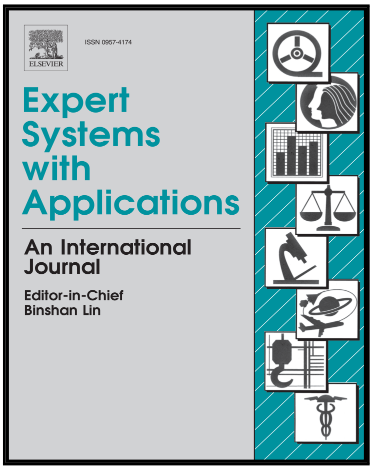
Communicated by Binshan Lin

Accepted Manuscript



Extractive Multi-Document Summarization Using population-based Multicriteria Optimization

John Ansamma, P.S. Premjith, M. Wilscy

PII: S0957-4174(17)30404-9
DOI: 10.1016/j.eswa.2017.05.075
Reference: ESWA 11367

To appear in: *Expert Systems With Applications*

Received date: 9 November 2015
Revised date: 28 May 2017
Accepted date: 29 May 2017

Please cite this article as: John Ansamma, P.S. Premjith, M. Wilscy, Extractive Multi-Document Summarization Using population-based Multicriteria Optimization, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.05.075

Highlights

- A system with feature extraction and multicriteria optimization technique.
- Uses objective functions that cover both statistical and semantic aspects.
- Population based approach with random weights for features in each generation.
- High Precision and Recall values compared to the other popular methods.

Extractive Multi-Document Summarization Using population-based Multicriteria Optimization

Ansamma John[a,*], Premjith P.S[b], Wilscy M[a]

[a]Department of Computer Science, University of Kerala, Kariavattom, Kerala, India
[b]Department of Computer Science, T.K.M College of Engineering, Kollam Kerala, India

Abstract

Multi-document summarization is the process of extracting salient information from a set of source texts and present that information to the user in a condensed form. In this paper, we propose a multi-document summarization system which generates an extractive generic summary with maximum relevance and minimum redundancy by representing each sentence of the input document as a vector of words in Proper Noun, Noun, Verb and Adjective set. Five features, such as TF_ISF, Aggregate Cross Sentence Similarity, Title Similarity, Proper Noun and Sentence Length associated with the sentences, are extracted, and scores are assigned to sentences based on these features. Weights that can be assigned to different features may vary depending upon the nature of the document, and it is hard to discover the most appropriate weight for each feature, and this makes generation of a good summary a very tough task without human intelligence. Multi-document summarization problem is having large number of decision parameters and number of possible solutions from which most optimal summary is to be generated. Summary generated may not guarantee the essential quality and may be far from the ideal human generated summary. To address this issue, we propose a population-based multicriteria optimization method with multiple objective functions. Three objective functions are selected to determine an optimal summary, with maximum relevance, diversity, and nov-

*Ansamma John
Email addresses: ansamma.john@gmail.com (Ansamma John), premjith1190@gmail.com (Premjith P.S), wilsyphilipose@hotmail.com (Wilscy M)

elty, from a global population of summaries by considering both the statistical and semantic aspects of the documents. Semantic aspects are considered by Latent Semantic Analysis (LSA) and Non Negative Matrix Factorization (NMF) techniques. Experiments have been performed on DUC 2002, DUC 2004 and DUC 2006 datasets using ROUGE tool kit. Experimental results show that our system outperforms the state of the art works in terms of Recall and Precision.

Keywords: Multi-document Summarization, Multicriteria optimization, Latent Semantic Analysis, Non Negative Matrix Factorization, DUC, ROUGE

Reviewers comments and point to point responses of second review.

Reviewer 1

Comment 1: The legend of Figure 1 in page 11 needs to be fixed.

Response : Figure is resized to fix the problem with the legend of Figure 1.

Reviewer 2:

Comment 1: The authorship page is incomplete.

Response : Authorship page is completed.

Reviewers comments and point to point responses of first review.

Reviewer 1

Comment 1: Overall, it is an interesting paper. It is not entirely clear if the proposed models would work well in other datasets. For this reason, I would like to see an additional round of experiments in at least another dataset(s)
Response : Additional rounds of experiments are done on DUC 2004 and DUC 2006 data sets and results are given in the paper.
Comment 2 : Authors need to include more recent state-of-the-art. It is missing references to well-know optimization methods such MMR, event-based summarization, and other recent methods:
Response: All specified well-known methods are included in appropriate positions in the paper
Comment 3: The conclusion needs to be expanded and explain why the new method is better.
Response : Conclusion is expanded with the explanation of why this method is better.
Comment 4: The paper would also benefit from the inclusion of example summaries.
Response : Sample summary generated for the topic D30001T Of DUC 2004 is included in the paper before the conclusion part.
Comment 5: The paper needs proofreading of nearly every page. It has multiple typos, missing commas, and it needs normalized language.
Response : Proof reading is done and corrections are made.
Comment 6 : section 3.2.5 - Does these formula works in another datasets?
Response : Yes. It is for giving more preferences to the sentences whose length is near to the average length of the sentences in the document sets. Yes it is working well for all datasets.

Reviewer 2:

Comment 1: My major concern is that authors need to compare the performance of their proposal to other popular datasets. Therefore, the amount of improvement of the proposal has not been correctly evaluated, and so the insight is not clear.

Response : We performed experimental evaluation with other datasets DUC 2004 and DUC 2006 to et a more insight about the result.

Comment 2: Additionally, to my perspective, authors should make an additional effort to better present their work since in its present form the benefits are not clearly expressed and quantified.

Response 2 : Tried to change the presentation of the work to express the benefits.

Comment 3: The authors should give the readers some concrete information to get them excited about their work. The current abstract only describes the general purposes of the article. It should also include the article's main (1) impact and (2) significance on expert and intelligent systems.

Response : Abstract is modified to include articles main impact.

Comment 4: I believe that it will make this paper stronger if the authors present insightful implications in at least one paragraph-based on their experimental outcomes.

Response : One paragraph is included at the end of the experimental evaluation section to present insightful implications on our experimental outcomes.

Comment 5: The authors also need to clearly provide 4-5 solid future research directions in the Conclusion section. These directions should be written as at least a separate paragraph and such directions need to be insightful for most of ESWA community.

Response : 4 future research directions are incorporated in the conclusion paragraph.

Comment 6 : Finally, the language and grammar also require some work, and I noted a large number of typographical errors. The paper needs a linguistic check, preferably by a native speaker.

Response : Efforts are put to correct the typographical errors. Linguistic check is done by native speaker.

---

1. Introduction

The meteoric increase of documents on the World Wide Web lead to the availability of multiple documents with same theme and concept, which in turn lead to the data overload. It is difficult to identify an appropriate single document related to a particular theme that satisfies user's complex information requirements within a limited time. One solution for this is to adopt a document summarization system. Automatic document summarization is the process of generating the compact version of the document/documents while maintaining the important aspects of the input document/documents.

There are two main methods for text summarization: abstractive and extractiveMani & Maybury (1999). Abstractive summarization system generates coherent summaries by analysing and understanding the contextual and linguistic information of the document, using natural language processing tools and rewriting the appropriate sentences such that generated single sentence may cover the information associated with one or more important sentences in the document Mani & Maybury (1999); Mei & Chen (2012); Otterbacher et al. (2002); Filatova & Hatzivassiloglou (2004); Marujo (2015); Marujo et al. (2015a); Christensen et al. (2013, 2014); Marujo et al. (2015b). Extractive summarization techniques choose and concatenate relevant sentences from the input document to form summaries. Extractive summarization system assigns a score to each sentence in the source document based on the information content and only highly scored novel sentences are selected to the summary. Extractive summarization systems are very popular due to its low cost and simplicity compared to other approaches.

Based on the content, summaries can be classified as task-specific (query based) or generic. Query based summarizer generates summaries by extracting sentences that are relevant to user's query, whereas generic summaries provide main concepts of the source document without any external interaction Mei & Chen (2012). Goldstein et. al. Goldstein et al. (2000); Carbonell & Goldstein (1998) developed query based summarizer using maximum marginal relevance

(MMR) principle to reduce redundancy and to enhance the diversity of the content in summary.

Depending upon the number of source documents to be summarized, summaries can be single document or multi-document. Single document summary contains salient information from a single document, while multi-document summaries contain relevant information that covers the entire concept of two or more documents without much redundancy (Dunlavy et al., 2007). Researchers developed different approaches in extending the single document summarization to multi-document summarization such as single-layer hierarchical and waterfall (Marujo et al., 2015c).

Summarization process can be supervised or unsupervised (Mani & Maybury, 1999; Fattah & Ren, 2009). In a supervised extractive summarization system, there are two phases: learning and testing. During the learning phase, training documents and their corresponding summaries are given, by which system learns to categorise the sentences into two classes, summary sentences, and non-summary sentences. In a supervised environment, summarization is considered as a binary classification problem. Many summarization systems are developed using Support Vector Machine (Chali et al., 2009), Neural Network based (Abdel Fattah & Ren, 2008), Random forest(John & Wilscy, 2013) and other classifiers. The main disadvantage of the supervised system is that it requires a large amount of training data for getting the good summary. Many extractive summarization systems were developed using unsupervised techniques such as clustering, which does not require any labelled training dataset.

By representing documents as a graph and identifying salient elements of the text using topology of the graph several graph-based summarization (Canhasi, 2014; Page et al., 1999; Erkan & Radev, 2004; Glavaš & Šnajder, 2014) methods are developed. graph-based algorithms are utilized to extract salient sentences from the input documents.

We propose a population-based extractive unsupervised and generic multi-document summarization system by extracting salient sentences from the input documents by considering summarization as a multicriteria optimization problem. Feature based sentence extraction method is adopted to determine the extent to which selected sentences can cover the entire collection of documents. Five features such as Term Frequency- Inverse Sentence frequency (TF_ISF), Aggregate Cross Sentence Similarity, Title Similarity, Proper Nouns Existence, and Sentence Length associated with the sentences in the documents are extracted, and sentences are scored based on the weights assigned to these features. The task of assigning an appropriate weight to the features is difficult due to the possibility of selecting an infinite number of weight values in the range 0 to 1 for each feature and varying dominance of features in different documents due to the divergent writing style and skills of different authors. Hence random weights are assigned to the features to score the sentences, and diverse population of summary is generated iteratively from the scored sentences. From the population of summaries generated, an optimal summary that is close to the human generated summary can be selected. Hence a multicriteria optimization with multiple objective functions is used to produce the most appropriate summary. Three objective functions are considered in our work to determine the quality of the summary in terms of content coverage and novelty, by monitoring the statistical and semantic aspects of the document. Three independent summaries each maximizing the value of the respective objective functions are maintained. The final summary is generated by the appropriate selection of sentences from these summaries. ROUGE tool kit is used for the performance analysis of our system using DUC 2002, DUC 2004 and DUC 2006 datasets. Precision and Recall values obtained for our system is better than the results of other state of the art systems.

