

Event-based summarization of news articles

Feride Savaroğlu TABAK*^{ORCID}, Vesile EVRİM^{ORCID}

Department of Computer Engineering, Faculty of Engineering, European University of Lefke,
Lefke, Northern Cyprus

Received: 12.04.2019

Accepted/Published Online: 09.01.2020

Final Version: 28.03.2020

Abstract: In recent years, with the increase of available digital information on the Web, the time needed to find relevant information is also increased. Therefore, to reduce the time spent on searching, research on automatic text summarization has gained importance. The proposed summarization process is based on event extraction methods and is called an event-based extractive single-document summarization. In this method, the important features of event extraction and summarization methods are analyzed and combined together to extract the summaries from single-source news documents. Among the tested features, six features are found to be the most effective in constructing good summaries. The constructed summaries are tested on benchmark Document Understanding Conferences 2001 and 2002 datasets, and the results outperformed most of the other well-known summarization methods.

Key words: Event-based summarization, event extraction, feature selection, single-text summarization, extractive summarization

1. Introduction

Advances in information technology enable us to share enormous amount of data on the web and its applications. Therefore, managing and using vast amounts of data for the needs of users becomes an important research area. Automatic document summarization is used to extract the main idea of the document by eliminating the less significant and redundant parts of the information. Depending on the number of documents it processes, automatic summarization is categorized as single or multidocument summarization. In 2000, the National Institute of Standards and Technology (NIST) introduced the Document Understanding Conferences (DUCs) [1] to encourage researchers to work on text summarization. However, in 2002, the DUC stopped supporting research on single-document summarization, considering that it is harder to extract summaries. Since then, most of the research has focused on multidocument summarization.

Although today there is a need for accessing lots of information in a short time, regardless of its presentation in multiple sources, this research uses single-document summarization. One of the best application areas of the summarization is information retrieval. In retrieving information, search engines (Google, Yahoo, etc.) use automatic summarization to help users find the most relevant information for their searches. Besides search engines, summarization can be used for social media marketing, automated content extraction for writing a good blog, tailoring financial documents for decision making, gathering important information for political moves, and detecting the emotions of users, among others.

Today online news sources provide instant information about the events around us, which feeds the articulations in social media channels such as Facebook and Twitter. Although summarization of news documents is

*Correspondence: ftabak@eul.edu.tr

studied intensively, few works have been conducted on event-based news summarization. As the summarization techniques aim to extract the main ideas of documents, analyzing an event and its context elements from news articles is not far from finding the most important information conveyed in the text. Therefore, in this research, an event-based summarization technique is used to extract the most meaningful information from long articles to ease the work of readers.

The rest of the paper is arranged as follows: Section 2 presents related work and Section 3 introduces the proposed method. The experimental setup and results of event-based extractive single-document summarization (EBDS) are explained in Section 4. Lastly, Section 5 is the conclusion.

2. Related work

In this paper, two related research tasks, event extraction and summarization, are combined together in the news domain to extract important parts of the text. Therefore, the research in both fields is investigated to be merged.

2.1. Summarization

The main goal of automatic text summarization is to construct a brief and coherent summary of one or more documents. The process of automatic summarization can be categorized in different ways: based on structure (extractive/abstractive) [2–5], the number of source documents, the use of external resources (knowledge-poor techniques, knowledge-rich techniques) [6], features such as query words [7], and chronological news.

Summarization can also be processed by supervised and unsupervised techniques. Supervised techniques [8,9] use a classifier trained on a set of documents. On the other hand, unsupervised techniques eliminate the need for training data for identifying the prominent sentences [10]. Graph-based methods are unsupervised extractive methods. There are many summarization algorithms that use graph methodology in their processes. A graph is a mathematical structure consisting of a set of vertices connected by edges [7]. The most frequently used graph-based ranking algorithms are Kleinberg's HITS [11] and Google's PageRank [12] algorithms. HITS and PageRank both operate on a set of pages connected by hyperlinks. The TextRank [10] and LexRank [13,14] algorithms are both graph-based algorithms derived from PageRank for ranking sentences and are frequently used in summarization. In both algorithms, vertices represent sentences and edges show the semantic similarity between the sentences. The two methods differ in similarity algorithms, such that the LexRank algorithm uses cosine similarity of TF-IDF vectors while TextRank uses varying similarity measures. Garcia et al. [15] also proposed a graph-based methodology to estimate the coherence of the sentences. Their method uses integer linear programming (ILP) for optimizing the selection of sentences based on maximizing the relevant concepts while minimizing the size of the summary.

