

A submodularity based game theory framework for multi-document summarization

Faizaan Ahmad

M. Tech, Computer Science Department, Jamia Hamdard, New Delhi-110062

E-mail : jahan.rahmat@gmail.com

Manuscript received online 05 January 2020, accepted on 16 May 2020

Abstract : The explosion of a large amount of data over the internet has made it inconceivable to effectively forage relevant information from it. To overcome the above problem, a fundamental tool known as summarization techniques have been used for understanding and organizing such large datasets. One approach of summarization widely discussed in the literature is called multi-document summarization, where extracts of texts are selected to best represent the essence of a document. A game theory framework for multi-document summarization is proposed and a combination of three similarity measures is utilized to capture similarity between textual units. The proposed framework ensures that the summary obtained is well covered and informative. Results of the proposed approach were evaluated with ROUGE evaluation metric on DUC and TAC datasets against human-generated summaries and it outperformed other competitive methods.

(Keywords : Multi-document summarization, game theory, similarity measures)

Introduction

The digital era has given rise to the massive amount of digital information which is increasing at an enormous rate¹, with most of the generated data being in the form of text. This makes it completely inconceivable to manually forage useful information from the internet and underscores the use of automatic methods to filter data according to specific needs. Text summarization² aims to selectively extract useful content from an information source (also called document), thereby, presenting the most important relevant information to the application according

to their needs³. To capture the gist of the document, is the prime aim of automatic text summarization, without jeopardizing the effective content present in it⁴. It is also required to maximize the coverage of the summarizable content, so that the essence obtained is not specific to any special region of the document. Finally, text summarization aims at removing information redundancy in the document by extracting only the important content from it. Summarization can be categorized into two domains i.e. multi-document summarization(mds) and single document summarization(sds) based on the number of documents [5]. Single document summarization aims at summarizing a single document with the help of available techniques such as Random Indexing[6], TextRank[7]etc. Multi-document summarization(mds) works quite similar to the single document summarization approach, except for the fact, that here a set of documents of similar contextual nature is considered as a whole, rather than each document as a different entity[8]. In case of multi-document summarization, due to the dispersion of information in the documents, the complexity of summarization increases and the redundancy of the information may also increase, as the documents share similar ideas and rise from the same contextual background. Moreover, a good multi document summarization system aims at keeping the quality of summary high, by reducing the redundant information seeping out in the summary, while performing summarization.

There are two components of multi-document summarization: abstractive and

extractive summarization. In abstractive summarization, summaries are formed from the significant portions of the document, whereas extractive summarization focuses on finding salient sentences that are used to form the summary of the document. Extractive summarization with sentence selection is the focus of this paper. It considers both redundancies among sentences and salience score of sentences and accordingly, sentences are chosen to generate a summary. Extractive summarization can be further classified into supervised and unsupervised.

The former approach is a two-stage process i.e. training and testing and generation of a good summary requires a lot of training data. During the training phase, the input documents and their corresponding summary is fed into the system, which is then further used by the system to classify them into non-summary and summary sentences. In the past, researchers^{9,10, 11, 12} have extensively adopted supervised and unsupervised environment for summarization using neural networks, random forest, clustering etc. Many graph-based summarization methods have been proposed, where documents are represented in the form of graphs, that use topology of the graph to identify salient elements from the underlying graph¹³.

The motivation behind using game theory is to incorporate submodularity concept and model nodes of the graph as rational players of an alliance game, where the aim of each player is to maximize its information spread. The global objective is to obtain a good summary that contains minimum redundancy, maximum coverage and is highly diversified.

1.1 Contributions and Outline of the Paper

This paper proposes a versatile game theory framework for summarization, where initially mds task is mapped as budgeted maximum coverage problem(bmcp)¹⁴, that is often encountered in job scheduling, social network etc.

A brief summary of the contributions of the proposed work is described below:

1. The proposed approach uses a combination of statistical, linguistic and co-reference similarity to capture similarity between two textual units.
2. A novel sentence ranking algorithm based on game theory concept is proposed to solve different mds tasks.

Experiments are conducted on DUC02, DUC04, DUC05, DUC06, TAC08A and TAC08B datasets and ROUGE-2 recall metric is used to evaluate the performance. A remarkable performance is achieved by the proposed approach where it outperformed DUC and TAC competitors and other methods. The remaining paper is organized in the following way: Section 2 discusses definitions and preliminaries, succeeded by section 3, highlighting the related work carried out in the domain. Section 4, describes in detail the framework for multi-document summarization. In section 5, the focus is on the evaluation of the proposed work carried out on DUC and TAC datasets followed by conclusion discussed in section 6.

2. Definitions and Preliminaries

2.1 Cooperative game theory

Game theory framework models strategic interaction among players of an alliance game where each player wants to maximize its payoff in terms of future favor. The analysis of a cooperative game can be made using solution concepts, which is basically a way of dividing the total profit obtained from an alliance game in a fair manner. We proceed with giving a brief introduction about the cooperative game. Numerous solution concepts have been defined in cooperative game such as nucleus, Shapley value, core etc. But we will only discuss Shapley value since it is related to the proposed work.