The rest of the paper is organized as follows: Section 2 discusses related works in text summarization, Section 3 deals with the proposed summarization system with feature extraction and multicriteria optimization, Section 4 explains the result and analysis of our work and Section 5 conclusion.

2. Related Works

A large number of extractive summarization systems are developed over the years. Sentence ranking is the most critical step of all extractive summarization system and researchers are putting much effort to improve the sentence ranking method to enhance the quality of the summary.
Radev et. al.(Otterbacher et al., 2002) developed centroid based multi-document summarization system applicable for both single and multiple documents. For every sentence, they extracted three features such as centroid score, positional score and overlap with the first sentence. Linear combination of these features scores is used to select the salient sentences to the summary. Josef Steinberger, Karel Jeek (Steinberger & Ježek, 2005) proposed a generic summarization system by detecting semantic aspects of the text using the singular value decomposition method. It improved the overall quality of the extracted summary. A new evaluation method is also developed which considers the similarity between the source document set and summary. Leonhard Hennig(Hennig & Labor, 2009) developed document summarization system based on Latent Semantic Analysis to capture the main topic of a document. Sentences are extracted to the summary based on their topic relevance and coverage. This method also satisfies the multi-lingual constraint.

Sun Park, Ju-Hong Lee, Deok-Hwan Kim, Chan-Min Ahn (Park et al., 2007) introduced a multi-document summarization system that is capable of extracting meaningful sentences into the summary by using non-negative matrix factorization (NMF)and k- means clustering. Using NMF, relevant semantic variables are extracted which in turn are employed to identify most topic relevant sentences. They incorporated k-means clustering to avoid the effect of biased inherent semantic of the document set.
Rasim M.Alguliev, Ramiz M.Aliguliyev and Nijat R Isazade (Alguliev et al., 2013) proposed an optimization based model for generic document summarization by considering sentence to document collection, the summary to document

collection and sentence to sentence relation to extract the salient sentences. Using Differential Evolutionary algorithm, they solved the optimization problem to come up with a good summary.

Above mentioned works perform the summarization by the analysis of statistical TF_IDF measure or by the analysis of semantic features (Binh Tran, 2013) associated with the sentences, or by the application of the evolutionary algorithm, where optimization is performed by considering the frequency of occurrence of terms in summary and the documents. We propose a generic multi-document summarization system, which generates summaries by using multicriteria optimization functions. We used three objective functions to find a summary with maximum coverage, novelty and semantic aspects. Semantic aspects are identified using the SVD and NMF techniques. population-based approach is used to come up with a good summary. The proposed method is explained in the next section.

3. Proposed Work

In multi-document summarization, a generic summary provides information about the salient and diverse concepts present in the documents. In this work, an extractive generic summarizer is presented which uses feature based sentence scoring and multicriteria optimization with multiple objective functions. Objective functions are defined to identify statistical and semantic coverage of the summary to improve the overall quality of the summary. The final summary is generated by the population-based approach, where each generation is created from two parent chromosomes. In each generation, parent chromosomes are identified by assigning random weights to the features and scoring the sentences based on these weights. Block diagram of the proposed method is given in Figure 1.

The methodology is explained in detail in the following subsections.

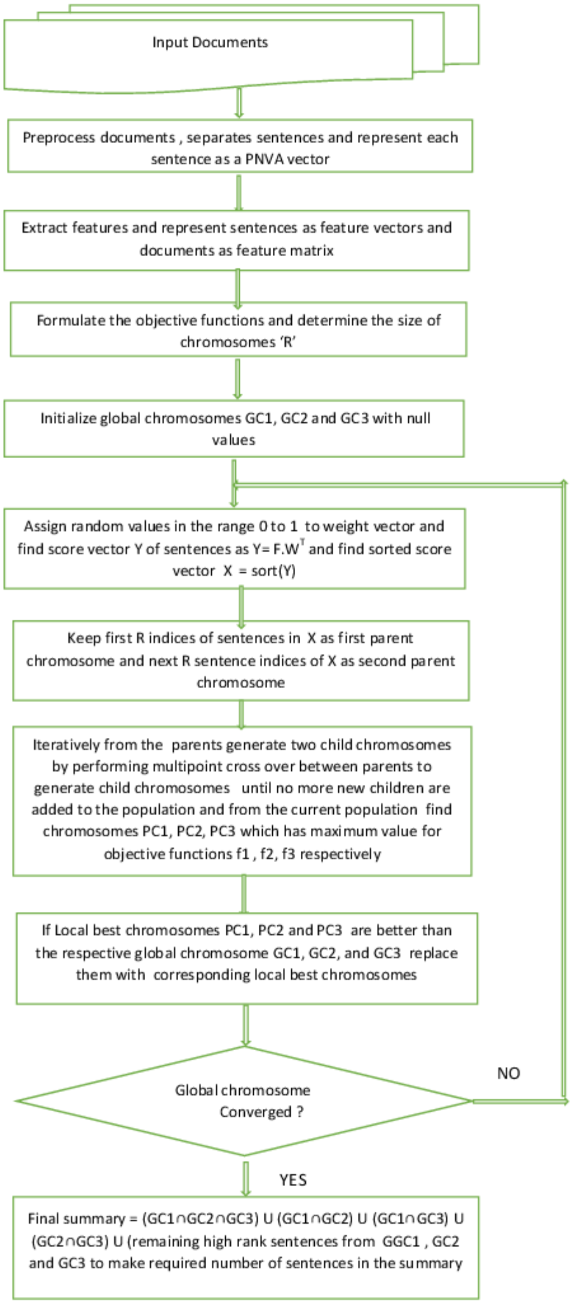


Figure 1: Diagrammatic representation of the proposed work

3.1. Document Pre-processing

Let $D = \{D_1, D_2, D_3, ..., D_d\}$, be the set of input documents presented for the summarization, in random order. Sentences from the source documents are extracted by the sentence segmentation and tokenization process, which results in the separation of sentence set $S = \{S_1, S_2, ..., S_n\}$, where each sentence $Si \epsilon S$ is a collection of tokens. If each document D_i consists of n_i number of sentences then total number of sentences in D is n, where $n = \sum_{i=1}^{d} n_i$. Documents in D may also contain information which are not relevant from summarization viewpoint. Therefore, a preprocessing step is performed to extract the required textual units and to represent them in a convenient format for further processing. The sentences may contain words that do not contribute any meaningful information. In general, those words are called stop-words, which are removed in the stop-word removal step using the stop-word collection provided by Brown corpus. By the analysis of documents, it is found that some words in the documents such as 'a.m.', 'p.m.', 'fla.', 'edt.' and all the tokens with less than three alphabets, do not make any significant contribution to the summarization process and hence they too are removed by creating a modified stop-word list of our own. Sentence set S is transformed into a stop-word removed set of sentences $S' = \{S'_1, S'_2, ..., S'_n\}$ such that $|S| = |S'|$ *and* $|S_i| \geq |S'_i|$.

As the size and number of documents increase, the number of words and its different forms also increase considerably which in turn increase the complexity of sentence representation. Hence the tokens are converted to its root form using Porter Stemming Algorithm (Willett, 2006; Porter, 1980) thereby reducing the number of distinct tokens in the document set. Stemming process on S' results in a set $ST = \{St_1, St_2, .., St_n\}$, where each St_i consist of a stemmed version of the tokens present in S'_i. Since we are processing multiple documents related to the same topic, there is a possibility of repetition of salient sentences in different documents. Summary produced by the selection of salient sentences may lead to the redundancy in summary. Redundancy of information in documents can appear directly or in the form of subsumption, in which the information in one sentence is subsumed in another sentence. If a sentence S_a is subsumed in S_b, then the selection of both the sentences into the summary will result in redundant information. From the subsumed sentence pair, only the longer sentence is added to the summary. Adopting Cosine Similarity to find sentence subsumption, is not an apt measure. If the number of words in the longer sentence is much higher some common word in shorter sentences, then Cosine Similarity between the sentence pair will be a small value. Hence a variant form of Simple Matching Coefficient scheme is formulated as follows. Sentence S_a is preprocessed to get reduced version of it say, St_a . Similarly St_b is obtained. Then subsumption measure γ is computed as

$$\gamma = |St_a \cap St_b|/|St_a| \tag{1}$$

where $St_a \leq St_b$. If $\gamma > 0.80$, then S_a is removed from further processing. This process is applied progressively to sentences in ST resulting in reduced sentence set STR, where $STR = \{ST_1, ST_2, ..., ST_{n'}\}$ and $n' \leq n$. Now The sentences in STR have variable lengths and this is not an appropriate representation of sentences for the statistical processing of extractive summarization. Vector Space Model(VSM) is a popular method used for the representation of sentences in extractive summarization. Let $T = \{t_1, t_2, t_3, ..., t_m\}$ represents all the distinct m terms appearing in the set STR. Most of the extractive summarization research work uses Vector Space Model (VSM), a bag-of-words approach in which a sentence Si is represented as $Si = \{tk_1, tk_2, ..., tk_m\}$ where m is the size of stemmed words set . Each $tk_i \,\epsilon\, ST_i$, can be either 1 or 0 indicating the presence or absence of i^{th} token of T in sentence ST_i, respectively or the frequency of occurrence of i^{th} token of T in sentence ST_i. When VSM model with stemmed words is used, every sentence is transformed into a m-dimensional vector, which increases space complexity. By the inspection of human generated summaries, it has been observed that majority of the tokens in the sentences of summaries fall into the Proper Noun, Noun, Verb and Adjective category. Based on this observation, Part-of-Speech (POS) tagging is employed to identify Proper Nouns, Nouns, Verbs and Adjectives. To reduce the size of the sentence vectors and to perform some computation in an efficient way, another VSM representation scheme for sentences is introduced by employing POS tagging on the set STR which results in the separation of set of Proper Nouns(PS), Nouns(NS), Verbs(VS), Adjectives(AS). POS tagging is case sensitive and may classify some verbs, adjectives and nouns as proper nouns. So better approximation of Proper Noun set (PNS) can be constructed using equation (2).

