{array}{ll}\min\limits_{X} & J = \mathrm{Tr}[V(X^TX + \lambda I)^{-1}V^T] \\ \text{s.t.} & X \subset V, |X| = m\end{array} \quad (5)$$

where V is the candidate sentence set, X is the selected sentence set, I is the identity matrix, and $\mathrm{Tr}[\cdot]$ is the matrix trace calculation. Please see (Yu, Bi, and Tresp 2006) for the detailed derivation from Eq. (4) to Eq. (5).

For the optimization problem (5), we use a greedy algorithm to find the approximate solution. Given the previously selected sentence set X_1, DSDR selects the next new sentence $\mathbf{x}_i \in V$ as follows:

$$\begin{array}{ll}\min\limits_{\mathbf{x}_i} & J(\mathbf{x}_i) = \mathrm{Tr}[V(X^TX + \lambda I)^{-1}V^T] \\ \text{s.t.} & X = X_1 \cup \mathbf{x}_i, \mathbf{x}_i \in V.\end{array} \quad (6)$$

Denoting $P = X_1^TX_1 + \lambda I$, Eq. (6) can be rewritten as:

$$\begin{aligned} J(\mathbf{x}_i) &= \mathrm{Tr}[V(X^TX + \lambda I)^{-1}V^T] \\ &= \mathrm{Tr}[V(P + \mathbf{x}_i\mathbf{x}_i^T)^{-1}V^T] \\ &= \mathrm{Tr}\left[VP^{-1}V^T - \frac{VP^{-1}\mathbf{x}_i\mathbf{x}_i^TP^{-1}V^T}{1 + \mathbf{x}_i^TP^{-1}\mathbf{x}_i}\right], \end{aligned} \quad (7)$$

where the Woodbury matrix identity (Riedel 1992) is applied in the second step.

Fixing the candidate sentence set V and the selected sentence set X_1, $\mathrm{Tr}[VP^{-1}V^T]$ is a constant, so the objective function is the same as maximizing the second part in the trace:

$$\max_{\mathbf{x}_i} \mathrm{Tr}\left[\frac{VP^{-1}\mathbf{x}_i\mathbf{x}_i^TP^{-1}V^T}{1 + \mathbf{x}_i^TP^{-1}\mathbf{x}_i}\right] = \frac{\|VP^{-1}\mathbf{x}_i\|^2}{1 + \mathbf{x}_i^TP^{-1}\mathbf{x}_i}. \quad (8)$$

To simplify the computation, we introduce a matrix $B = VP^{-1}V^T$. Then the index of the new sentence $\mathbf{x}_i$ can be obtained by:

$$i = \arg\max_i \frac{\|B_{*i}\|^2}{1 + B_{ii}}, \quad (9)$$

where i is the index of the new sentence $\mathbf{x}_i$ in the candidate sentence set V, B_{*i} and B_{ii} are the ith column and diagonal entry of matrix B.

Once we find the new sentence $\mathbf{x}_i$, we add it into X_1 and update the matrix B as follows:

$$\begin{aligned} B^t &= VP_t^{-1}V^T \\ &= V(P_{t-1} + \mathbf{x_i}\mathbf{x_i}^T)^{-1}V^T \\ &= B^{t-1} - \frac{VP_{t-1}^{-1}\mathbf{x_i}\mathbf{x_i}^TP_{t-1}^{-1}V^T}{1 + \mathbf{x_i}^TP_{t-1}^{-1}\mathbf{x_i}} \\ &= B^{t-1} - \frac{B_{*i}^{t-1}[B_{*i}^{t-1}]^T}{1 + B_{ii}^{t-1}}. \end{aligned} \quad (10)$$

where the matrix B^{t-1} denotes the matrix B at the step $t-1$.