A cooperative game $[M; p]$ is a pair of two parameters $(M; p)$, where M is a non-empty and finite set such that $M = \{1, 2, 3, \dots, m\}$ and $p(V)$ defines the profit obtained from any alliance such that $V \subseteq M$. The Shapley value $[p]$ of an alliance game represents the average weighted marginal contribution made by each player during the game. Mathematically, it can be defined in the following way described below.

The Shapley value of a node $q \in M$ in a network is given:

$$\phi_q = \frac{1}{m!} \sum_{\eta \in \Omega} [p(D_q(\eta) \cup \{q\}) - p(D_q(\eta))]$$

where m is the total number of nodes in a network, Ω is a set of $m!$ permutations on M , $p(V)$ is a set of real valued mapping of contribution by a set of nodes and is defined as $p(D \cup \{q\}) - p(D)$; $D \subseteq M - \{q\}$, η is a unique permutation of nodes selected by a subset D ($D \subseteq M_q$ and $q \in D$).

2.2 Sub modularity

Game theory is based on the concept of sub modularity that possesses decreasing marginal utility value:

$$f(R \cup \{a\}) - f(R) \leq f(S \cup \{a\}) - f(S)$$

where $S \subseteq R$, S and $R \subseteq T$ and $a \in T - R$. According to this property, the difference in the incremental value of the function f that a single element makes when added to an input set decreases as the size of the input set increases.

3. Problem Setup

The proposed research work focusses on sub modular function maximization under budget constraint and is defined as:

$$\max_{S \subseteq D} \{ f(S) : \sum_{s \in S} \text{cst}_s \leq B_g \}$$

where $f(\cdot)$ represents the sub modular function to achieve summary quality, D comprises of the sentences of the documents, S is the generated summary, cst denotes the cost associated with

selecting sentence s_m , within budget constraint B_g . From document summarization perspective, it is often required that the generated summary should be of limited length. To be specific, B_g represents the number of terms contained in summary S_y , and cst_m represents the number of words or terms in sentence s_m . Sub modular optimization can be beneficial, only if the objective function that measures summary quality is also sub modular. There are two ways to apply sub modularity optimization in any domain. The first way is to apply it forcefully that may result in objective function performing poorly. The second way is to apply it in domains where sub modularity exists naturally. Extractive summarization is one of the domain where sub modularity exists naturally that will be seen further. Since sub modularity is well defined for graphs and can be used to measure the summary quality. As mentioned in section 1, in the proposed approach, graph model is adopted to capture the interactions between textual units and sub modularity optimization is used to measure summary quality. Given a set of multiple documents, they are concatenated to form a single document D . Then, this document D is converted into a graph $G = (N; E)$, where N represents the sentences of the documents and E denotes the weighted edge $w(s_m; s_n)$ between two textual units s_m and s_n of N . Graph cut function is a sub modular function that is used to compute the similarity of generated summary S_y to other documents $D \in \mathcal{D}$.

$$f(S_y) = \sum_{s_m \in D \setminus S_y} \sum_{s_n \in S_y} w(s_m, s_n)$$

Now from bmc perspective, let budget B_g represent the total number of words that occur in summary S_y and cost associated with each sentence be represented by cst , which is the number of words that appear in a sentence s_m of a document D . The aim of a good quality summary should be to (i) minimize redundancy and (ii) maximize coverage, and to generate a summary within budget. One of the approach, that is widely used, to achieve both purpose at the same time is MMR (Maximum Marginal Relevance)⁸. To

measure the quality of generated summary, we use an approach similar to MMR [8] and is given as:

$$f(S_y) = \frac{\sum_{sm \in D \setminus S_y} \sum_{sn \in S_y} w(sm, sn)}{\sum_{sm, sn \in S_y, sm \neq sn} w(sm, sn)}$$

where D is the document set and $w(sm; sn)$ is the normalized sum of the three similarity scores discussed in eq.7,8,9. The first term of the above equation is used to ensure information coverage and second term ensures redundancy removal. One of the property possessed by submodular functions is, they are closed for linear combination. Both terms of the above equation is submodular and hence, using this property we can conclude that $f(S_y)$ is also submodular.

4. Related Work

4.1 Multi-document summarization tasks

The goal of MDS is to create non-redundant, diversified summary representing the gist of the original document. Usually, the summarizer utilizes the underlying document to compute the saliency score of the sentence, based on semantic or statistical similarity. Keeping in view, the requirements of the user, there are three ways to categorize multi-document summarization: generic, query-focussed and update summarization. Many researchers have proposed submodular function based methods for query-focussed¹⁵, update, and generic summarization¹⁶. A sub-modularity based algorithm was proposed by Lin¹⁷ for generic summarization. The inspiration for the proposed approach comes from their work. However, cosine similarity based submodular functions were proposed to capture similarity between two textual units. But in the proposed approach, a combination of statistical, linguistic and co-reference similarity measures based submodular function is proposed. Moreover, a game theory framework is proposed for different multi-document summarization tasks.