$$PNS = PS - (NS \cup VS \cup AS) \tag{2}$$

Also, during POS tagging, some nouns are classified into both noun and proper noun categories. In the context of extractive summarization, these words are having more significance than the words that appear in the noun set alone. Let PNNS be the set of words that appear in both noun set and proper noun set categories and is constructed using (3).

$$PNNS = PS \cap NS \tag{3}$$

Pictorial representation of the above mentioned sets is given below. After PNNS

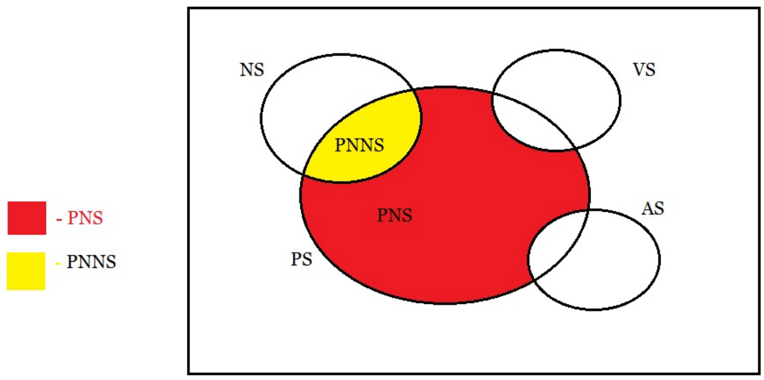


Figure 2: Diagramatic representation of proposed work

is created, words in PNNS are removed from Noun Set (NS) which results in $PNNS \cap NS \cap VS \cap AS = \phi$. For further processing, a list PNVA is generated from the sets PNS, PNNS, NS, VS and AS by performing stemming on terms in these sets and keeping them in the respective order. PNVA act as the dictionary of terms for sentence representation and each sentence $St_i \epsilon STR$ can be represented as $St_i = \{r_1, r_2, r_3, , r_j, ..., r_{m'}\}$, where $m' = |PNVA|, r_j$- is the term-weight of j^{th} PNVA token in St_i. Clearly $m' < m$, therefore using the newly introduced PNVA representation method, the dimension of each sentence is reduced to *m'*, which is much less than *m*. After preprocessing of documents, we proceed to the next stage, Feature Based Scoring, where sentences are scored, based on the features such as TF_ISF, title resemblance, aggregate similarity, sentence length and proper noun.

3.2. Feature Based Sentence Scoring

In this work, five features are utilized for scoring individual sentences. From the examination of the existing works in the literature and from the analysis of Gold Standard Summaries from DUC 2002, DUC 2004 and DUC 2006 datasets, it is observed that scoring of sentences using features such as Term Frequency Inverse Sentence Frequency (TF_ISF) (Blake, 2006), Aggregate Cross Sentence Similarity, Title Similarity (Salton & Lesk, 1965), Proper Nouns (Pitler et al., 2010), and Sentence Length (Kupiec et al., 1998) provide a better numerical score estimation for sentence. These five features contribute heavily to the extractive generic summarization of multiple documents.

To facilitate the scoring process, sentences are transformed to suitable formats. Here we use two kinds of sentence representation based on PNVA scheme. One is using the frequency of PNVA terms in the sentence, in which a sentence $St_i = \{TF(r_1), TF(r_2), TFr_3), ..., TF(r_j), ..., TF(r'_m)\}$, where $TF(r_j)$ is the frequency of occurrence of j^{th} token of PNVA in St_i. We name it as PNVA_TF representation. The second representation is using the TF_ISF value of a token in PNVA with reference to document set D. TF_ISF value of a token gives the importance of that token in the sentence. Generally, TF_ISF of a term can be found out using the equations (4), (5), and (6)

$$TF(r_j) = (count(r_j) \ in \ STR)/N \tag{4}$$

$$ISF(r_j) = log(N/(1 + N(r_j))) \tag{5}$$

$$TF\_ISF(r_j) = TF(r_j) * ISF(r_j) \tag{6}$$

where N is the sum of occurrences of all tokens in T and $N(r_j)$ is the number of sentences containing the token r_j. TF_ISF values are computed for every token in PNVA. A sentence St_i is represented in PNVA_ISF as $St_i = \{TF\_ISF(r_1), TF\_ISF(r_2), TF\_ISF(r_3), ..., TF\_ISF(r_j), ..., TF\_ISF(r_{m'})\}$, where $TF\_ISF(r_j)$ has the TF_ISF value of j^{th} token of PNVA if that token is present in St_i and zero otherwise. We name this format as PNVA_ISF. The feature scoring using TF_ISF and Aggregate Cross Sentence Similarity employs PNVA_ISF while Proper Nouns and Title Similarity employs PNVA_TF. The reason behind the application of different formats for sentences is based on the studies performed on scoring patterns of those features. The features used for scoring are formulated as follows:

3.2.1. TF_ISF scoring (FS1)

Usually, sentence score heavily depends upon the number of tokens within the sentences. A summary is generated by extracting sentences corresponding to each topic and number of sentences extracted depends upon the significance of that topic(Luhn, 1958). Hence Term Frequency and Inverse Sentence Frequency is the first score used which is denoted by FS1 and is computed for a sentence St_i as the sum of the TF_ISF of terms as in equation (7) below.

$$FS1(St_i) = \sum_{j=1}^{m'} TF\_ISF(r_j) \ , \quad St_i \ \epsilon \ STR, \quad r_j \ \epsilon \ PNVA \tag{7}$$

3.2.2. Aggregate Cross Sentence Similarity (FS2)

A sentence which has maximum similarity with all other sentences in the input documents can be an appropriate candidate for the summary. Hence the second feature score FS2 is the sum of Cosine Similarity of a sentence with all other sentences in the documents and is computed as in equation (8).

$$FS2(St_i) = \sum_{j=1}^{n'} sim(St_i, St_j), \qquad St_i, St_j \; \epsilon \; STR \tag{8}$$

where Cosine Similarity between sentences St_i and St_j i.e. $sim(St_i, St_j)$ is computed as in equation (9).

$$sim(St_i, St_j) = \frac{\sum_{k=1}^{m'} St_{ik} * St_{jk}}{\sqrt{\sum_{k=1}^{m'} St_{ik}} * \sqrt{\sum_{k=1}^{m'} St_{jk}}} \tag{9}$$

3.2.3. Title Similarity (FS3)

Generally title of a document is one of the most important entity that conveys information about document's content. A sentence is salient from summarization aspect if it has maximum similarity with the titles of the documents. Title similarity feature score represented by FS3 is computed by extracting the tokens of all titles of the document set D during the preprocessing stage. A list TL is created in such a way that TL stores the stemmed tokens in the title along with its frequency of occurrences. Now a title vector TV is constructed using PNVA_TF scheme of representation i.e, $TV = \{r_1, r_2, r_3, , r_j, ...r'_m\}$, each r_j is the frequency of occurrence of j^{th} token of in TL. The title similarity of a sentence is the Cosine Similarity between that sentence and TV which is computed as shown in equation (10).

$$FS3(St_i) = sim(St_i, TV), \qquad St_i \; \epsilon \; STR \tag{10}$$

3.2.4. Proper Noun Scoring (FS4)

The fourth feature score FS4 for each sentence is computed based on the occurrence of tokens of the sentence in the sets PNS and PNNS. Tokens in PNS are more significant from the summarization view point than tokens in PNNS. Tokens in other sets are not considered for this scoring. By observation, it is identified that sentences with tokens in PNS have high probability to be placed in the summary. So more weight is given to tokens in PNS than tokens in PNNS.

Feature score FS4 is computed as follows.

$$FS4(St_i) = \sum_{j=1}^{l1} TF(r_j) + \sum_{k=1}^{l2} (0.50 * TF(r_k)) \tag{11}$$

$$St_i \;\epsilon\; STR, \;\; r_j \;\epsilon St_i \; AND \; r_j \;\epsilon PNS, \;\; r_k \;\epsilon St_i \; AND \;\; r_k \epsilon PNNS$$

where $l1 = length(PNS)$ and $l2 = length(PNNS)$.

3.2.5. Sentence Length (FS5)

The fifth feature score FS5 is employed to give more preferences to sentences having length within a particular range, based on the average length of sentences in STR. Short and long sentences are considered less significant. The score of each sentence is computed using equation(12)

$$\begin{aligned} FS5(St_i) &= (0.80 * len(St_i)/Avg && if\ len(S_i) \;<\; Avg \\ &= 1.2 * Avg/len(St_i) && if\ Avg \leq len(S_i) \leq 1.60 * Avg \\ &= 0.80 * Avg/len(St_i) && if\ len(S_i) \;>1.60 * Avg) \end{aligned} \tag{12}$$

$$where \;\; Avg = \frac{(\sum_{i=i}^{n'} len(St_i)}{n'}$$

The five feature scores associated with each sentence is computed and each sen-
180 tence in STR is transformed into a five dimensional vector $[fs_{i1}, fs_{i2}, fs_{i3}, fs_{i4}, fs_{i5}]$, where fs_{ij}, is the j^{th} feature score of i^{th} sentence ST_i. Computation of score associated with each of the sentence using feature scores is explained in the following section.

185 *3.2.6. Feature weight computation*

For extractive summarization, in order to prioritize sentences, appropriate sentence ranking mechanisms are required. Many research work try to produce an optimal ranking function to rank the sentences based on features, thereby producing the summary using high ranked sentences, but all these efforts were not fully successful due to the lack of efficient mechanism to provide appropriate weights for the features. For n' sentences in STR, a feature matrix F of order n'

x 5 is constructed using the five scores computed for each sentence. The feature scores are normalized to avoid the dominance of features having higher magnitude over others. The Feature Matrix F is multiplied with a weight vector W given by $W^T = [w_1, w_2, w_3, w_4, w_5]$, where $w_i \epsilon$ [0,1]. For ranking the sentences, an aggregate single score y_i is to be computed for each sentence by applying weights to the five features. The aggregate score vector $Y = [y_1, y_2, ..., y_{n'}]^T$ is computed as below

$$Y = \begin{bmatrix} y1 \\ y2 \\ ... \\ ... \\ y_{n'} \end{bmatrix} = \begin{bmatrix} fs_{11} & fs_{12} & fs_{13} & fs_{14} & fs_{15} \\ fs_{21} & fs_{22} & fs_{23} & fs_{24} & fs_{25} \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ fs_{n'1} & fs_{n'2} & fs_{n'3} & fs_{n'4} & fs_{n'5} \end{bmatrix} \cdot \begin{bmatrix} w1 \\ w2 \\ w3 \\ w4 \\ w5 \end{bmatrix} \tag{13}$$

Identifying the most appropriate weight values for features is still an unanswered problem. Many researchers used supervised learning process which is time consuming to assign appropriate weights for the features in sentence ranking. In reality, the sentence scoring using weight values depends heavily on the nature of documents, knowledge and vocabulary of the author and style of writing. As the matter is subjective, varying from individuals to individuals, finding the most appropriate weights to features becomes very difficult. To overcome this problem, feature weights (W) are maintained randomly. A problem independent and a multicriteria optimization process is employed to find a good solution by exploring the solution space. In the following section we discuss the multicriteria optimization method used to find a good extractive summary using three objective functions.

