corpus	# documents	# summaries	summ len	doc size (KB)	type
DUC'02	533	2-3	100	1-20	SD
DUC'04	50	4	100	20-89	MD (50x10)
DUC'07	23	4	250	28-131	SD

Table 2: Corpora statistics.

by the ROUGE-1 and ROUGE-2 (Lin, 2004) recall scores, with the word limit indicated in Table 2, without stemming and stopword removal.

The results of the introduced algorithm (Gamp) may be affected by several input parameters: minimum support ($Supp$), codes limit (C), and the maximal gap allowed between frequent words (Gap). In order to find the best algorithm settings for a general case, we performed experiments that explored the impact of these parameters on the summarization results. First, we experimented with different values of support count in the range of $[2, 10]$. The results show that we get the best summaries using the sequences that occur in at least four document sentences.

A limit on the number of codes is an additional parameter.We explored the impact of this parameter on the quality of generated summaries. As we could conclude from our experiments, the best summarization results are obtained if this parameter is set to the maximal number of words in the summary, W. Consequently, we used 100 codes for summarizing DUC 2002 and DUC 2004 documents and 250 codes for DUC 2007 documents.

The maximal gap ratio defines a pattern for generating the frequent sequences and has a direct effect on their structure and number. Our experiments showed that allowing a small gap between words of a sequence helps to improve slightly the ranking of sentences, but the improvement is not significant. Thus we used $Gap = 0.8$ in comparative experiments for all corpora. The resulting settings for each corpus are shown in Table 3.

Corpus	Supp	Max. codes	Gap
DUC 2002 and 2004	4	100	0.8
DUC 2007	4	250	0.8

Table 3: Best settings.

We compared the Gamp algorithm with the two known unsupervised state-of-the-art summarizers denoted by Gillick (Gillick and Favre, 2009) and McDonald (McDonald, 2007). As a baseline, we used a very simple approach that takes first sentences to a summary (denoted by TopK). Table 4 contains the results of comparative evaluations. The best scores are shown in bold. Gamp outperformed the other methods on all datasets (using ROUGE-1 score). The difference between the scores of Gamp and Gillick (second best system) on DUC 2007 is highly significant according to the Wilcoxon matched pairs test. Based on the same test, the difference of scores obtained on the DUC 2004 is not statistically significant. On the DUC 2002, Gamp is ranked first, with insignificant difference from the second best (McDonald's) scores. Based on this result, we can conclude that MDL-based summarization using frequent sequences works better on long documents or multi-document domain. Intuitively, it is a very logical conclusion, because single short documents do not contain a sufficient number of frequent sequences.

It is noteworthy that, in addition to the greedy approach, we also evaluated the global optimization with maximizing coverage and minimizing redundancy using Linear Programming (LP). However, experimental results did not provide any improvement over the greedy approach. Therefore, we report only the results of the greedy solution.

	ROUGE-1 Recall			ROUGE-2 Recall		
Algorithm	DUC'02	DUC'04	DUC'07	DUC'02	DUC'04	DUC'07
Gamp	**0.4421**	**0.3440**	**0.3959**	0.1941	**0.0829**	**0.0942**
Gillick	0.4207	0.3314	0.3518	0.1773	0.0753	0.0650
McDonald	0.4391	0.2955	0.3500	**0.1981**	0.0556	0.0672
TopK	0.4322	0.2973	0.3525	0.1867	0.0606	0.0706

Table 4: Comparative results.

4 Conclusions

In this paper, we introduce a new approach for summarizing text documents based on their Minimal Description Length. We describe documents using frequent sequences of their words. The sentences with the highest coverage of the best compressing set are selected to a summary. The experimental results show that this approach outperforms other unsupervised state-of-the-art methods when summarizing long documents or sets of related documents. We would not recommend using our approach for summarizing single short documents which do not contain enough content for providing a high-quality description. In the future, we intend to apply the MDL method to keyword extraction, headline generation, and other related tasks.

Acknowledgments

This work was partially supported by the U.S. Department of the Navy, Office of Naval Research.