Dimension reduction methods are also used to eliminate the least important features to gather the important sentences for summaries. Latent semantic analysis (LSA) [16] was introduced to use semantic relations among the words and sentences in the text. By the usage of singular value decomposition (SVD) in LSA, orthogonal dimensions ensure the sentence by eliminating the redundancy, which is important for multidocument summarization. Gong and Liu [16] applied SVD together with LSA for an automatic summary generation. Steinberger and Karel [17] combined the SVD method of Steinberger and Karel [18] with a sentence compression algorithm to remove unimportant parts of sentences. In order to capture the latent topics in the documents, Hidayta et al. [19] used the latent Dirichlet allocation (LDA) [20] statistical model and Lee et al. [21] used nonnegative matrix factorization (NMF) [21,22] to produce additive parts-based representation of data for interpretation of semantic features.

Mendoza et al. [23] introduced a memetic optimization model based on weighting of statistical features of each sentence (position, sentence length, sentence to title relationship, cohesion, and coverage) combined with group features (candidate sentence to summary/original document similarity). Their experiments showed that position, title to sentence relationship, and length features have the best influence to get improved results. Long et al. [24] investigated the relation of the sentences in a document and used entailment among the sentences in extracting the important sentences from the text. Their method was based on the assumption that the words in a sentence can be aligned with the other words in the other sentences. Once the sentences are scored, greedy or ILP approaches are used to select the sentences for summaries. Through the tests they evaluated combined scoring methods and the best results were obtained by the model that combined the entailment scores normalized by logarithmic functions and position scores.

More recent studies on summarization use hybrid methods including machine learning techniques. Al-guliyev et al. [25] proposed a method named COSUM, which uses two stages for sentence selection. In the first stage sentences are clustered based on k-means. In the second stage, salient sentences are selected from the clusters based on a proposed optimization algorithm that focuses on the coverage and diversity of the selected sentences. Mao et al. [26] combined graph-based techniques along with supervised learning and unsupervised learning to produce extractive single-document summaries. In this method, the importances of sentences are measured by exploring statistical features and the relationships between sentences obtained by using graph methods. Mao proposed three techniques to score the sentences: a graph-based model and supervised model, a graph-based model and independent feature of the supervised model, and sentence importance with supervised model. Liu et al. [27] proposed a learning-based algorithm called the structured summarization model (SUMO). SUMO uses iterative refinement to induce the multiroot tree into summaries by repeatedly refining the results predicted in the previous iterations.

2.2. Event detection

Starting from the early 1990s with the widespread usage of the Internet, research on event extraction and representation gained acceleration. In 2005, the Automatic Content Extraction (ACE) program annotated events from broadcast transcripts and newswire [28] for specific domains with intense annotation work. Consequently, Li et al. [29] used a cooccurrence graph between named entities and event terms. They assigned initial relevance scores to each of the named entities and event terms and ran the PageRank algorithm to define the context-dependent relevance of named entities and event terms. Kolya [30] used semantic role-labeling (SRL) [31], WordNet [32], and handcrafted rules with machine learning techniques to extract deverbal nouns as events from the TempEval 2010 dataset.

Similarly, event detection has been studied in information retrieval to retrieve event-based documents as a result of query searches [33]. Lin et al. [34] used semantic roles such as predicate-argument in queries with the ones in the main document and proved that semantic roles perform better than syntactic dependencies. Pai et al. [35] used predicate-subject-object triplets to construct content maps and find the similarity among the triplets as a representation of the event score of the documents.

Recent research on detection of events also gains importance in behavior analysis of people [36]. Extracting the reaction of people in regards to events and the effect of the conveyed information on people's emotions increases the need for combining event extraction methodologies with summarization methodologies.

2.3. Event-based summarization

While summarization and event detection research was ongoing in different fields, at the beginning of 2000, researchers start to use event information in summarization. Filatova and Hatzivassiloglou [37] defined events

as verbs/action nouns and named entities. They obtained the weight of events in sentences to identify the important sentences for construction of the summaries from multidocument text.

Filotova and Hatzivassiloglou [37] and Wu [38] showed that relevance between similar event terms improves the performance of the summarization. Xu et al. [39] derived event relevance from event ontology constructed with formal concept analysis (FCA) [29] from multidocument text. The ontology is built from a set of relevant documents and according to the named entities associated with the events. They showed that the relevant named entity recognition improves event-based summarization.