3.3. Multicriteria Optimization

Generating an extractive summary of multiple documents is a complex problem with multiple solutions. Multiple summaries generated can be evaluated according to various criteria based on statistical and semantic features. A summary which is better according to one criterion, can be worse according to another. Consequently, there is no unique optimal solution but rather a set of incomparable alternatives or trade-offs exist. The field of multicriteria optimization deals with all the methods which take into account more than one objective function in the problem modelling, and seek to find one particular or a representative set of the trade-offs. Multicriteria optimization aims for covering or approximating the efficient solution set.

Generally, statistical measures are used to produce extractive summary which might not guarantee the production of best summary. Most of the statistical techniques used in the extractive summarization systems cannot estimate the impact of contextual usage of words meaning in passages of discourse. Most of the statistic based summarization systems treat every sentence as independent units of processing while they are correlated in reality. Extractive summarizer strive to generate a summary that is relevant and cover the entire concept of the document. Quality of the summary can be enhanced by exploiting semantic aspects of the documents. Hence we approached this summarization task as a multicriterion optimization problem which does not have an exact single solution but multiple possible solutions from which best one is selected. In this work three objective functions covering the statistical and semantic aspects are used independently or in groups to generate the summary. Using population-based approach three candidate summaries which has maximum values corresponding to the three objective functions are generated. A final required size summary is generated by selecting the sentences which are common in more than one candidate summary and if the required size is not met, then the high ranked sentences from the remaining sentences in the three candidate summaries are selected.

In the following subsections we discuss in detail the representation of summaries as chromosomes, formulation of three objective functions and population-based multicriteria optimization (MCO) method to generate the single best summary.

3.3.1. Chromosome Encoding

In genetic algorithms chromosomes represent the candidate solutions of a problem. Each chromosome is subdivided into genes, where genes contribute significantly to the solution. In this work, each chromosome represents a candidate summary or partial candidate summary and each gene is the index of a sentence in the document set. The size of the chromosome which is the number of genes in the chromosome is represented by R and it is determined by the size of the summary to be generated and the number of objectives functions used. The required summary size is specified by the end-user and the required size summary can be generated by controlling the size of the chromosome R, which in turn depends upon the number of objective functions used for the optimization. R can be computed as $R = \theta \ * \ Required\ summary\ size$, where value of θ is to be selected based on the number of objective functions. If only one objective function is used for the optimization, then the size of chromosome R should be equal to the number of sentences required for the summary, so the value of θ is 1. If a combination of functions is used for optimization then by keeping θ to 1, we may get a summary of length at most n_f times the *required summary size*, where n_f is the number of objective functions used in the combination. If the size of chromosome R is high, then selection of appropriate sentences from global optimal solutions become a critical problem due to the existence of a large number of sentences in feasible summaries. If θ has a very small value then required size summary cannot be generated from the chromosomes. A detailed analysis is performed to find the most appropriate value of θ. Experimental analysis using DUC 2002 dataset and evaluation of system generated summary with ROUGE tool kit is performed to find out appropriate value for θ, for different combination of objective functions, and it is observed that better Precision and Recall values are obtained with the value 0.65 for θ when three objective functions are used for optimization and with a value of 0.8 when two objective functions are used.

3.3.2. Population generation by crossover

Population is a collection of chromosomes in which each chromosome is a candidate summary for generating an optimized summary. For each generation the population is created by identifying two most appropriate parent chromosomes and subsequently the children are generated by the crossover operation on the parent chromosomes. This work maintains a randomness in selecting the parent chromosomes in every generation, which results in the production of diverse population for every generation and it ensures the search on the global solution space.

Since each chromosome represents the candidate sentences for the summary, the parent chromosomes of each generation is constructed by the feature based ranking scheme with random feature weight using equation (13). A sorted score vector X, which contains the indices of sentences in the original document set, sorted based on their scores in descending order is created. For each generation, two parent chromosomes are constructed in such a way that the first R sentence indices of X forms the first parent chromosome and the following R sentence indices forms the second one. Thus the feature based sentence ranking act as a guide for identifying the parent chromosomes. The population is created by adding two child chromosomes in every iteration by random multipoint crossover of the parent chromosomes until no new children are added.

3.3.3. Formulation of Objective Functions

Objective functions determine the fitness of an individual chromosome. This fitness indicates the goodness of the solution and closeness of the solution to the optimal one. The goodness of the summary is purely subjective and can be considered as a multicriterion optimization problem where it is difficult to determine whether one solution for one criterion is better than another solution for another criterion. Multicriterion optimization problem does not have an exact single solution but multiple possible solutions from which best one is to be selected. In this work in order to include the statistical and semantic aspects in the summarization process, three objective functions are coined based on the summary centroid- document set centroid relationship and words/sentence-topic relationships. First function takes care of centroid relationship and next two functions cover the semantic structure of documents using Latent semantic analysis (LSA) based Singular Value Decomposition (SVD) (Steinberger & Ježek, 2005) and Non Negative Matrix Factorization (NMF). Experiments in (Stevens et al., 2012) indicates that SVD provides an accurate, stable, and reliable approximation to human judgements of similarity and relatedness between word pairs in comparison to other topic models but in terms of topic coherence scores its performance is poor. SVD represent a sentence as a linear combination of many non-intuitive and less important semantic feature vectors and NMF represent a sentence as a linear combination of few intuitive and directly related semantic feature vectors. NMF provides better performance in identifying subtopics of a document than SVD (Lee et al., 2009). To get the advantages of both these aspects two functions based on them are also used for optimization.

These three objective functions are non-commensurable functions which cannot be combined to a single meaningful scalar function. So they are used independently to generate three candidate summaries, each give maximum value for the respective function and final summary is generated by selecting required number of relevant sentences from the candidate summaries. Following subsections deal with the formulation of three objective functions used in our work.

3.3.4. Centroid based Objective Function f1

A summary is meaningful if (i) the relevance of the sentence in the summary with respect to the entire document set is maximum (ii) the overall concepts in the summary cover the entire concept of the document set (iii) the redundancy within the summary is minimum. First objective function is based on the work of Alguliev et al (Alguliev et al., 2013) a statistical approach and is given as,

$$f1 = \frac{f_{cover}(S_{chrom})}{f_{diver}(s_{chrom})}, \quad \textit{where } S_{chrom} \textit{ denote a chromosome.} \tag{14}$$

The function f_{cover} (S_{chrom}) is intended to capture the main contents of the input document set and is defined as

$$f_{cover}(S_{chrom}) = Sim(O, O^s) \ . \sum_{i \ \epsilon \ S_{chrom}} Sim(O, S_i) \tag{15}$$

where O is the mean vectors of the entire sentence set and O^s is the mean vector of the summary S_{chrom}. The diversity function is defined as follows:

$$f_{diver}(S_{chrom}) = \sum_{i \ \epsilon \ S_{chrom}} \ \sum_{j \ \epsilon \ S_{chrom}} Sim(S_i, S_j) \tag{16}$$

Lower value of $f_{diver}(S_{chrom})$ corresponds to higher novelty in the summary. Now we have the first objective function as:

$$f1 = \frac{Sim(O, O^s) \ . \ \sum_{i \ \epsilon \ S_{chrom}} Sim(O, S_i)}{\sum_{i \ \epsilon \ S_{chrom}} . \sum_{i \ \epsilon \ S_{chrom}} Sim(S_i, S_j)} \tag{17}$$

3.3.5. LSA (SVD) based Objective function f2

The centroid based objective function only checks the similarity between the centroid concepts in the summary and the entire document set and diversity of sentences in the summary. The second objective function is introduced to maximize the semantic coverage of the sentences in candidate summaries. This function is based on Latent Semantic Analysis (LSA). LSA is a fully automatic mathematical/statistical technique for extracting and representing the contextual usage of word's meanings in passages of discourse. LSA produces measures of word-word, word-passage and passage-passage relations that are well correlated with several human cognitive phenomena involving association or semantic similarity (Dennis et al., 2003). Latent Semantic Analysis is an application of Singular Value Decomposition for text summarization. The basic idea of LSA is to find out the similarity of meaning of words in the documents using word co-occurrence relationship and represent each sentence as a linear combination of many non-intuitive and less important semantic feature vector. LSA consists of mainly two steps. In the first step, the entire document is represented as term by sentence matrix A, where each column of the matrix A represents the PNVA_TFISF vector of a sentence in stemmed sentence set STR.

In the second step Singular Value Decomposition (SVD) is applied on matrix A. SVD can capture interrelationships among terms, so that terms and sentences can be clustered on a semantic basis rather than on the basis of words alone and a weight is assigned to each sentence based on the semantic clustering.

$$A = U \cdot \Sigma \cdot V \tag{18}$$

Matrix V describes the polarity of correlation between the sentences and concepts. In the SVD decomposition V may contain negative values which also indicates the semantic relationship between the sentence and reflections of the concepts identified by SVD. By exploiting the concept-sentence relationship from V, the SVD semantic score for a sentence is given by :

$$Sc_j = \sum_{i=1}^{nc} |V_{ij}| \tag{19}$$

where nc is the no of relevant concepts used for the scoring which is much less than n. In this work it is taken as length of R. V_{ij} is the weight of i^{th} semantic concept of j^{th} sentence. and objective function f2 is formulated as follows:

$$f2 = \sum_{i \epsilon S_{chrom}} Sc_i \tag{20}$$

This objective function is used to find a summary with maximum semantic
315 coverage.