Initially the previously selected sentence set X_1 is empty. So the matrix P is initialized as:

$$P_0 = \lambda I. \quad (11)$$

Then the initialization of the matrix B can be written as:

$$B^0 = VP_0^{-1}V^T = \frac{1}{\lambda}VV^T. \quad (12)$$

Algorithm 1 DSDR with linear reconstruction

```
Input:
    • The candidate data set: $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]^T$
    • The number of sentences to be selected: $m$
    • The trade off parameter: $\lambda$
Output:
    • The set of $m$ summary sentences: $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^T \subseteq V$
1:  initialize $X \leftarrow \emptyset$;
2:  $B^0 \leftarrow VV^T/\lambda$;
3:  for $t = 1$ to $m$ do
4:      for $i = 1$ to $n$ do
5:          $score(\mathbf{x_i}) \leftarrow \|B_{*i}^{t-1}\|^2/(1 + B_{ii}^{t-1})$
6:      end for
7:      $\mathbf{x}_i \leftarrow \arg\max_{\mathbf{x}_i} score(\mathbf{x}_i)$
8:      $X \leftarrow X \cup \mathbf{x}_i$
9:      $B^t \leftarrow B^{t-1} - B_{*i}^{t-1}[B_{*i}^{t-1}]^T/(1 + B_{ii}^{t-1})$
10: end for
11: return $X$;
```

We describe our sequential method for linear reconstruction in Algorithm 1. Given a document with n sentences, Algorithm 1 generates a summary with m sentences with the complexity as follows:

- $O(n^2d)$ to calculate the initialization B^0 according to Step (2).
- $O(n^2m)$ for the Step (3) to Step (10).
  - $O(n)$ to calculate $score(\mathbf{x_i})$ in Step (5)
  - $O(n^2)$ to update the matrix B in Step (9).

The overall cost for Algorithm 1 is $O(n^2(d + m))$.

Nonnegative Linear Reconstruction

The linear reconstruction optimization problem Eq. (4) in the previous section might come up with a_{ij}'s with negative values, which means redundant information needs to be removed from the summary sentence set X. To minimize the redundant information, in this section, we use the nonnegative linear reconstruction which adds nonnegative constraints on the coefficients.

Nonnegative constraints on data representation has received considerable attention due to its psychological and physiological interpretation of naturally occurring data whose representation may be parts-based in the human brain (Palmer 1977; Wachsmuth, Oram, and Perrett 1994; Cai et al. 2011). Our nonnegative linear reconstruction method leads to parts-based reconstruction because it allows only additive, not subtractive, combinations of the sentences.

For the sake of efficient optimization, following (Yu et al. 2008; Cai and He 2012),we formulate the objective function of nonnegative DSDR as follows:

$$\begin{array}{ll} \min_{\mathbf{a}_i,\beta} & J = \sum_{i=1}^{n} \left\{ \|\mathbf{v}_i - V^T\mathbf{a}_i\|^2 + \sum_{j=1}^{n} \frac{a_{ij}^2}{\beta_j} \right\} + \gamma\|\beta\|_1 \\ \text{s.t.} & \beta_j \geq 0, \quad a_{ij} \geq 0 \quad \text{and} \quad \mathbf{a}_i \in R^n, \end{array} \tag{13}$$

where $\beta = [\beta_1, \ldots, \beta_n]^T$ is an auxiliary variable to control the candidate sentences selection. Similar to LASSO (Tibshirani 1996), the L_1 norm of β will enforce some elements to be zeros. If $\beta_j = 0$, then all $a_{1j}, \ldots, a_{nj}$ must be 0 which means the j-th candidate sentence is not selected. The new formulation in Eq. (13) is a convex problem and can guarantee a global optimal solution.