1. Generic summarization- Generic summarization aims to produce a compact summary from the overall documents content. In

case of generic summarization, rank of the sentences are computed based on statistical features and then accordingly, salient sentences are extracted. A well known centroid based algorithm known as MEAD was proposed by¹⁸, for document summarization, that generated salient sentences based on inter-sentence and sentence level features. Different methods based on latent information¹³ and probabilistic approaches¹⁹ were explored by research community.

2. Query-focussed summarization- In query focussed summarization, a compact summary is generated based on the initial search query given by user. Some works of²⁰ proposed for generic summarization, was extended by incorporating query information. A query focussed summarizer was proposed by²¹, that utilized both relatedness between sentences and relationship between query and sentence to generate compact summary. Different probabilistic approaches have been proposed by researchers [8], based on different concepts of queries and document generation.

3. Update summarization - Update summarization aims at creating a compact summary of the document corpus, under the hypothesis that the user has prior knowledge of a set of topics. In update summarization, the user/reader is made aware of new information about a particular topic. One of the main issues that arise in update summarization is the occurrence of redundant information that has already occurred in the main summary. Maximum MR[8], one of the well-known method is used to remove the redundancy, occurring in summarization task. Other methods to deal with update summarization tasks, include information distance based method²², graph-based method²³, semantic-based methods²⁴, and submodularity based methods²⁵.

5. Proposed framework

This section discusses in detail about the game theory framework for mds. The

procedure for extractive summarization is divided into following modules:

1. Graph model: an explanation about the creation of nodes and weighted edges in the proposed approach using a graph model.
2. Submodularity modeling: the submodularity existing in the sentences of the underlying document is explored and the mds task is mapped as bmcp¹⁴.
3. Salient sentence selection: Shapley value²⁶ is used to extract salient sentences and it is further optimized using the proposed algorithm. A comprehensive explanation about every module is given in following subsections:

5.1 Graph Model

Three dimensions of the graph are used in the proposed system. The nodes correspond to the sentences present in the documents and three different kinds of edges are used to incorporate more than one similarity measure. The actual weight on the edge is the normalized sum of the three similarity scores, discussed below:

1. **Statistical similarity** - A bag of words model is used to create multi-dimensional vector for every sentence of the document. Then, to measure the extent of content overlap between two sentences, cosine similarity (cosim) is used. The formula for computing cosine similarity between two sentence vectors s_i and s_j is given below:

$$\text{cosim}(s_i; s_j) = (S_i \cdot S_j) / (|S_i| |S_j|)$$

where s_i and s_j are weighted term-frequency vectors of sentences s_i and s_j .

2. **Semantic similarity**: The relations such as hypernyms, synonyms, and hyponyms in sentences are measured by semantic similarity. Calculation of semantic similarity between two

sentence can be accomplished in following steps:

- (a) The nouns of the sentences are represented as word vectors.
- (b) Between two sentences, calculation of semantic similarity score is done, for each pair of words.
- (c) Similarity scores are obtained by summing up and normalizing the result.

Wordnet is used to compute the semantic similarity for every pair of words in a sentence²⁷. For considering the similarity, only nouns in the sentences are considered. If $d_{\text{wordnet}}(x; y)$ represents the distance between words x and y , then the semantic similarity between sentences S_m and S_n is calculated as:

$$W_{\text{semantic}}(S_m; S_n) = \sum_{x \in S_m, y \in S_n} d_{\text{wordnet}}(x, y)^{-1}$$

Co reference resolution: The process of identifying different nouns representing the same entity is called co reference resolution²⁸. In the proposed work, Stanford Core NLP was used to establish co reference. When a co reference is found between two sentences, an edge is drawn between them, otherwise they are kept disjoint. In other words,

$$w_{\text{co reference}}(S_m; S_n) = 1$$

5.2 Submodularity Modeling

Let us assume that a collection of set E is given, with a budget B_g and associated costs cst . The objective of bmcp is to find $E_0 \subseteq E$, such that the total weight of elements in E_0 is maximized and the total cost of sets in E_0 does not exceed B_g . BMCP is NP-hard problem, but a greedy approach has been proposed by¹⁴, that selects those salient sentences, that make the maximum marginal contribution within budget B_g and take $(1 - (1e))$ approximation time. The selection of salient sentences is determined by the submodularity. Now from bmcp and submodularity perspective, mds can be modeled as:

1. Generic summarization -Let budget B_g represent the total number of words, that would occur in summary S_y and cost associated with each sentence be represented by cst , which is the number of words that appear in a sentence sm of a document D . The quality increase in summary occurs with the addition of one candidate sentence and the cost associated with each sentence. The quality incurred by the summary is defined in following way:

$$f(S_y) = \sum_{sm \in D \setminus S_y} \sum_{sn \in S_y} w(sm, sn) - \sum_{sm, sn \in S_y, sm \neq sn} w(sm, sn)$$

where D is the document set and $w(sm; sn)$ is the normalized sum of the three similarity scores discussed. The first term of the above equation is used to ensure information coverage and the second term ensures redundancy removal. One of the property possessed by submodular functions is, they are closed for linear combination. Both terms of the above equation is submodular and hence, using this property we can conclude that $f(S_y)$ is also submodular.