Liu et al. [40] worked on multidocument event-based summarization. They used semantic relevance obtained from the VerbOcean external linguistic resource together with the PageRank algorithm to evaluate the significance of sentences, in which the sentences with more event terms are assigned higher significance. Aksoy et al. [41] used semantic features like agent, patient, and instrument features obtained from SRL [42] on the DUC 2004 dataset. They observed that the SRL-based sentence scoring approach outperformed a term-to-term-based sentence scoring system. By considering the semantics of events, Glavas and Snajder [43] used event graphs to obtain summaries from multiple texts. In their model, sentences are selected based on event importance, which is constituted by participant importance, event informativeness, and temporal relations among events.

Many of the above event-based summarization techniques are used in multidocument summarization and also require different methodologies to eliminate the redundancy in summarization. Through the summarization methods, it is hard to identify event-based summarization of single documents. Therefore, despite event identification not being mentioned in their processes, because of the relatedness of the used techniques, some summarization methods are found to be related to event-based methods. One of the related method was proposed by Ferreira et al. [14]. In the proposed system 15 sentence-scoring methods are combined within three categories (word-based, sentence-based, graph-based) and tested on news articles, blogs, and scientific documents. In the experiments some of the combinations of these methods are also tested. As a result, it is observed that features such as TS-ISF, sentence position, lexical similarity, and sentence-title resemblance achieved the best performance for the news articles in single-document summarization. Again Oliveira et al. [44] presented the comparative analysis of 18 generic sentence-scoring techniques for single and multiple documents as an extension to Ferreira et al.'s work [14]. In their experiments each individual technique and different combinations of these techniques are tested on CNN, DUC 2001, and DUC 2002 news datasets. Although their combined techniques had a better performance, none of the combinations performed the best for all datasets.

Through the literature, it is easy to observe that news documents are the main evaluation ground for summarization methods. We believe that the main idea of news articles is highly dependent on events and articulated in event-intense sentences. However, the research on detecting events in news documents for summarization is still not well investigated. Therefore, in this project, summarization techniques are combined with various event-extraction methodologies. The event features that are most prominent in determining the main idea are identified to construct the most effective context elements of events to be used in obtaining the relevant summaries from single-document articles.

3. Proposed method

News sites such as CNN,¹ Rappler,² and BBC³ inform users about current events. An event can contain several events or other dependent events, which indicates a semantic relationship among them [45]. Therefore,

¹CNN - Breaking News, Latest News and Videos [online]. Website <https://www.cnn.com>. [Accessed 4 January 2019].

²News - Rappler [online]. Website <https://www.rappler.com>. [Accessed 4 January 2019].

³Home - BBC News [online]. Website <https://www.bbc.co.uk>. [Accessed 4 January 2019].

in this research the relationships between events and event semantics are investigated to extract the meaningful summaries as presented in the EBDS framework in the Figure.

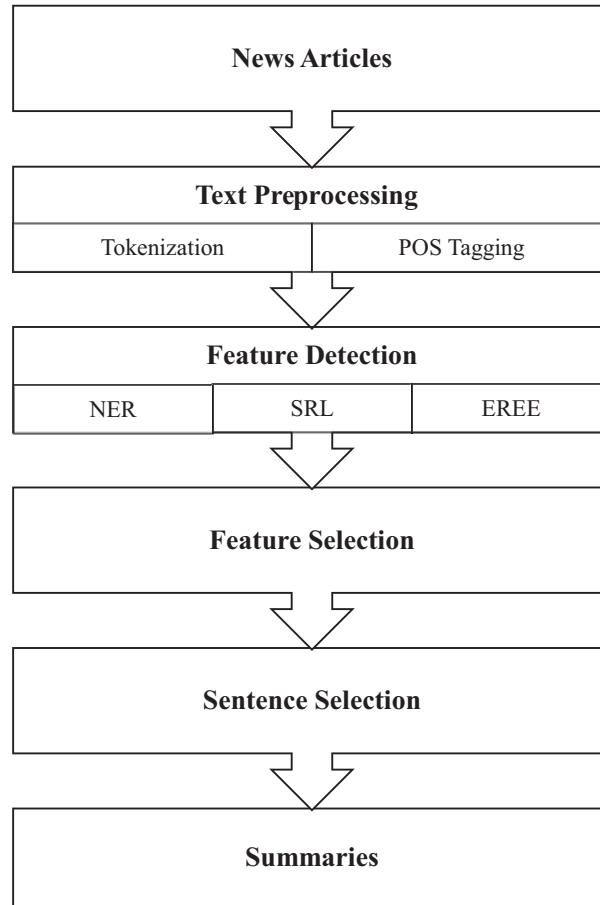


Figure 1. Event-based extractive single-document summarization process.