3.3.6. NMF base Objective function f3

To improve the semantic quality of the summary to be generated, third objective function based on Non-Negative Matrix Factorization(NMF) is employed. NMF groups the terms into different semantic features and represent each sentence as a linear combination of few intuitive and directly related semantic feature vectors (Lee et al., 2009). NMF identifies subtopics better than the SVD. The entire document is represented as term by sentence matrix B, where each column of the matrix B represents the PNVA_TF vector of a sentence in stemmed sentence set STR, NMF decomposes B into two matrices given by :

$$B_{m'Xn} \approx W_{m'Xnc} \cdot H_{ncXn} \tag{21}$$

where W is the non negative Semantic Feature matrix of order $m'Xnc$ and H is the Non Negative semantic variable matrix H of order $nc\ X\ n,\ nc < min(m', n)$ is the number of concepts to be found out. Both H and W matrices are used to find, $Weight(W_{*i})$ the weight of i^{th} semantic feature based on the weight of terms present in it, and $Weight(H_{*i})$ the relevance of i^{th} semantic features across the sentences in document set.

$$weight(W_{*i}) = \sum_{j=1}^{m'} W_{ji} \tag{22}$$

$$weight(H_{i*}) = \frac{\sum_{j=1}^{n} H_{ij}}{\sum_{(i=1)}^{nc} \sum_{j=1}^{n} H_{ij}} \tag{23}$$

The NMF score of a j^{th} sentence using the above mentioned measures is computed as follows:

$$Sc_j = \sum_{i=1}^{nc} (H_{ij}\ .\ weight(W_{*i})\ .\ weight(H_{i*}) \tag{24}$$

Now objective function f3 is formulated as follows:

$$f3 = \sum_{j \epsilon S_{chrom}} Sc_j \tag{25}$$

As explained earlier, objective functions are non-commensurable functions and therefore three optimal candidate summaries are generated by maximising three objective functions independently. The proposed multi-document summarization procedure named MCO-CLN is given in Algorithm 1. C, L, N stands for the centroid, LSA and NMF based methods using which the three objective functions are formulated. Final summary is generated by selecting common sentences in the candidate summaries and the remaining sentences from the high ranked sentences in these candidte summaries to make the required size.

Algorithm 1: Multicriteria Optimization Algorithm

Procedure : MCO-CLN (f1, f2,f3)

1. Let $GC1, GC2, GC3$ represent the global best summaries corresponding to

the three objective functions f1, f2, f3 respectively and are initialised with empty sets.

2. Let $PC1, PC2, PC3$ represent the local best summaries corresponding to the objective functions f1, f2 and f3 in every generation and are initialized with empty sets.

3. Initialise weight vector $W^T = [w1, w2, w3, w4, w5]$ with random, $wi \epsilon [0, 1]$.

4. Calculate scores for sentences using equation (13)

5. Construct a sorted score vector X containing indices of sentences sorted based on the its scores in Y.

6. Construct two parent chromosomes, $p1 := X[1 : R]$ and $p2 := X[R + 1 : 2R]$

7. Perform *Build_Next_Population(p1,p2)* to create a new generation of chromosomes

8. Select from the new generation three chromosomes PC1, PC2 and PC3 corresponding to the three local best summaries with maximum values for the functions f1, f2, and f3 respectively.

9. If local best summaries PC1, PC2 and PC3 are better than the respective global best summaries GC1, GC2 and GC3 replace the global ones with respective local ones.

10. Repeat steps 3 to 9 until the global best summaries converge.

11. Generate final summary as $(GC1 \cap GC2 \cap GC3) \cup (GC1 \cap GC2) \cup (GC1 \cap GC3) \cup (GC2 \cap GC3) \cup (required\ high\ ranked\ sentences\ from\ the\ remaing\ sentences\ in\ (GC1 \cup GC2 \cup GC3))$

12. Return final summary

13. End **procedure**

Algorithm 2: Create a new generation by cross over operations

---

procedure : *Build_Next_Population(p1,p2)*

1. Initialise the set for population, P with empty set

2. Generate a vector V of size R and initialise its elements randomly from set

{ 0, 1}

3. Initialise two vectors C1,C2 as C1 := p1 and C2 := p2

4. For j = 1 to R

   if (V(j) = 1) then $C1(j) \leftrightarrow C2(j)$

5. Add the child chromosomes C1, C2 to P and remove duplicate entries from P

6. Repeat steps 2 to 5 until no more new children are added to the population

7. Return P the Generated Population

8. End **procedure**

Experimental evaluation of the proposed method is given below

4. Experimental Evaluation

4.1. Experimental Data

Performance of proposed system is evaluated using open benchmark datasets from Document Understanding Conference(DUC) DUC2002, DUC 2004 and DUC 2006. DUC 2002 dataset provides their model summaries of varying size such as 50 words, 100 words, 200 words and 400 words against which the summary generated by our system can be compared. The dataset contains four categories of documents: documents about a single natural disaster event(Category 1) and created within at most seven day window, documents about a single event in any domain created with at most a seven day window(Category 2), document about multiple distinct event of a single type (no limit on the time window) (Category 3)and documents about biological information mainly about a single individual(Category 4). Evaluation is done using each category of data and average score is computed. DUC 2004 dataset consist of data given for five different tasks. For our evaluation, we have used the data provided for the task 2, which consist of 50 clusters of 10 documents each and a generic summary of 665 bytes per cluster. Average score of 50 clusters is taken as result. In DUC

2006 dataset, there are 50 datasets with an average of 25 documents per set and system generated summaries of 250 words for each set.

4.2. Evaluation Metrics

For evaluation of our system generated summaries, N-gram co-occurrence statistical measure of ROUGE toolkit is used (Lin, 2004). The formulation of ROUGE-N is defined as follows :

$$ROUGE - N = \frac{\sum_{S \epsilon Summ_{ref}} \sum_{N-gram \epsilon S} Count_{match}(N - Gram)}{\sum_{S \epsilon Summ_{ref}} \sum_{N-gram \epsilon S} Count(N - Gram)} \tag{26}$$

where N is the size of the N-gram which can be unigram, bigram or trigram. $Count_{match}(N - Gram)$ is the number of N-grams occurring in both candidate and reference summaries and count(N-Gram) is the number of N-Grams in the reference summary. The updated ROUGE evaluation methods can generate three types of scores for a system generated summary such as Recall, Precision, and F-measure.

4.3. Results

Summaries for DUC 2002, DUC 2004 and DUC 2006 datasets are generated using the multicriteria optimization algorithm with the three objective functions individually, all pairs of functions and all the three functions together. Detailed analysis is performed using generated summaries to see whether summaries generated by individual functions or combinational functions are better. The average ROUGE-1 and ROUGE-2 score are computed for summaries generated from all data sets using individual and combinational functions.

Dataset based ROUGE results using individual functions.

The results obtained by using the objective functions f1, f2, f3 individually are given in Table 1, Table 2, and Table 3 respectively. It can be observed from the results that summaries generated using function f2 has highest ROUGE scores. In ROUGE, evaluation is done by comparing system generated summaries with human generated summaries. Human generates summaries by considering the

Table 1: ROUGE-1 and ROUGE-2 scores obtained for DUC 2002, DUC 2004 and DUC 2006 dataset using centroid function f1 **MCO-CLN (f1)**)

	Average ROUGE-1 Score			Average ROUGE-2 Score		
Dataset	Recall	Precision	F measure	Recall	Precision	F measure
DUC 2002	0.5311	0.5216	0.5259	0.2339	0.2283	0.2308
DUC 2004	0.45601	0.46906	0.46020	0.12109	0.13338	0.1263
DUC 2006	0.360163	0.303089	0.328065	0.062024	0.055366	0.058312

Table 2: ROUGE-1 and ROUGE-2 scores obtained for the datasets DUC 2002, DUC 2004 and DUC 2006 using the LSA (SVD) based function f2 (**MCO-CLN(f2)**)

	Average ROUGE-1 Score			Average ROUGE-2 Score		
Dataset	Recall	Precision	F measure	Recall	Precision	F measure
DUC 2002	0.5524	0.5321	0.5414	0.2565	0.2448	0.2502
DUC 2004	0.5264	0.4992	0.5122	0.1664	0.1669	0. 0.1666
DUC 2006	0.3865	0.2975	0.3352	0.0730	0.0608	0.6628

Table 3: ROUGE-1 and ROUGE-2 scores obtained for DUC 2002, DUC 2004 and DUC 2006 datasets using NMF based function f3 (**MCO-CLN(f3)**)

	Average ROUGE-1 Score			Average ROUGE-2 Score		
Dataset	Recall	Precision	F measure	Recall	Precision	F measure
DUC 2002	0.5477	0.5284	0.5374	0.2559	0.2472	0.2513
DUC2004	0.5018	0.4966	0.4991	0.1487	0.1564	0.1524
DUC2006	0.3699	0.2985	0. 0.3302	0.0648	0.0566	0.0604

explicit representation of terms, contextual meaning of the terms and relationship between the terms. Overall score of summaries generated by function f2 uses the semantic scores of sentences provided by LSA (SVD) using the word sentence relationship matrix. Besides having the capability of modelling relationships among words and sentences, SVD has the capability of noise reduction, which helps to improve accuracy. Hence the ROUGE scores obtained by function f2 is better than functions f1 and f3.

Summaries generated by f3 function is better compared to f1 since NMF is more capable to identify directly related semantic features. But these summaries are not good as LSA based method, because LSA is more capable to generate summaries which are reliable approximation of human generated summaries by using non intuitive semantic features than NMF.

ROUGE results for pairs of objective functions on datasets.