Let k be the sentence that is selected for summary, then the quality increase of the summary can be represented as:

$Q(k) = f(S_y \cup \{k\}) - f(S_y)$ (10) The quality increase function is submodular and its proof is given in³⁸.

2. Query-focussed Summarization(QFS) -

The problem of QFS can be formulated in the following way:

A query q is given by user and a dataset D is given, then the quality of the summary is defined as:

$$f(S_y; q) = f(g) + P_{sm2Sy} w(q; sm)$$

where $f(g)$ represents general information coverage explained in eq.9, except for the last component, that computes the information coverage based on given query. The quality

function is submodular, since it is closed for linear combination.

3. Update summarization(US) - This problem can be formulated as: A query q is given by user, and two sets of datasets are presented D_h : history dataset and D_u : update dataset. The aim of US is to generate summary S_u from document D_u , which is relevant to the query q , given D_h .

5.3 Sentence scoring methodology

To identify salient nodes in the underlying graph, sentences are scored according to their Shapley values²⁶, which is used to compute the marginal contribution of every node made during information diffusion process. A brief description about the calculation of Shapley value with terminologies has been discussed in section 1. We proceed with the explanation of algorithm 1, which comprises of two main steps:

1. For every node in the underlying network, Shapley value is computed and accordingly, a rank list $r[i]$ of nodes is prepared in non-increasing order.
2. From the rank list, the top nodes are chosen using algorithm 2.

Algorithm 1 Shapley value

Let n be the nodes in graph G , $R = 10000$ iterations. Also let $\phi[a]$ contain the Shapley values of all nodes.

Let t be the temporary variable Mc the marginal contribution of each node and reflects the change in coverage due to the addition of node a in the set of activated nodes.

procedure Shapley Value(G)

for $a = 0$ to n **do**

$\Phi[a] \leftarrow 0$

$Mc[a] \leftarrow 0$

end for

for $b = 1$ to z **do**

for $a = 1$ to n **do**

```

t[a]=0
end for
for c = 1 to R do
  Random threshold is assigned to nodes
  for a = 1 to n do
    t[a] t[a] + [v(Da( $\pi$ b)) [ fag - v(Da( $\pi$ b))]]
  end for
end for
for a = 1 to n do
  Mc[a] Mc[a] + t[a]=R
end for
end for
for a = 1 to n do
   $\Phi$ [a] Mc[a]=z
end for
return  $\phi$ 
end procedure

```

The above two steps are described as follows: if a node is able to adopt the information, then it is said to be active. It is assumed that a node continues to be active later, once it has become active. If a node a is able to activate more number of nodes as compared to node b, then it is said to be more influential. The salient nodes of the network can be determined, if the activation capability of the node is known, in the underlying diffusion process. During the information diffusion process, the marginal contribution made by each node is computed and accordingly, salient nodes are identified. For calculating marginal contribution made by nodes, Shapley value solution concept of the game theory is used. The Shapley value calculation is computationally a NP-hard problem, so a sample-based approach is used for the calculation of Shapley values of nodes that works in polynomial time. The sample based approach would provide approximate Shapley values for nodes z (where cardinality of z is a polynomial of n) permutation of the nodes are randomly generated in the network to achieve the purpose. Let $z = j\theta j$ so that $z = O(n)$. This sampled set θ is used, to compute approximate Shapley values as explained by algorithm 1.

In this algorithm:

1. In $\pi(b)$, let $Da(\pi b)$ represent the set of nodes that occur before node a .
2. Mc reflects the change in coverage due to the addition of node a in the set of activated nodes.
3. $\Phi(a)$ represents the Shapley value of node.

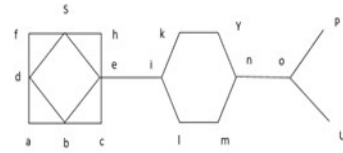


Fig. 1: Selection of silent influential nodes

We proceed with an explanation of algorithm 1, that discusses the procedure for scoring sentences. Let π , be a unique permutation of nodes and is a subset of z and $\pi(p)$ be a representative of node p in the permutation δ . Now each node's contribution to the spread of influence is computed according to their order dictated by π . Initially, all nodes are considered to be inactive in the network and threshold is assigned to each node randomly. The first stage begins with the activation of $\pi(1)$, due to which more number of nodes are activated in the network and this counts for the contribution of $\pi(1)$. The second stage proceeds with the activation of $\pi(2)$ and its contribution sums to 0, if it was already activated by $\pi(1)$. Otherwise, $\pi(2)$ is activated and the number of nodes that get activated due to $\pi(2)$ comprises of the contribution of $\pi(2)$. Likewise, it is continued upto $\pi(n)$ and the procedure is repeated using the same π , for R times (10; 000 times). Furthermore, in the set of sampled permutations, the above step is repeated for each permutation and the average contribution of each node is calculated, that determines the information spread in the underlying network. Finally, a rank list $r[i]$ is constructed that constitutes of nodes arranged in decreasing order according to their Shapley values.