The first step of the framework is to preprocess the text through tokenization and part-of-speech (POS) [46] tagging to be used in further modules: feature detection, feature selection, and sentence selection of the system.

3.1. Feature detection

The preprocessed data are used to find the different features in the area of event extraction and summarization to be tested and used in the proposed summarization method.

The named entity recognition (NER) system is a well-known technique used for extracting meaningful pieces from text that are concerned with events. Named entities are also known as “atomic elements” in a sentence. In our study, 18 entity types of NER [47] such as person, organization, location, expressions of time, and money are extracted from sentences to constitute the named entity features by using the NLTK tool that is suitable for text processing libraries to provide tokenization, stemming, tagging, parsing, etc.

As verbs are the main entities of events, another set of features is extracted through SRL [42] to find predicate-argument structures including main verbs and participants in each input sentence. The Practical

Natural Language Processing Tools (practnlptools) are used for extracting SRL features. By using SRL, it is possible to identify 17 features [48]. Among the 17 features the most relevant features to the event information, namely agent (A0), patient (A1), and verb (V) and their adjuncts temporal (TMP) and location (LOC), are chosen to be used in this work. The difference between the location attribute obtained by NER and SRL is that NER-location detects the location names whereas SRL-location identifies the location or spatial orientation of a state or action. An example of temporal and locative is presented in Example 1:

Example 1 *John broke the window [at a school] SRL-LOC in the [Middle-East] NER-LOC.*

To extract the events, which are the nominalization of the predicate (verb), semantic roles are used. Nominalizations are commonly known as nouns derived from verbs, with suffixation [49]. Three known categories exist for nominalization. They are event, result, and agent/patient nominalizations. In this research, agent nominalizations [50] (e.g., officer) are identified by suffixes such as ‘-or’, ‘-er’, ‘-itor’, ‘-eer’, ‘-ier’, ‘-ior’, or ‘-yer’, while patient nominalizations [50] (e.g., employee) used end with ‘-ee’ or ‘-ed’.

Also, additional features such as non-deverbal and deverbal nouns are obtained via WordNet [32] semantic ontology. In the detection of events, some of the rules mentioned by Kolya et al. [30] and their variations are used as follows:

Noun and Verb Senses (NAVS): The words that are in the form of a noun (NN) POS categories are searched in WordNet and the ones with both noun and verb senses are considered as an event. For example, film, dream, etc.

After Stemming Verb Senses (ASV): After stemming, nouns are looked for in WordNet. If the sense of the obtained word is a verb, the word is also being considered as an event. For instance, landings, heading, stands, etc.

Deverbal Nouns (DEV): Noun words that end with such suffixes as ‘ing’, ‘ly’, ‘ious’, ‘tion’, ‘ion’, ‘sion’, ‘ment’, ‘ance’, ‘ence’, ‘ancy’, or ‘ness’ are considered as events [50]. For example: agreement, opening, direction.

Words After Preposition (WAPREPOS): The word tokens that come after prepositions such as ‘after’, ‘before’, ‘from’, ‘about’, ‘across’, ‘against’, ‘along’, ‘among’, ‘around’, ‘at’, ‘behind’, ‘below’, ‘beneath’, ‘beside’, ‘between’, and ‘during’ [50] are accepted as event words. For example: school, Academy, six.

Overall, through the above-mentioned methods, 11 features (NOUN, NE, NAVS, ASV, DEV, WAPREPOS, A0, A1, V, TMP, and LOC) are extracted to be candidates in the proposed event-based summarization as listed in Table 1.

3.2. Feature selection

The feature selection module is used to obtain important event features for summarization. Feature selection is an important process for data mining and pattern recognition. A good feature selection process helps to increase the performance of summarization [51].

Many well-developed statistical and mathematical methods are offered for feature selection [23,52]. As the aim of this work is to find the set of features that perform the best together, for selection of important

Table 1. Candidate feature set.

Feature Order	Abbreviation	Name
0	NOUN	Nouns
1	NE	Named Entities
2	NAVS	Noun and Verb Senses
3	ASV	After Stemming Verb Senses
4	DEV	Deverbal Nouns
5	WAPREPOS	Words After Prepositions
6	A0	Agent
7	A1	Patient
8	V	Verb
9	TMP	Temporal Marker
10	LOC	SRL-Location

features, genetic approaches are proposed. In genetic approaches, each solution is represented as a string of binary numbers known as a chromosome. In this paper, since the initial set of features consists of 11 candidates, the chromosome is interconnected with 11 genes. Each gene refers to a feature represented in binary format, in which zero (0) refers to an inactive feature and one (1) an active feature (e.g., 10100101011). Since the aim is to extract the best-performing subset of 11 features, 2^{11} combinations of chromosomes are tested to determine the best feature set.