By fixing $\mathbf{a}_i$'s and setting the derivative of J with respect to β to be zero, we can obtain the minimum solution of β:

$$\beta_j = \sqrt{\frac{\sum_{i=1}^{n} \mathbf{a}_{ij}^2}{\gamma}}. \tag{14}$$

Once the β is obtained, the minimization under the nonnegative constraints can be solved using the Lagrange method. Let α_{ij} be the Lagrange multiplier for constraint $a_{ij} \geq 0$ and $A = [a_{ij}]$, the Lagrange L is:

$$L = J + \mathrm{Tr}[\alpha A^T], \quad \alpha = [\alpha_{ij}].$$

The derivative of L with respect to A is:

$$\frac{\partial L}{\partial A} = -2VV^T + 2AVV^T + 2A\mathrm{diag}(\beta)^{-1} + \alpha.$$

Setting the above derivative to be zero, α can be represented as:

$$\alpha = 2VV^T + 2AVV^T - 2A\mathrm{diag}(\beta)^{-1},$$

where $\mathrm{diag}(\beta)$ is a matrix with diagonal entries of $\beta_1, \ldots, \beta_n$. Using the Kuhn-Tucker condition $\alpha_{ij}a_{ij} = 0$, we get:

$$(VV^T)_{ij}a_{ij} - (AVV^T)_{ij}a_{ij} - (A\mathrm{diag}(\beta))_{ij}a_{ij} = 0.$$

This leads to the following updating formula:

$$a_{ij} \leftarrow \frac{a_{ij}(VV^T)_{ij}}{[AVV^T + A\mathrm{diag}(\beta)]_{ij}}. \tag{15}$$

Algorithm 2 DSDR with nonnegative linear reconstruction

```
Input:
    • The candidate sentence set: $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]^T$
    • The trade off parameter: $\gamma > 0$
Output:
    • The set of the summary sentences: $X \subseteq V$
Procedure:
initialize $a_{ij}$, $\beta_j$;
initialize $X \leftarrow \emptyset$;
repeat
    $\beta_j = \sqrt{\frac{\sum_{i=1}^{n} \mathbf{a}_{ij}^2}{\gamma}}$;
    repeat
        $a_{ij} \leftarrow \frac{a_{ij}(VV^T)_{ij}}{[AVV^T + A\mathrm{diag}(\beta)]_{ij}}$;
    until converge;
until converge;
$X \leftarrow \{\mathbf{v}_i | \mathbf{v}_i \subset V, \beta_j \neq 0\}$;
return $X$;
```

The Eq. (14) and Eq. (15) are iteratively performed until convergence. For the convergence of this updating formula, we have the following Theorem 1.

Theorem 1. *Under the iterative updating rule as Eq. (15), the objective function J is non-increasing with fixed β, and that the convergence of the iteration is guaranteed.*

Proof. To prove Theorem 1, we introduce an auxiliary function as:

$$G(\mathbf{u}, \mathbf{a}_i) = \sum_{j=1}^{n} \left\{ \frac{(P\mathbf{a}_i)_j}{a_{ij}} u_j^2 - 2(VV^T)_{ij}u_j \right\}, \tag{16}$$

where $P = VV^T + \mathrm{diag}(\beta)$, and $\mathbf{u} = [u_1, \ldots, u_n]^T$ is a positive vector. $G(\mathbf{u}, \mathbf{a}_i)$ can also be identified as the sum of singular-variable functions:

$$G(\mathbf{u}, \mathbf{a}_i) = \sum_{j=1}^{n} G_j(u_j). \tag{17}$$

Let $F(\mathbf{a}_i) = \mathbf{a}_i^T P\mathbf{a}_i - 2(VV^T)_{i*}\mathbf{a}_i$, Sha *et al.* (2007) have proved that if a_{ij} updates as:

$$a_{ij} \leftarrow \arg\min_{u_j} G_j(u_j), \tag{18}$$

$G(\mathbf{u}, \mathbf{a}_i)$ converges monotonically to the global minimum of $F(\mathbf{a}_i)$.