Recall and precision values obtained for ROUGE-1 and ROUGE-2 for sum-

Table 4: ROUGE-1 and ROUGE-2 scores obtained for DUC 2002, DUC 2004 and DUC 2006 datasets using the combined Centroid and LSA functions (f1&f2) (**MCO-CLN (f1f2)**)

	Average ROUGE-1 Score			Average ROUGE-2 Score		
Dataset	Recall	Precision	F measure	Recall	Precision	F measure
DUC 2002	0.5532	0.5316	0.5419	0.2586	0.2467	0.2524
DUC 2004	0.5274	0.5096	0.5181	0.1671	0.1699	0.1684
DUC 2006	0.4593	0.2589	0.3249	0.08876	0.05581	0.06739

Table 5: ROUGE-1 and ROUGE-2 scores obtained for DUC 2002, DUC 2004 and DUC 2006 datasets using Centroid and NMF functions (**MCO-CLN(f1f3**)

	Average ROUGE-1 Score			Average ROUGE-2 Score		
Dataset	Recall	Precision	F measure	Recall	Precision	F measure
DUC 2002	0.5512	0.5287	0.5392	0.2576	0.2473	0.2521
DUC 2004	0.5221	0.4967	0.5088	0.1667	0.1677	0.1671
DUC 2006	0.4429	0.2663	0.3278	0.0842	0.0569	0.0670

Table 6: ROUGE-1 and ROUGE-2 scores obtained for DUC 2002, DUC 2004 and DUC 2006 datasets using combination of LSA and NMF functions f2 and f3 (**MCO-CLN (f2f3)**)

	Average ROUGE-1 Score			Average ROUGE-2 Score		
Category	Recall	Precision	F measure	Recall	Precision	F measure
DUC 2002	0.5471	0.5284	0.5365	0.2585	0.2485	0.2527
DUC 2004	0.5221	0.4983	0.5095	0.1642	0.1662	0.1650
DUC 2006	0.4437	0.2697	0.3289	0.086222	0.05822	0.068169

maries generated with pairs of functions f1f2, f1f3 and f2f3 are given in Table 4, Table 5, and Table 6 respectively. It can be observed that summaries generated by considering the functions f1 and f2 together are better than the summaries generated by the individual functions f1 or f2 or f3 and other paired combinational functions f1f3 and f2f3. Reason is that, combined f1f2 function generates summaries by considering the sentences in the input documents which satisfy centroid, novelty and semantic features together.

From Table 5 it is observed that ROUGE-1 score of f1f3 function pair is less than the scores of f1f2 and f2 but it is better than the scores obtained for function f1 or f3 independently. This indicates that summary generated by LSA is more closer to human generated summary than summary generated by other individual functions and other function pairs without f2. ROUGE scores obtained by the combinational function of f2f3 is less than that of individual functions f1, f3 and other pairs of functions because here summaries are generated by considering semantic aspect alone without considering other aspects explicitly. But the ROUGE-2 scores provided by these combinational function are better than all other individual functions since LSA (SVD) and NMF can consider both indirectly and directly related semantic features or word combinations by finding the relationship between words and sentences.

ROUGE results obtained for the combined objective functions f1f2f3 on datasets

Table 7: ROUGE-1 and ROUGE-2 scores obtained for DUC 2002, DUC 2004 and DUC 2006 datasets using combination of Centroid, LSA and NMF functions f1,f2 and f3 (**MCO-CLN f1f2f3**))

	Average ROUGE-1 Score			Average ROUGE-2 Score		
Datasets	Recall	Precision	F measure	Recall	Precision	F measure
DUC 2002	0.554	0.5423	0.5487	0.277	0.2679	0.2718
DUC 2004	0.5280	0.5159	0.5214	0.1694	0.1749	0.1719
DUC 2006	0.4673	0.2539	0.3254	0.0931	0.05584	0.06902

ROUGE-1 and ROUGE-2 scores obtained for summaries generated by the combination of three objective functions f1f2f3 are given in Table 7. It is observed that scores of f1f2f3 are better than scores of all individual functions and all other pairs of functions since this combination considers all the features coverage, novelty and semantic aspects which is very much essential in multidocument summaries.

Comparison of our best method, combination of f1f2f3, with our other methods are shown in Table 8. Here the percentage of relative improvement is used for the comparison which is calculated using the formula:
$(((score\ of\ (f1f2f3) - scoreof(other\ method))/(score\ of\ other\ method)*100)$.
"+" in the table indicate that the result outperforms and "-" indicates the opposite. From Table 8 it is evident that summaries generated by combined function f1f2f3 gives good ROUGE-1 and ROUGE-2 scores. All the entries in the table are "+", so combined f1f2f3 function is better than individual and all other pairs of functions.

Further we compared our method with other popular methods, whose results are taken from Table 2 of (Alguliev et al., 2013) and the results are given in Table 9 and Table 10. The method we have taken for comparison are **DUCbest**, **Random**, **NMF**, **WCS**, **WF-NMF**. Method **DUCbest** takes the scores of the best participating systems in DUC, **Random** selects sentences for the summary through a random selection, **NMF** ranks sentences based on the weighted score

Table 8: Percentage wise relative improvement of our function f1f2f3 with respect to individual and all other pairs of functions

Dataset	Method	Metric	f1	f2	f3	f1f2	f1f3	f2f3
DUC 2002	f1f2f3	ROUGE-1	+4.133	+0.29	+1.137	+0.144	+0.505	+1.25
DUC 2002	f1f2f3	ROUGE-2	+15.56	+7.40	+7.62	+6.64	+7.00	+6.68
DUC 2004	f1f2f3	ROUGE-1	+13.64	+0.30	+4.96	+0.11	+1.12	+1.12
DUC 2004	f1f2f3	ROUGE-2	+28.52	+1.77	+12.22	+1.36	+1.59	+3.069
DUC 2006	f1f2f3	ROUGE-1	+22.94	+17.29	+20.84	+1.71	+5.22	+5.05
DUC 2006	f1f2f3	ROUGE-2	+33.37	+21.59	+30.39	+4.66	+9.56	+7.41

obtained by the application of Non Negative Matrix Factorization on term by sentence matrix, **Centroid** based system ranks sentences based on the features such as centroid, position and first sentence similarity. **WCS** combines single document summary of multiple documents and **WF-NMF** is an extension of NMF method. It can be observed from Table 9 and Table 10 that our method is better than other popular methods.

Table 11, Table 12, Table 13 and Table 14 gives the percentage relative improvement of our method for ROUGE-1 and ROUGE-2 score respectively. It can be observed that the percentage improvement of the proposed method for the ROUGE scores are better than other methods. Hence most of the entries in the tables are '+' .

From table 12 it is clear that, for some individual and pairs of our functions, system exhibit a lower performance (-ve values) than other famous methods for the ROUGE-2 scores. However, all the values corresponding to f1f2f3 are all '+ve' which show that combinational functions with f1, f2 and f3 together gives much better performance than the other popular methods.

From Table 9, Table10, Table 11, Table 12, Table 13and Table 14 it is evident that our method outperforms other popular methods. Better performance of