5.4 Selection of salient influential nodes (Proposed algorithm)

Selection of salient influential nodes in the underlying graph reduces to the problem of finding top nodes within budget B_g , that would comprise of summary S of document D . To select topnodes, a naive approach would be to select nodes that make maximum marginal contribution, i.e., those nodes which have highest Shapley values, but this may lead to the selection of longer nodes, since they contain more terms and contribute more towards marginal contribution. To avoid such situation, marginal contribution of nodes is divided by the length of the node and is placed in $[]$ list. An array $Rlist[]$ is constructed in which the nodes are placed according to decreasing order of their quality increase values. We may proceed with the selection of nodes that have highest value in $Rlist[]$, but it may lead to the selection of nodes that are too localized over a part of the text and thus, do not represent a generalized picture. For this reason, a different method is presented here, that avoids the problem of selecting nodes centralized at a point. This could be well explained through an example shown in Fig.1, where we consider that $W_{xy} = d1i, 8 Y 2 Ni$,

Algorithm 2 Selection of salient nodes(Proposed algorithm) procedure

```

GetTopNodes( $G, \Phi, b$ )
Let  $n$  be the nodes in graph  $G$ 
topnodes  $\leftarrow \emptyset$ ,  $a \leftarrow 0$ , count  $\leftarrow 0$ ,  $S \leftarrow \emptyset$ ,
 $c \leftarrow 0$ ,  $0$ ,  $Rlist \leftarrow \emptyset$ 
for  $a = 1$  to  $n$  do
     $\psi[a]$ 
 $\psi[a] \rightarrow \frac{\psi[a]}{(\text{len}(\psi[a]))^r}$ 
end for
Rlist Sort nodes in  $[a]$  according to decreasing
value
while  $c < b$  and  $a \neq n$  do
if  $G[a]$  is not a neighbor in  $Rlist$  then

```

```

Add  $G[a]$  to topnodes
count count + 1
 $c = c + \text{length}(a)$ 
 $S \leftarrow S \cup \text{topnodes}$ 
end if
 $a \leftarrow a + 1$ 
end while
return  $S$ 
end procedure
 $\forall X \in N$  and  $W_{ab}=W_{ad}=1/2$ ,  $W_{ba}=W_{bc}=W_{bd}=W_{be}=1/4$  etc.

```

For the above network, $Rlist[] = (e; d; s; b; o; i; m; y; n; k; l; a; f; c; h; p; q)$.

When top-4 nodes are chosen from $Rlist[]$ they would be $(e; d; s; b)$, but it is observed that these nodes are localized over the same portion of the network and the cascade of information would increase, only if the nodes are spread in an appropriate manner over the network. To overcome the above limitation, the following approach is adopted for the selection of top nodes from the $Rlist[]$:

1. the first node is taken from the $Rlist[]$ and is added to the topnodes[] list.
2. the second node is taken from the $Rlist[]$ only if it does not have any neighbors already present in the top nodes[] list and the cost of addition of that node does not exceed budget b . Likewise, following the above procedure, after a certain number of nodes are added to the topnodes[] list, there arrives a situation when there is no possibility of the addition of another node, since it would exceed the budget B_g and hence, the procedure is stopped. Algorithm 2 discusses about the selection of topnodes from the $Rlist[]$ in the same way explained above. The list of nodes occurring in the topnodes[] list represent the summary S of document D . Two modifications are made to the naive algorithm[]: first is the introduction of proposed salient selection algorithm based on Shapley value, a game theory concept to solve $bmcp$ and second one is the introduction of a scaling factor $r(0 \leq r \leq 1)$, which

is used to adjust the scale of cost. It might be possible that objective function values (Shapley value given by eq. 1) and cost are not comparable. For eg., it is difficult to predict whether a sentence S_a constituting of 10 words with Shapley value 3 is better than another sentence S_b constituting of 6 terms with Shapley value 1. This problem can be reduced by using scaling. In the real-world environment, there might exist certain words that are not occurring very frequently, we eliminate such terms since they make no contribution in the summary generation and name it as Y .