Taking the derivation of $G_j(u_j)$ with respect to u_j and setting it to be zero, we obtain the updating formulation as:

$$a_{ij} \leftarrow \frac{a_{ij}(VV^T)_{ij}}{[AVV^T + A\mathrm{diag}(\beta)]_{ij}}, \tag{19}$$

which agrees with Eq. (15).

We can rewrite the objective function J as:

$$J = \sum_{i=1}^{n} F(\mathbf{a}_i) + \mathrm{Tr}[VV^T] + \gamma\|\beta\|_1. \tag{20}$$

Fixing β, we can obtain the minimizer of J by minimizing each $F(\mathbf{a}_i)$ separately. Since the objective function J is the sum of all the individual terms $F(\mathbf{a}_i)$ plus a term independent of $\mathbf{a}_i$, we have shown that J is non-increasing with fixed β under the updating rule as Eq. (15).

$\square$

Algorithm 2 describes the DSDR with nonnegative linear reconstruction. Suppose the maximum number of iterations for Step (4) and Step (6) are t_1 and t_2 respectively, the total computational cost for Algorithm 2 is $O(t_1(n + t_2(n^3)))$.

Experiments

In this study, we use the standard summarization benchmark data sets DUC 2006 and DUC 2007 for the evaluation. DUC 2006 and DUC 2007 contain 50 and 45 document sets respectively, with 25 news articles in each set. The sentences in each article have been separated by NIST [1]. And every sentence is either used in its entirety or not at all for constructing a summary. The length of a result summary is limited by 250 tokens (whitespace delimited).

Evaluation Metric

We use the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit (Lin 2004) which has been widely adopted by DUC for automatic summarization evaluation. ROUGE measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the peer summary (produced by algorithms) and the model summary (produced by humans). We choose two automatic evaluation methods ROUGE-N and ROUGE-L in our experiment. Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries and ROUGE-L uses the longest common subsequence (LCS) matric. ROUGE-N is computed as follows:

$$ROUGE - N = \frac{\sum\limits_{S \in Ref} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in Ref} \sum\limits_{gram_n \in S} Count(gram_n)}$$

where n stands for the length of the n-gram, Ref is the set of reference summaries. $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, and $Count(gram_n)$ is the number of n-grams in the reference summaries. ROUGE toolkit reports separate scores for 1, 2, 3 and 4-gram, and also for the longest common subsequence. Among these different scores, the unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most (Lin and Hovy 2003). Due to limited space, more information can be referred to the toolkit package.

Compared Methods

We compare our DSDR with several state-of-the-art summarization approaches described briefly as follows:

- **Random**: selects sentences randomly for each document set.

Table 1: Average F-measure performance on DUC 2006. ”DSDR-lin” and ”DSDR-non” denote DSDR with the linear reconstruction and DSDR with the nonnegative reconstruction respectively.

Algorithm	Rouge-1	Rouge-2	Rouge-3	Rouge-L
Random	0.28507	0.04291	0.01023	0.25926
Lead	0.27449	0.04721	0.01181	0.23225
LSA	0.25782	0.03707	0.00867	0.23264
ClusterHITS	0.28752	0.05167	0.01282	0.25715
SNMF	0.25453	0.03815	0.00815	0.22530
DSDR-lin	**0.30941**	**0.05427**	**0.01300**	**0.27576**
DSDR-non	**0.33168**	**0.06047**	**0.01482**	**0.29850**

Table 2: Average F-measure performance on DUC 2007. ”DSDR-lin” and ”DSDR-non” denote DSDR with the linear reconstruction and DSDR with the nonnegative reconstruction respectively.