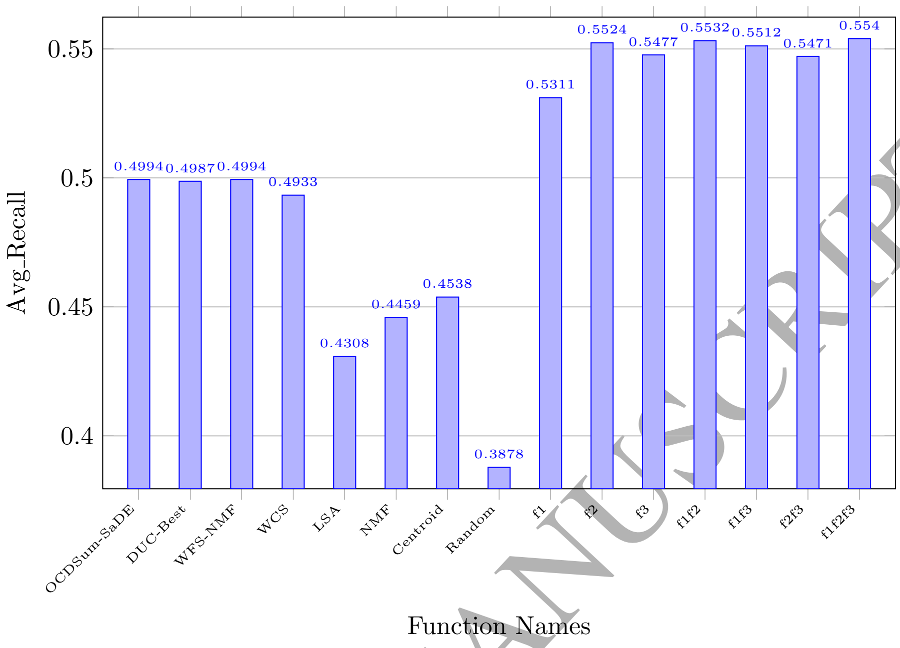


Figure 3: Comparison of ROUGE-1 scores of the methods on DUC 2002 dataset

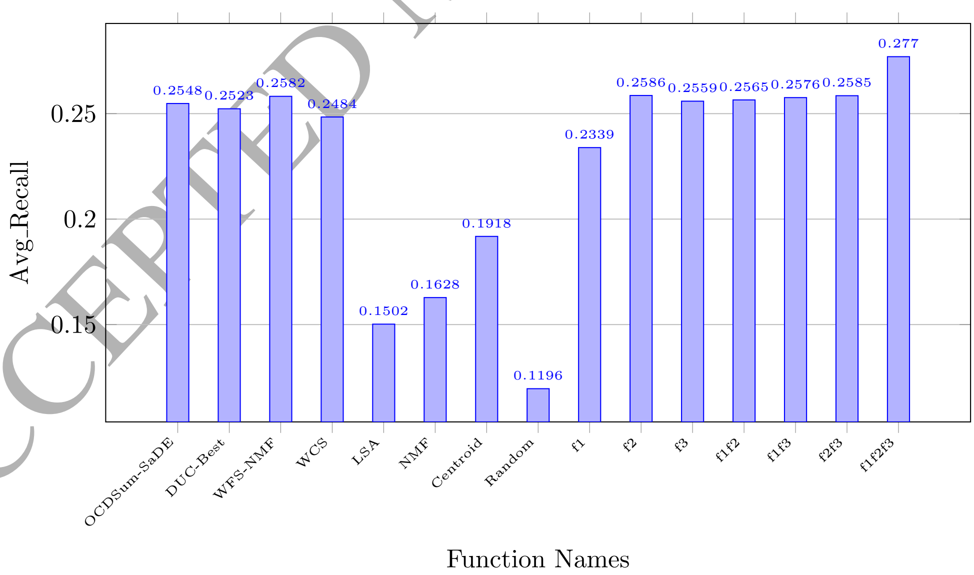


Figure 4: Comparison of ROUGE-2 scores of the methods on DUC 2002 dataset

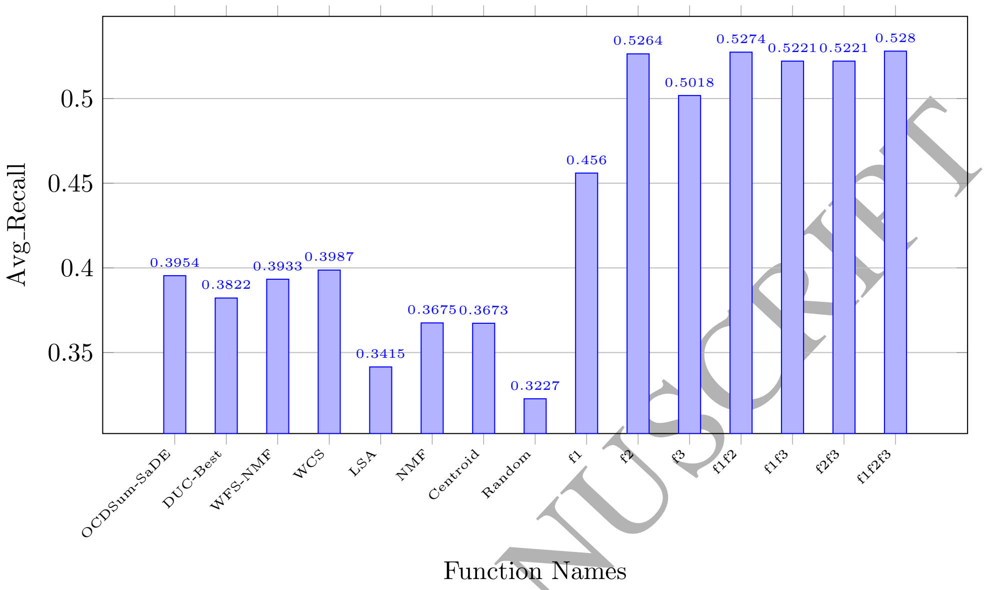


Figure 5: Comparison of ROUGE-1 scores of the methods on DUC 2004 dataset

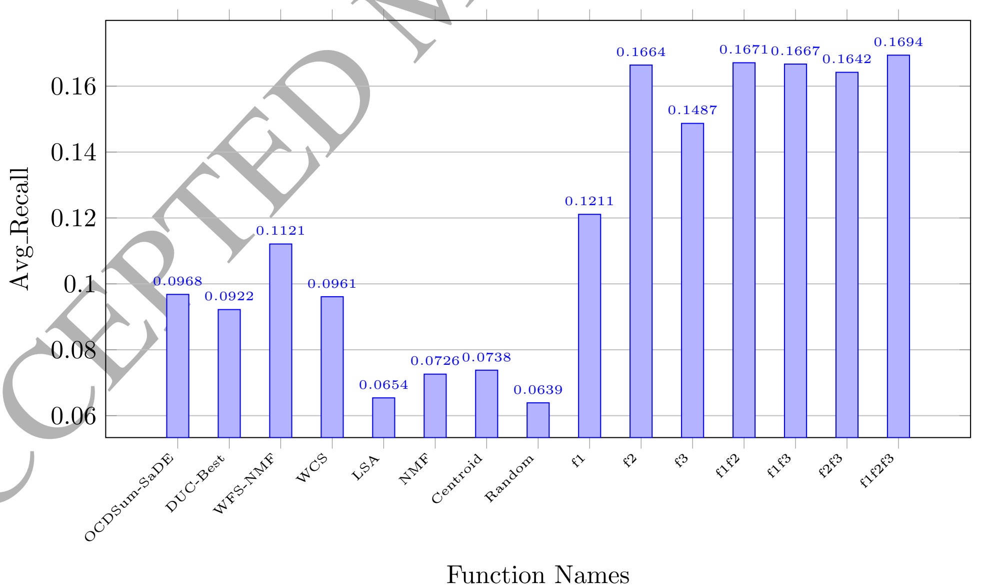


Figure 6: Comparison of ROUGE-2 scores of the methods on DUC 2004 dataset

Table 9: ROUGE-1 and ROUGE-2 scores of our methods with other popular methods on DUC 2002 dataset

Method	ROUGE-1	ROUGE-2
OCDSum-SaDE	0.499	0.2548
DUC-best	0.4987	0.2523
WFS-NMF	0.4994	0.2582
WCS	0.4933	0.2484
LSA	0.4308	0.1502
NMF	0.4459	0.1628
Centroid	0.4538	0.1918
Random	0.3878	0.1196
MCO-CLN f1	**0.5311**	**0.2339**
MCO-CLN f2	**0.5524**	**0.2565**
MCO-CLN f3	**0.5477**	**0.2559**
MCO-CLN f1f2	**0.5532**	**0.2586**
MCO-CLN f1f3	**0.5512**	**0.2576**
MCO-CLN f2f3	**0.5471**	**0.2585**
MCO-CLN f1f2f3	**0.554**	**0.277**

our system is due to the 1) Representation of sentences by PNVA vector model, 2) Initialization of weight vector W with random values in every generation to overcome the problem associated with the prediction of weight values associated with the features, 3) Creation of most promising parent chromosomes in each generation, 4) Application of population-based methods to generate a diverse global solution space in each iteration, 5) Application of multicriteria optimization method by selecting and utilizing the objective functions that take care of semantic aspects at different levels, 6) Appropriate selection of sentences from three candidate summaries in each iteration to find the final optimal summary. An example of summary produced and its corresponding reference summary by our multi-document method is given for Topic D3001T of DUC 2004 dataset.

Table 10: ROUGE-1 and ROUGE-2 scores of our methods with other popular methods on DUC 2004 dataset

Method	ROUGE-1	ROUGE-2
OCDSum-SaDE	0.3954	0.0969
DUC-best	0.3822	0.0922
WFS-NMF	0.3933	0.1121
WCS	0.3987	0.0961
LSA	0.3415	0.0654
NMF	0.3675	0.0726
Centroid	0.3673	0.0738
Random	0.3227	0.0639
MCO-CLN f1	**0.45601**	**0.121**
MCO-CLN f2	**0.5264**	**0.1664**
MCO-CLN f3	**0.5018**	**0.1487**
MCO-CLN f1f2	**0.5274**	**0.1671**
MCO-CLN f1f3	**0.5221**	**0.1667**
MCO-CLN f2f3	**0.5221**	**0.1642**
MCO-CLN f1f2f3	**0.5280**	**0.1694**

5. Conclusion

In this work, we have explored a feature based automatic generic extractive multi-document summarization method, which generates summary with maximum relevance and novelty. This process considered summarization task as a multicriteria optimization problem, which is solved by coining three objective functions to explore i) the centroid relationship between document and summary ii) similarity and relatedness between words and iii) indirectly related semantic features. To solve the optimization problem, population-based method is

Table 11: Percentage improvement of proposed method for ROUGE-1 score on DUC 2002 dataset.

Methods	Percentage improvement of MCO with other methods on DUC 2002							
Method	OCDSum-SaDE	DUC-best	WFS-NMF	WCS	LSA	NMF	Centroid	Random
f1	+6.044	+6.1	+5.96	+7.12	+18.89	+16.042	+14.55	+26.98
f2	+9.67	+9.72	+9.59	+10.69	+22.01	+19.27	+17.85	+29.79
f3	+8.89	+8.95	+8.82	+9.93	+21.34	+18.59	+17.14	+29.19
f1f2	+9.79	+9.85	+9.72	+10.82	+22.12	+19.39	+17.97	+29.89
f1f3	+9.47	+9.52	+9.39	+10.50	+21.84	+19.10	+17.67	+29.64
f2f3	+8.79	+8.85	+8.72	+9.83	+21.26	+18.49	+17.05	+29.12
f1f2f3	+10.57	+10.63	+10.50	+11.59	+22.79	+20.08	+18.67	+30.50

Table 12: Percentage improvement of proposed method for ROUGE-2 score on DUC 2002 dataset

Methods	Percentage improvement of MCO with other methods on DUC 2002							
Method	OCDSum-SaDE	DUC-best	WFS-NMF	WCS	LSA	NMF	Centroid	Random
f1	-8.94	-7.87	-10.39	-6.19	+35.78	+30.39	+17.99	+48.87
f2	+0.66	+1.64	-0.66	+3.16	+41.44	+36.53	+25.22	+53.37
f3	+0.43	+1.41	-0.89	+2.93	+41.31	+36.38	+25.05	+53.26
f1f2	+1.47	+2.44	+0.15	+3.94	+41.92	+37.05	+25.83	+53.75
f1f3	+1.09	+2.06	-0.23	+3.57	+41.69	+36.80	+25.54	+53.5
f2f3	+1.43	+2.39	+0.12	+3.90	+41.89	+37.02	+25.80	+53.73
f1f2f3	+8.01	+8.92	+6.79	+10.32	+45.77	+41.23	+30.76	+56.82

adopted. Three summaries that maximize the value of the corresponding objective function are identified by generating a diverse population of candidate summaries from the most promising parents in each iteration. Final optimal summary is developed from the three candidate summaries obtained after the iteration process.

Table 13: Percentage improvement of proposed method for ROUGE-1 score on DUC 2004 dataset.

Methods	Percentage improvement of MCO with other methods on DUC 2004							
Method	OCDSum-SaDE	DUC-best	WFS-NMF	WCS	LSA	NMF	Centroid	Random
f1	+13.29	+16.19	+13.75	+12.57	+25.11	+19.41	+19.45	+29.23
f2	+24.89	+27.39	+25.29	+24.26	+35.13	+30.19	+30.22	+38.69
f3	+21.20	+23.83	+21.62	+20.55	+31.95	+26.76	+26.80	+35.69
f1f2	+25.03	+27.53	+25.43	+24.40	+35.25	+30.32	+30.36	+38.81
f1f3	+24.27	+26.79	+24.67	+23.64	+34.59	+29.61	+29.65	+2138.19
f2f3	+24.27	+26.79	+24.67	+23.64	+34.59	+29.61	+29.65	+38.19
f1f2f3	+25.11	+27.61	+25.51	+24.49	+35.32	+30.39	+30.44	+38.88

Table 14: Percentage improvement of proposed method for ROUGE-2 score on DUC 2004 dataset

Methods	Percentage improvement of MCO with other methods on DUC 2004							
Method	OCDSum-SaDE	DUC-best	WFS-NMF	WCS	LSA	NMF	Centroid	Random
f1	19.92	23.81	0	20.58	45.95	40	39.01	47.19
f2	41.87	44.69	27.41	42.35	60.77	56.45	55.73	61.67
f3	34.84	37.99	18.63	35.37	56.02	51.18	50.37	57.028
f1f2	42.01	44.82	27.59	42.49	60.86	56.55	55.83	61.76
f1f3	41.87	44.69	27.42	42.35	60.77	56.45	55.73	61.67
f2f3	40.99	43.85	26.31	41.47	60.17	55.79	55.06	61.08
f1f2f3	42.79	45.57	28.57	43.27	61.39	57.14	56.43	62.28

By the experimental evaluation of our method using ROUGE tool kit with DUC 2002, DUC 2004 and DUC 2006 datasets, we observed that proposed approach with combination of three objective functions generates good quality summary and performs better than previous state-of-the-art methods in terms of ROUGE scores. There are several possible direction of future research:

In extraction based summarization process feature based sentence extraction method is relevant to determine the extent to which the selected sentence can cover the entire collection of documents. Hence multi-document summarizer can be further improved by identifying more appropriate features associated with the text documents, in the feature extraction phase.

