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).

□

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