4.2 Compared Methods

As our framework is fully unsupervised, we do not compare it with supervised methods. Document summarization based on data reconstruction, along with the gradient descent algorithm in He et al. [2012]'s work, naturally becomes the direct baseline for comparison, denoted as DSDR. We denote our sparse optimization formulation of $\ell_{2,1}$ regularization (1) as SpOpt-$\ell_{2,1}$ (the version with diversity term denoted as SpOpt-Δ), and the compressive solution as SpOpt-comp. We also report experimental results on the same datasets (if reported in their paper) given by several recent state-of-the-art unsupervised systems, including matrix-factorization [Wang *et al.*, 2008], the document-sensitive graph model DsR-Q [Wei *et al.*, 2010], the bi-mixture probabilistic latent semantic analysis method (BI-PLSA) [Shen *et al.*, 2011], graph based multimodality learning [Wan and Xiao, 2009], and a very recent work on two-level sparse representation [Liu *et al.*, 2015].

We denote a weeker baseline that extract the leading sentence from each document as LEAD. This baseline is included in the official evaluation. PEER 24 and PEER 15 are the DUC 2006/2007 participants with highest ROUGE performance respectively. For DUC 2007 main task, there is an extractive baseline named CLASSY04 that ignores topic narrative but achieved the highest performance in general multi-document summarization task of DUC 2004.

4.3 Evaluation and Results

We run the commonly used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [6] metrics for summarization tasks [Lin and Hovy, 2003; Lin, 2004]. The ROUGE metric automatically scores a candidate summary by measuring the amount of ngram overlap between this candidate and manually-created reference summaries. ROUGE-k uses k-gram overlap.

We report ROUGE recall of summaries generated by all systems in comparison. These results are listed in Table 1.

SpOpt-$\ell_{2,1}$ is still superior to DSDR, even though the motivations are similar and the optimization problems are both convex. One probable reason is that the $\ell_{2,1}$ form is directly optimizing coefficient matrix with row sparsity, while the original formulation tries to guide this sparsity indirectly with another group of variables.

Our unsupervised compressive framework outperforms all other unsupervised systems in comparison and achieves very competitive results against the best peer system in DUC 2006/2007. We also observe that adding the sentence dissimilarity term (-Δ) can indeed improve performance.

We also ask three annotators (who are not among the authors of this paper and are fluent in English) to carry out human evaluation for the generated summarization in terms of different aspects of quality, including Grammaticality (GR), Non-Redundancy (NR), Referential Clarity (RC), Topic Focus (TF) and Structural Coherence (SC), similar to the evaluation in DUC 2006 [Dang, 2006]. Each aspect is rated with scores from 1 (poor) to 5 (good). This evaluation is performed on the same random sample of 10 topics in DUC 2006.

Table 1: Results of ROUGE evaluation on DUC 2006/2007

DUC 2006	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
LEAD	0.30217	0.04947	0.09788
MatrixFacto.	0.39551	0.08549	N/A
DsR-Q	0.39550	0.08990	N/A
BI-PLSA	0.39384	0.08497	N/A
MultiModal.	0.40503	0.08545	N/A
[Liu *et al.*, 2015]	0.34034	0.05233	0.10730
DSDR	0.37695	0.07312	0.11678
SpOpt-$\ell_{2,1}$	0.39069	0.08336	0.13791
SpOpt-Δ	0.39962	0.08682	0.14227
SpOpt-comp	0.41331	0.09136	0.15046
SpOpt-comp-Δ	**0.41534**	0.09455	0.15310
PEER 24	0.41095	**0.09551**	**0.15523**

DUC 2007	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
LEAD	0.31250	0.06039	0.10507
CLASSY04	0.40562	0.09382	0.14641
DsR-Q	0.42190	0.11230	N/A
MultiModal.	0.42609	0.10438	N/A
[Liu *et al.*, 2015]	0.35399	0.06448	0.11669
DSDR	0.39765	0.08679	0.13732
SpOpt-$\ell_{2,1}$	0.41833	0.10627	0.16304
SpOpt-Δ	0.42360	0.11109	0.16474
SpOpt-comp	0.44517	0.12025	0.17072
SpOpt-comp-Δ	**0.44607**	**0.12454**	0.17429
PEER 15	0.44515	0.12448	**0.17715**

The evaluated summaries include the summaries produced by the compressive framework and those from their extractive counterpart, i.e. SpOpt-$\ell_{2,1}$ that only involves sentence selection and extraction. Also one of the official reference summaries generated by human is also in comparison and can be treated as an upper bound for all aspects. The average score and standard deviation for each metric are displayed in Table 2.

From the comparison between compressive summarization and the extractive version, there exist slight improvements of non-redundancy. This exactly matches what we can expect from sentence compression that keeps only important part and drop redundancy. We also observe certain amount of improvements on structural coherence. This may be a result of iterative joint optimization of sentence and word selection that simultaneously considers more global and local coherence.

There may be multiple reasons behind the loss of grammaticality, such as errors of dependency outputs given by the dependency parser, the incompleteness of constraint sets, etc.

Table 3 shows an example summary produced by our compressive system. Words in grey are not selected in the final compressed summaries. In most cases, the removed phrases do not hurt the overall readability of the summary.

In the experiments, the time consumption of our methods is significantly less than the original reconstruction formulation with gradient descent algorithm. Even for the compressive case, the acceleration ratio achieves more than 60 under the same single machine computing environment.

[6] Parameter options of ROUGE are set to be consistent with official evaluation of corresponding tasks at DUC 2006 and DUC 2007.