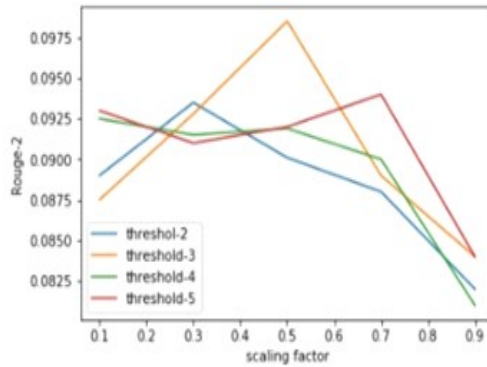


Fig. 2: Scaling factor

Table 1:
Statistics of DUC and TAC datasets

System	ROUGE-1	ROUGE-2
Centroid	0.1918	0.2362
Random	0.1196	-
DUC best	0.2523	0.2840
NMF	0.1628	0.2168
LSA	0.1502	0.2022
WCS	0.2484	0.2789
OCDSum	0.2548	-
Proposed	0.3348	0.3656

6. Evaluation

6.1 Dataset

Experiments are conducted to evaluate three types of summarization tasks and the proposed framework has performed competitively, in comparison with other methods. Experiments are performed on DUC02 and DUC04 datasets for generic summarization, for query focussed summarization DUC06 and DUC05 datasets are used, and TAC08 datasets are used for update summarization. The details of the datasets are given in Table 1.

6.2 Evaluation Metric

The proposed work uses ROUGE toolkit^{29,30} for evaluating the quality of generated summary with human generated summary. ROUGE metrics are based on the comparison of n-grams between the summary to be evaluated and one or several human written reference summaries. ROUGE-2(R-2) recall is considered as the main metric for comparison, since it has high capability to evaluate automatic summaries. In addition, ROUGE-SU4(R-SU4) recall score is also computed for analysis purpose. However, for some data sets R-SU4 score is not available.

Table 2:
Results of DUC2002

System	ROUGE-1	ROUGE-2
Centroid	0.1918	0.2362
Random	0.1196	-
DUC best	0.2523	0.2840
NMF	0.1628	0.2168
LSA	0.1502	0.2022
WCS	0.2484	0.2789
OCDSum	0.2548	-
Proposed	0.3348	0.3656

Table 3:
Results for DUC04

System	ROUGE-1	ROUGE-2
Centroid	0.7981	0.1251
Random	0.0639	0.1177
DUC	0.0922	0.1377
NMF	0.0726	0.1291
LSA	0.0654	0.1194
WCS	0.0961	0.1353
OCDSum	0.0969	-
Greedy-kl	0.0856	-
Js-Swarm-2	0.0818	-
Ts Sum	0.0816	-
Freq.Sum	0.0811	-
Classy 11	0.0921	-
Classy 04	0.0901	-
DPP	0.0961	-
Submodular	0.0936	-
Js-Gen-2	0.0930	-
CSEM	0.0974	-
Proposed	0.0984	0.1355

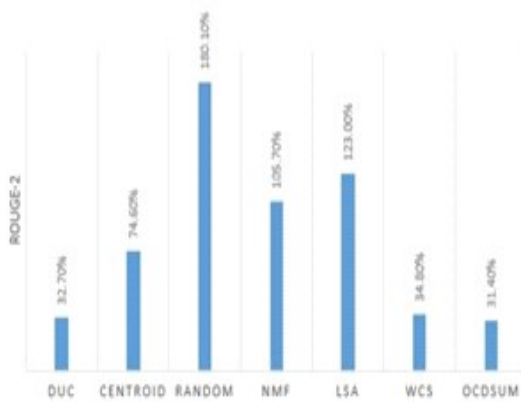


Fig. 3: Relative improvement on DUC02 for R-02

6.3 Experimental Analysis

From generic summarization perspective, extensive experiments are conducted on DUC02 and DUC04 dataset. After performing extensive experiments, it is observed that (see Fig.2)(a) under consistent experiment setting, different scaling factors give different results (b) it is found that best summary results are achieved when $r = 0.5$ and $\tilde{O} = 3$. To achieve best results, this value of scaling factor is used for other datasets also.

6.3.1 Generic Summarization

From generic summarization perspective, extensive experiments are conducted on DUC02 and DUC04 dataset.

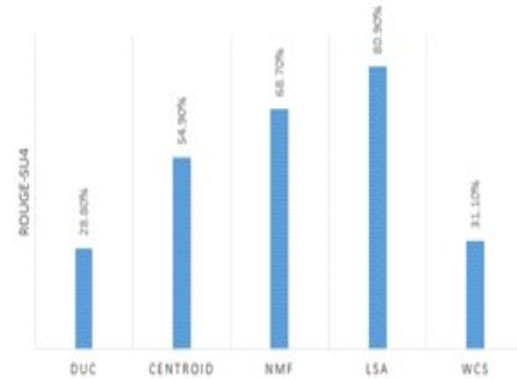


Fig. 4: Relative improvement on DUC02 for R-SU4

The empirical results obtained are compared with some of the well-known methods and their results are taken from Table 9 and Table 10 of [31] and Table 5 of [32]. The competitive methods considered for comparison are mainly based on either submodular functions or ontology and are given as: (i) DUC best: Represents the best participant of DUC 2002 and DUC 2004 conference; (ii) Random (baseline method): Summaries are constructed by selection of sentences in a random manner from document

set;(iii) Centroid (baseline method): Different features of sentences are used to construct summary and then similarity of every candidate sentence is computed with the first sentence; (iv) FreqSum: An extractive summarization approach was proposed where key sentences were identified using various features of the textual content;(v) NMF: A term by sentence matrix is created by applying non-negative matrix factorization on the document and salient sentences are identified according to the weighted value;(vi)LSA: In this approach, singular value decomposition(SVD) was applied to the document-term matrix and key sentences were identified based on the obtained weighted value;(vii) WCS:

The ranks obtained from various single documents are combined to form the summary of multi-documents; (viii) OCDSum:It was based on the concept of differential evolution algorithm (xi) GreedyKl: It used hierarchical LDA(linear discriminant analysis) concept for summarization; (x) Classy04: Sentences were scored using HMM(hidden markov model); (xi) TsSum: The approach focussed on the identification of topic signatures for summarization purpose; (xii) Classy11: An extractive summarizer based on probabilistic topic model was proposed for a multi-lingual system; (xiii) DPP: A probabilistic approach was introduced for summarization; (xiv) Submodular: Focussed on the submodularity that is hidden in the underlying text to generate summary; (xv) JS-Gen2 and Js-Swarm2: Both of them generated extractive summaries based on evolutionary algorithms; and (xvi)CSEM[32]: A sentence extraction model was proposed based on the semantic knowledge, that is mined from the underlying corpus. As is evident from Table 2 and Table 3, that the R-2 recall results of the proposed system are best, in comparison son to other rivals on DUC02 and DUC04 datasets. The percentage of relative improvement is computed to analyze the performance of the proposed approach and is presented in Fig. 3-6.

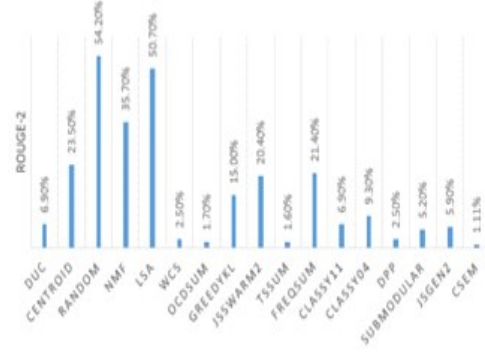


Fig. 5: Relative improvement on DUC02 for R-SU4

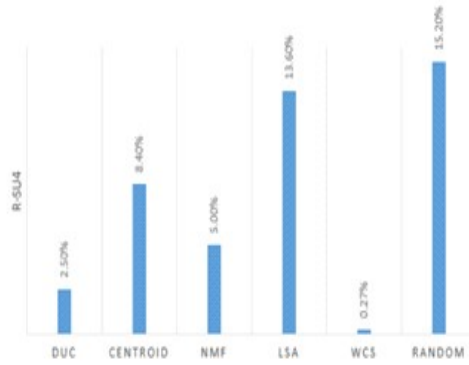


Fig. 6: Relative improvement on DUC02 for R-SU4

Table 4: Results for DUC'05

Systems	R-2	R-SU4
Random	0.04143	0.09066
DUC best	0.06024	0.11488
LSA	0.04079	0.09352
Qs-MRF	0.07790	0.13660
SNMF	0.06043	0.12298
wiki	0.07070	0.13002
Proposed	0.07731	0.13000

Table 5:
Results for DUC'06

Systems	R-2	R-SU4
Random	0.05022	0.10226
DUC best	0.07543	0.13206
LSA	0.04892	0.10083
Qs-MRF	0.08917	0.14329
SNMF	0.08549	0.13981
wiki	0.08091	0.14022
Proposed	0.09202	0.14612

Conclusion

The main contribution of this paper is the proposal of a novel sentence selection

algorithm based on an unsupervised game theory framework for the multi document summarization. Submodular functions based on a combination of statistical, linguistic and co-reference similarity measures is utilized to deal with different mds tasks. The proposed algorithm ensures that the sentences identified as the summary not only covers the most relevant information, but also provides a generalized perspective of the documents to be summarized. In comparison with other multi-document summarization systems (DUC, TAC competitors, baseline methods, and state-of-the-art methods), the proposed system showed a remarkable improvement. However, the limitation of the proposed approach is, it is applicable for English text only and can be further enhanced by incorporating the idea of a multilingual language detection system.