Algorithm	Rouge-1	Rouge-2	Rouge-3	Rouge-L
Random	0.32028	0.05432	0.01310	0.29127
Lead	0.31446	0.06151	0.01830	0.26575
LSA	0.25947	0.03641	0.00854	0.22751
ClusterHITS	0.32873	0.06625	0.01927	0.29578
SNMF	0.28651	0.04232	0.00890	0.25502
DSDR-lin	**0.36055**	**0.07163**	**0.02124**	**0.32369**
DSDR-non	**0.39573**	**0.07439**	**0.01965**	**0.35335**

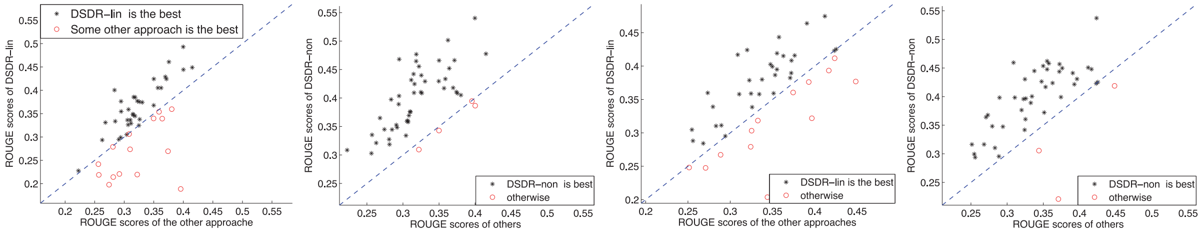
- **Lead** (Wasson 1998): for each document set, orders the documents chronologically and takes the leading sentences one by one.
- **LSA** (Gong and Liu 2001): applies the singular value decomposition (SVD) on the terms by sentences matrix to select highest ranked sentences.
- **ClusterHITS** (Wan and Yang 2008): considers the topic clusters as hubs and the sentences as authorities, then ranks the sentences with the authorities scores. Finally, the highest ranked sentences are chosen to constitute the summary.
- **SNMF** (Wang et al. 2008): uses symmetric non-negative matrix factorization(SNMF) to cluster sentences into groups and select sentences from each group for summarization.

It is important to note that our algorithm is unsupervised. Thus we do not compare with any supervised methods (Toutanova et al. 2007; Haghighi and Vanderwende 2009; Celikyilmaz and Hakkani-Tur 2010; Lin and Bilmes 2011).

Experimental Results

Overall Performance Comparison ROUGE can generate three types of scores: recall, precision and F-measure. We get similar experimental results using the three types with DSDR taking the lead. In this study, we use F-measure to compare different approaches. The F-measure of four ROUGE metrics are shown in our experimental results: ROUGE-1, ROUGE-2, ROUGE-3 and ROUGE-L. Table 1 and Table 2 show the ROUGE evaluation results on DUC 2006 and DUC 2007 data sets respectively. ”DSDR-lin” and

[1] http://www.nist.gov/index.html



(a) ROUGE scores on DUC 2006.
(b) ROUGE scores on DUC 2006.
(c) ROUGE scores on DUC 2007.
(d) ROUGE scores on DUC 2007.

Figure 1: The scores of all algorithms on each document set of DUC 2006 and DUC 2007, the black stars denote our proposed methods are best while the red circles denote otherwise. ”DSDR-lin” and ”DSDR-non” denote DSDR with the linear reconstruction and DSDR with the nonnegative reconstruction respectively.

Table 3: The associated p-values of the paired t-test on DUC 2006.

	Random	Lead	LSA	ClusterHITS	SNMF
DSDR-lin	$4.6 * 10^{-14}$	$7.1 * 10-6$	$9.2 * 10^{-14}$	$4.0 * 10^{-9}$	$9.3 * 10^{-8}$
DSDR-non	$2.6 * 10^{-25}$	$6.7 * 10^{-17}$	$2.3 * 10^{-30}$	$6.0 * 10^{-23}$	$1.8 * 10^{-25}$

Table 4: The associated p-values of the paired t-test on DUC 2007.