Sentence representation method also influences the overall efficiency of the summarizer and the quality of the summary. Hence investigation can be made to find new sentence representation methods, which are better than the PNVA vector model.

Quality of the optimal summary can be further improved by identifying new objective functions and by performing more semantic analysis.

References

Abdel Fattah, M., & Ren, F. (2008). Probabilistic neural network based text summarization. In *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE'08. International Conference on* (pp. 1–6). IEEE.

Alguliev, R. M., Aliguliyev, R. M., & Isazade, N. R. (2013). Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications*, *40*, 1675–1689.

Binh Tran, G. (2013). Structured summarization for news events. In *Proceedings of the 22Nd International Conference on World Wide Web Companion* WWW '13 Companion (pp. 343–348). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Blake, C. (2006). A comparison of document, sentence, and term event spaces. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 601–608). Association for Computational Linguistics.

Canhasi, E. (2014). *Graph-based models for multi-document summarization*. Ph.D. thesis University of Ljubljana.

Carbonell, J., & Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* SIGIR '98.

Chali, Y., Hasan, S. A., & Joty, S. R. (2009). A svm-based ensemble approach to multi-document summarization. In *Advances in Artificial Intelligence* (pp. 199–202). Springer.

Christensen, J., Mausam, Soderland, S., & Etzioni, O. (2013). Towards coherent multi-document summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1163–1173). Atlanta, Georgia: Association for Computational Linguistics.

Christensen, J., Soderland, S., Bansal, G., & Mausam (2014). Hierarchical summarization: Scaling up multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 902–912). Baltimore, Maryland: Association for Computational Linguistics.

Dennis, S., Landauer, T., Kintsch, W., & Quesada, J. (2003). Introduction to latent semantic analysis. In *Slides from the tutorial given at the 25th Annual Meeting of the Cognitive Science Society, Boston*.

Dunlavy, D. M., OLeary, D. P., Conroy, J. M., & Schlesinger, J. D. (2007). Qcs: A system for querying, clustering and summarizing documents. *Information processing & management*, *43*, 1588–1605.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Fattah, M. A., & Ren, F. (2009). Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. *Computer Speech & Language*, *23*, 126–144.

Filatova, E., & Hatzivassiloglou, V. (2004). Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization* (pp. 104–111).

Glavaš, G., & Šnajder, J. (2014). Event graphs for information retrieval and multi-document summarization. *Expert Systems with Applications*, .

Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4* (pp. 40–48). Association for Computational Linguistics.

Hennig, L., & Labor, D. (2009). Topic-based multi-document summarization with probabilistic latent semantic analysis. In *RANLP* (pp. 144–149).

John, A., & Wilscy, M. (2013). Random forest classifier based multi-document summarization system. In *Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in* (pp. 31–36). IEEE.

Kupiec, J. M., Pedersen, J. O., Chen, F. R., Brotsky, D. C., & Putz, S. B. (1998). Automatic method of generating feature probabilities for automatic extracting summarization. US Patent 5,778,397.

Lee, J.-H., Park, S., Ahn, C.-M., & Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, *45*, 20–34.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. volume 8.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*, 159–165.

Mani, I., & Maybury, M. T. (1999). *Advances in automatic text summarization* volume 293. MIT Press.

Marujo, L. (2015). *Event-based Multi-document Summarization*. Ph.D. thesis Carnegie Mellon University and Instituto Superior Técnico.

Marujo, L., Ling, W., Ribeiro, R., Gershman, A., Carbonell, J., de Matos, D. M., & Neto, J. a. P. (2015a). Exploring events and distributed representations of text in multi-document summarization. *Knowledge-Based Systems*, (pp. –).

Marujo, L., Portêlo, J., Ling, W., de Matos, D. M., Neto, J. P., Gershman, A., Carbonell, J., Trancoso, I., & Raj, B. (2015b). Privacy-preserving multi-document summarization. In *PIR15: Privacy-Preserving IR (SIGIR 2015 Workshop)*. ACM.

Marujo, L., Ribeiro, R., Martins de Matos, D., Neto, J. a., Gershman, A., & Carbonell, J. (2015c). Extending a single-document summarizer to multi-document: a hierarchical approach. In *Proceedings of the* 4^{th} *Joint Conference on Lexical and Computational Semantics* (pp. 176–181). Denver, Colorado: Association for Computational Linguistics.

Mei, J.-P., & Chen, L. (2012). Sumcr: a new subtopic-based extractive approach for text summarization. *Knowledge and information systems*, *31*, 527–545.

Otterbacher, J. C., Radev, D. R., & Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4* (pp. 27–36). Association for Computational Linguistics.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.*. Technical Report Stanford InfoLab.

Park, S., Lee, J.-H., Kim, D.-H., & Ahn, C.-M. (2007). Multi-document summarization based on cluster using non-negative matrix factorization. In *SOFSEM 2007: Theory and Practice of Computer Science* (pp. 761–770). Springer.

Pitler, E., Louis, A., & Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 544–554). Association for Computational Linguistics.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137.

Salton, G., & Lesk, M. E. (1965). The smart automatic document retrieval systemsan illustration. *Communications of the ACM*, *8*, 391–398.

Steinberger, J., & Ježek, K. (2005). Text summarization and singular value decomposition. In *Advances in Information Systems* (pp. 245–254). Springer.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952–961). Association for Computational Linguistics.

Willett, P. (2006). The porter stemming algorithm: then and now. *Program*, *40*, 219–223.

Sample System Generated Summary

The Sam Rainsy Party, in a statement released Friday, accused Hun Sen of being "unwilling to make any compromise" on negotiations to break the deadlock. HunSen's party won 64 of the 122 seats in parliament in July's national election, but not the two-thirds majority necessary to form a government on its own. The alleged assassination attempt came during massive street demonstrations organized by the opposition after Hun Sen's Cambodian People's Party narrowly won the election. Cambodian leader Hun Sen's ruling party and the two-party opposition had called on the monarch to lead top-level talks, but disagreed on its location. Hun Sen said Monday that the CPP and FUNCINPEC had agreed that the Senate would be half as large as the 122-seat National Assembly. Cambodian prime minister Hun Sen rejects demands of 2 opposition parties for talks in Beijing after failing to win a 2/3 majority in recent elections. Sihanouk refuses to host talks in Beijing. Hun Sen said his current government would remain in power as long as the opposition refused to form a new one. Sihanouk reported that Hun Sen responded positively to a request by Ranariddh for five of his key supporters to receive political pardons. Hun Sen has rejected the opposition's reservations, saying it would be inappropriate to hold a summit outside the country. FUNCINPEC has demanded from Hun Sen written guarantees for the safety of its members and activities as a precondition for re-entering negotiations. Ranariddh's party and opposition ally Sam Rainsy held back their support, claiming the CPP won due to fraud and intimidation. In his speech, Hun Sen blamed the violence on opposition leaders, saying the demonstrations instigated social and economic chaos. Ranariddh and his core supporters did not return to Cambodia until a few months before an election in July this year that the ruling party narrowly won. "A meeting outside Cambodia, as suggested by the opposition, could place all parties on more equal footing," said the statement. Ranariddh and Sam Rainsy have remained outside the country since the Sept. 24 ceremonial convening of parliament. Sam Rainsy, under investigation by a Phnom Penh court for his role in the demonstrations, has remained abroad. The remaining senators, he said, should be selected by a method agreed upon by the new government and the National Assembly. His assurances come a week before the first session of Cambodia's new parliament, the National Assembly. "But the ruling party refuses to negotiate unless it is able to threaten its negotiating partners with arrest or worse."

ACCEPTED MANUSCRIPT

Reference Summary

Cambodian prime minister Hun Sen rejects demands of 2 opposition parties for talks in Beijing after failing to win a 2/3 majority in recent elections. Sihanouk refuses to host talks in Beijing. Opposition parties ask the Asian Development Bank to stop loans to Hun Sen's government. CCP defends Hun Sen to the US Senate. FUNCINPEC refuses to share the presidency. Hun Sen and Ranariddh eventually form a coalition at summit convened by Sihanouk. Hun Sen remains prime minister, Ranariddh is president of the national assembly, and a new senate will be formed. Opposition leader Rainsy left out. He seeks strong assurance of safety should he return to Cambodia. Cambodian prime minister Hun Sen rejects demands of 2 opposition parties for talks in Beijing after failing to win a 2/3 majority in recent elections. Sihanouk refuses to host talks in Beijing. Opposition parties ask the Asian Development Bank to stop loans to Hun Sen's government. CCP defends Hun Sen to the US Senate. FUNCINPEC refuses to share the presidency. Hun Sen and Ranariddh eventually form a coalition at summit convened by Sihanouk. Hun Sen remains prime minister, Ranariddh is president of the national assembly, and a new senate will be formed. Opposition leader Rainsy left out. He seeks strong assurance of safety should he return to Cambodia. Cambodia King Norodom Sihanouk praised formation of a coalition of the Countries top two political parties, leaving strongman Hun Sen as Prime Minister and opposition leader Prince Norodom Ranariddh president of the National Assembly. The announcement comes after months of bitter argument following the failure of any party to attain the required quota to form a government. Opposition leader Sam Rainey was seeking assurances that he and his party members would not be arrested if they return to Cambodia. Rainey had been accused by Hun Sen of being behind an assassination attempt against him during massive street demonstrations in September. Cambodian elections, fraudulent according to opposition parties, gave the CPP of Hun Sen a scant majority but not enough to form its own government. Opposition leaders fearing arrest, or worse, fled and asked for talks outside the country. Han Sen refused. The UN found evidence of rights violations by Hun Sen prompting the US House to call for an investigation. The three-month governmental deadlock ended with Han Sen and his chief rival, Prince Norodom Ranariddh sharing power. Han Sen guaranteed safe return to Cambodia for all opponents but his strongest critic, Sam Rainsy, remained wary. Chief of State King Norodom Sihanouk praised the agreement.