References

1. B. Marr Big Data: 20 Mind-Boggling *Facts Everyone Must Read*. (2015)
2. D. Wang, T. Li, S. Zhu, C. Ding Multi-document summarization via sentence-level semantic analysis and symmetric factorization. In: Proceedings of SIGIR, pp. 307 (2008)
3. M. Inderjeet Summarization Evaluation: An Overview. In: Proceedings of the NTCIR Workshop (2001)
4. R. Hi, B. Qjn , T. Lin A novel approach to update summarization using evolutionary manifold ranking and spectral clustering. Expert Systems with Applications, vol. 39, pp. 2375 (2012)
5. A. Nenkova, R. Passonneau, K. Mckeown The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Language and Speech Processing, vol. 4(2),1 (2007)
6. N. Chatterjee, S. Mohan S. Extraction-Based Single-Document Summarization Using Random Indexing. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, pp. 448 (2007)
7. R. Mihalcea, P. Tarau TextRank: Bringing order into texts. Association for Computational Linguistics (2004)
8. J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz Multi-document Summarization by Sentence Extraction. In: *Proceedings of the 2000 NAACL/NLP Workshop on Automatic Summarization*, pp. 40 (2000)
9. A. Abuobieda , N. Salim , A.T. Albaham, A.H. Osman, Y.J. Kumar Text summarization features selection method using pseudo Genetic-based model. *Information Retrieval Knowledge Management*,193 (2012)
10. ST. Davis, JM. Conroy, JD. Schlesinger OCCAMS|an optimal combinatorial covering algorithm for multi-document summarization. IEEE, *International Conference on Data Mining Workshops*. IEEE, pp. 454 (2012)
11. Y. Wenpeng, P. Youlong Optimizing Sentence Modelling and Selection for Document Summarization. In: Proceedings of the 24th *International Joint Conference on Artificial Intelligence*, pp. 1383 (2015)
12. L. Yang, X. Cai, Y. Zhang, P. Shi Enhancing sentence-level clustering with ranking-based **summarization**. *Information Sciences* vol. **260**, pp. 37-50 (2014)
13. R. Ferreira, L. de Souza Cabral, R.D. Lins, G.P. e Silva, F. Freitas, G.D.C. Cavalcanti, R. Lima, S.J. Simske , L. Favaro Assessing sentence scoring

- techniques for extractive text summarization. *Expert Systems with Applications*, vol. 40, pp. 5755-5764 (2013)
14. S. Khuller, A. Moss, JS Naor The budgeted maximum coverage problem. *Inf Process Lett*, **70(1)**, pp.39-45 (1999)
 15. T. Li, S. Zhu, M. Ogihara Hierarchical document classification using automatically generated hierarchy. *Journal of Intelligent Information Systems*, **29(2)**, pp. 211 (2007)
 16. D. Sanchez, M. Balet, K. Gilbert Ontology-driven web-based semantic similarity. *Journal of Intelligent Information systems*, pp. 1 (2009)
 17. H. Lin, J. Bilmes Multi-document summarization via budgeted maximization of submodular functions. In: NAACL/HLT 2010 Annual Conference of the North American Chapter of the ACL, pages 912-920 (2010)
 18. DR Radev, HY Jing, M Sty's, D. Tam Centroid based summarization of multiple documents. *Inf Process Manag*, pp. 919-938 (2004)
 19. E. Canhasi, I. Kononeko Weighted archetypal analysis of the multi-document summarization. *Expert Systems with Applications*, vol. **41(21)**, pp.535 (2014)
 20. A. Celikyilmaz, D. Hakkani-Tur, G. Tur, A Fidler, D. Hillard Exploiting distance based similarity in topic models for user intent detection. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2011)
 21. C.C. Yang, F.L. Wang Hierarchical summarization of large documents. *Journal of the American Society for Information Science and Technology*, vol. **59(6)**, pp. 887 (2008)
 22. L. Chong, H. Min-Lie, Z. Xiao-Yan, Li. Ming A New Approach for Multi-Documnet Update Summarization. *Journal of Computer Science and Technology*, vol. **25**, pp.739 (2010)
 23. Li. Xuan, D. Liang, S. Yi-Dong Update Summarization via Graph-based Sentence Ranking. *IEEE Transaction on Knowledge and Data Engineering*, vol. **25**, pp. 1162 (2013)
 24. Li. Rumeng, S. Hiroyuki A Hierarchical Tree Model for Update Summarization. In: *European Conference on Information Retrieval*, pp.660 (2015)
 25. H. Lin, J. Bilmes Multi-document summarization via budgeted maximization of submodular functions. In: NAACL/HLT 2010 Annual Conference of the North American Chapter of the ACL, pages **912-920** (2010)
 26. R. Narayanam, Y. Narahari A shapley value based approach to discover influential nodes in social networks. *IEEE Transactions on Automation Science and Engineering*, pp. 1-18 (2011)
 27. T. Pedersen Information Content Measures of Semantic Similarity Perform Better Without Sense-tagged Text. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 329 (2010)
 28. A. Radford English syntax: An introduction. Cambridge University Press (2004)
 29. D. Das, A.F.T. Martins A survey on automatic text summarization. In: *Literature Survey for the Language and Statistics II course at CM* (2007)
 30. D. Radev, H. Jing H., Stys M. Centroid based summarization of multi-documents. *Information Processing and Management*, pp. **919** (2004)
 31. A. John, P.S. Premjith, M. Wilscy Extractive multi-document summarization using population based multi-criteria optimization. *Expert System with applications* (2016)
 32. Liu Weidong, Luo Xiangfeng, Jun Zhang, Xue Ruihong, Yi Da Xu Richard Semantic summary automatic generation in news event. *Concurrency and Computation: Practice and experience* (2017)