	Random	Lead	LSA	ClusterHITS	SNMF
DSDR-lin	$5.2 * 10^{-14}$	$1.7 * 10^{-8}$	$5.6 * 10^{-12}$	$3.4 * 10^{-10}$	$1.9 * 10^{-9}$
DSDR-non	$2.5 * 10^{-17}$	$8.0 * 10^{-13}$	$1.4 * 10^{-14}$	$7.9 * 10^{-15}$	$1.1 * 10^{-14}$

”DSDR-non” denote DSDR with the linear reconstruction and DSDR with the nonnegative reconstruction respectively.

As shown by the highest ROUGE scores in bold type from the two tables, it is obvious that DSDR takes the lead followed by ClusterHITS. ClusterHITS considers topics as hubs and sentences as authorities where hubs and authorities can interact with each other. So that the correlations between topics and sentences can improve the quality of summary. Besides, selecting sentences randomly is a little better than just selecting the leading sentences. Among all the seven summarization algorithms, LSA and SNMF show the poorest performance on both data sets. Directly applying SVD on the terms by sentences matrix, summarization by LSA chooses those sentences with the largest indexes along the orthogonal latent semantic directions. Although SNMF relaxes the orthogonality, it relies on the sentence pairwise similarity. Whereas, our DSDR selects sentences which span the intrinsic subspace of the candidate sentence space. Such sentences are contributive to reconstruct the original document, and so are contributive to improve the summary quality. Under the DSDR framework, the sequential method of linear reconstruction is suboptimal, so DSDR-non outperforms DSDR-lin.

Evaluations on Different Document Sets In Figure 1, we illustrate the ROUGE-1 scores for each document set from DUC 2006 and DUC 2007 respectively. In each panel, the vertical axis describes the scores of the DSDR approach and the horizontal axis contains the best scores of other methods. The black stars indicate that DSDR gets the best scores on the corresponding document sets while the red circles suggest the best scores are obtained from other methods. It can be obviously observed that both the proposed reconstruction methods are better than others, since the number of black stars are much more than that of red circles in each panel.

To check whether the difference between DSDR and other approaches is significant, we perform the paired t-test between the ROUGE scores of DSDR and that of other approaches on both data sets. Table 3 and Table 4 show the associated p-values on DUC 2006 and DUC 2007 data sets respectively. The test at the 99% confidence interval demonstrates that our proposed framework can obtain very encouraging and promising results compared to the others.

Conclusion

In this paper, we propose a novel summarization framework called *Document Summarization based on Data Reconstruction* (DSDR) which selects the most representative sentences that can best reconstruct the entire document. We introduce two types of reconstruction (linear and nonnegative) and develop efficient optimization methods for them. The linear reconstruction problem is solved using a greedy strategy and the nonnegative reconstruction problem is solved using a multiplicative updating. The experimental results show that out DSDR (with both reconstruction types) can outperform other state-of-the-art summarization approaches. DSDR with linear reconstruction is more efficient while DSDR with nonnegative reconstruction has better performance (by generating less redundant sentences). It would be of great interests to develop more efficient solution for DSDR with nonnegative reconstruction.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No: 61125203, 61173185, 90920303), National Basic Research Program of China (973 Program) under Grant 2011CB302206, Zhejiang Provincial Natural Science Foundation of China (Grant No: Y1101043) and Foundation of Zhejiang Provincial Educational Department under Grant Y201018240.

References

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems* 30(1-7):107–117.

Cai, D., and He, X. 2012. Manifold adaptive experimental design for text categorization. *IEEE Transactions on Knowledge and Data Engineering* 24(4):707–719.

Cai, D.; He, X.; Ma, W.-Y.; Wen, J.-R.; and Zhang, H. 2004. Organizing WWW images based on the analysis of page layout and web link structure. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*.

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph regularized non-negative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8):1548–1560.

Celikyilmaz, A., and Hakkani-Tur, D. 2010. A hybrid hierarchical model for multi-document summarization. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*.

Choi, Y. 2011. Tree pattern expression for extracting information from syntactically parsed text corpora. *Data Mining and Knowledge Discovery* 1–21.

Conroy, J., and O'leary, D. 2001. Text summarization via hidden markov models. In *Proc. of the 24th ACM SIGIR*, 407. ACM.

Gong, Y., and Liu, X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. of the 24th ACM SIGIR*, 19–25. ACM.

Haghighi, A., and Vanderwende, L. 2009. Exploring content models for multi-document summarization. In *Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Harabagiu, S., and Lacatusu, F. 2005. Topic themes for multi-document summarization. In *Proc. of the 28th ACM SIGIR*, 209. ACM.

He, X.; Cai, D.; Wen, J.-R.; Ma, W.-Y.; and Zhang, H.-J. 2007. Clustering and searching www images using link and page layout analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* 3(1).

Hoerl, A., and Kennard, R. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 55–67.

Hu, M.; Sun, A.; and Lim, E. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In *Proc. of the 31st ACM SIGIR*, 291–298. ACM.

Huang, Y.; Liu, Z.; and Chen, Y. 2008. Query biased snippet generation in xml search. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46(5):604–632.

Lin, H., and Bilmes, J. 2011. A class of submodular functions for document summarization. In *The 49th ACL-HLT, Portland, OR, June*.

Lin, C., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 71–78. Association for Computational Linguistics.

Lin, C. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. of the WAS*, 25–26.

Miller, G. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Natarajan, B. 1995. Sparse approximate solutions to linear systems. *SIAM journal on computing* 24(2):227–234.

Nenkova, A.; Vanderwende, L.; and McKeown, K. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proc. of the 29th ACM SIGIR*, 580. ACM.

Palmer, S. 1977. Hierarchical structure in perceptual representation. *Cognitive Psychology* 9(4):441–474.

Park, S.; Lee, J.; Kim, D.; and Ahn, C. 2007. Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization. *SOFSEM: Theory and Practice of Computer Science* 761–770.

Riedel, K. 1992. A sherman-morrison-woodbury identity for rank augmenting matrices with application to centering. *SIAM Journal on Matrix Analysis and Applications* 13(2):659–662.

Sha, F.; Lin, Y.; Saul, L.; and Lee, D. 2007. Multiplicative updates for nonnegative quadratic programming. *Neural Computation* 19(8):2004–2031.

Shen, D.; Sun, J.; Li, H.; Yang, Q.; and Chen, Z. 2007. Document summarization using conditional random fields. In *Proc. of IJCAI*, volume 7, 2862–2867.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Toutanova, K.; Brockett, C.; Gamon, M.; Jagarlamudi, J.; Suzuki, H.; and Vanderwende, L. 2007. The pythy summarization system: Microsoft research at duc 2007. In *Proc. of DUC*, volume 2007.

Turpin, A.; Tsegay, Y.; Hawking, D.; and Williams, H. E. 2007. Fast generation of result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*.

Wachsmuth, E.; Oram, M.; and Perrett, D. 1994. Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cerebral Cortex* 4(5):509.

Wan, X., and Yang, J. 2007. CollabSum: exploiting multiple document clustering for collaborative single document summarizations. In *Proc. of the 30th annual international ACM SIGIR*, 150. ACM.

Wan, X., and Yang, J. 2008. Multi-document summarization using cluster-based link analysis. In *Proc. of the 31st ACM SIGIR*, 299–306. ACM.

Wang, D.; Li, T.; Zhu, S.; and Ding, C. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proc. of the 31st ACM SIGIR*.

Wasson, M. 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In *Proc. of the 17th international conference on Computational linguistics-Volume 2*.

Yu, K.; Zhu, S.; Xu, W.; and Gong, Y. 2008. Non-greedy active learning for text categorization using convex ansductive experimental design. In *Proc. of the 31st ACM SIGIR*, 635–642. ACM.

Yu, K.; Bi, J.; and Tresp, V. 2006. Active learning via transductive experimental design. In *Proc. of the 23rd ICML*, 1081–1088. ACM.