Lin et al. [53] experimented on the DUC 2001 and DUC 2002 single-document datasets to find the Pearson correlation [54] between the ROUGE [53] performance evaluation measures and human summaries. The correlations in the DUC 2001 dataset were obtained as 0.78 for ROUGE-1 and 0.86 for ROUGE-2, and were outperformed by the correlation results of the DUC 2002 dataset, which were obtained as 0.98 for ROUGE-1 and 0.99 for ROUGE-2. Therefore, to get more accurate results in testing the performances of the chromosomes with 11 event-related features, the ROUGE-2 evaluation metric is used with the DUC 2002 dataset.

In this method, for each chromosome, the set of terms corresponding to the active feature of the chromosome is selected as a representative of each sentence, which then constitutes the summary of a text. Later the summary constructed for each chromosome is compared with DUC 2002 reference summaries by using the ROUGE-2 metric; Table 2 shows the ordered list of chromosomes based on their F-measures that are most effective to obtain relevant summaries.

The “SequenceofFeatures” column in Table 2 represents the chromosomes with active features of sequence such as {9, 10}, which corresponds to the chromosome (0000000011). Again the “Position” column represents the order of the sequences in terms of F-measure values. Note that the selected sequences in the top 10 do not have the first 10 position order. The selection of the sequences is based on position and also its identifying features. For example, although the sequence {4, 7, 9, 10} has position 6, as a result of the evaluation, it is not selected in the top 10 sequence list. The reason is that the sequence {4, 7, 9, 10} already contains the sequence {9, 10}, which has a position 1, and inclusion of features 4, 7 does not add any improvement in F-measure results.

Again in Table 2, since only a few feature terms are selected to represent a sentence, F-measure values are very low. However, we would like to note that the presented F-measure results are not used to judge the success

Table 2. Ordered list of selected sequence of features for summarization.

SequenceofFeatures	Position	F-measure
{9,10}	1	0.0931
{9}	18	0.0794
{10}	250	0.0581
{1,7}	996	0.0409
{1,6}	1285	0.0343
{1}	1321	0.0335
{6,7}	1493	0.0290
{7}	1596	0.0261
{5,6}	1761	0.0214

of summarization but rather to evaluate the effectiveness of the features, which is used for feature selection. Therefore, from the top sequences, it is observed that summaries with various sequences of 6 features, namely TMP (9), LOC (10), NER (1), A0 (6), A1 (7), and WAPREPOS (5), are effective in extracting event-based summaries.

3.3. Sentence selection

In recent summarization tasks, the sentence position is widely used [55,56]. It has been argued that the first sentence of the document contains the most important information about the document that is crucial for solid summaries [57,58]. Therefore, in the constructed extract, the first sentence of the text is added to the summaries regardless of the other features.

Apart from the first sentence, the other sentences in the document are evaluated based on the sequence of features as presented in Table 2. From the top, each sequence of features is checked in a sentence as presented in Algorithm 1, and once the feature sequence is detected, the rest of the sequences are neglected. Therefore, the found features from a sequence are used to calculate the importance score of a sentence as follows:

$$SentenceImportanceScore = \sum_{i=0}^{n-1} F(i), \quad (1)$$

where $i \in \{0, 1, 2, \dots, n-1\}$, and n is a number of features in a sequences for a sentence. $F(i)$ is a number of occurrences of the i th feature in a sentence. Once the sentences' importance values are calculated, the sentences with top scores are selected to construct the 100-word summaries of a document. To prevent the construction of longer summaries, in case the added sentence exceeds 100 words it is eliminated from the summary.

4. Experiments

In this section, the performance of the proposed method is compared with different summarization approaches. The evaluation is done based on the DUC [1] datasets.

Algorithm 1 Sentence selection

Input:
A[i]: Sentence array of each sequence of feature from 0 to 9
B[i]: Frequency array of sequence of feature from 0 to 9
S[i]: Sorting array that sorts sentence and frequency array according to descending order

Output: Summary

Initialization:
 $s \leftarrow 0$; $f \leftarrow 1$; $ct \leftarrow 1$

for k in range (2, len(filename)) **do**
 if $s \leftarrow no[k]$ **then**
 first.append(senten[f])
 countsent \leftarrow countsent + 1
 ct \leftarrow ct + 1
 if $temp[k] \geq 1$ and $loc[k] \geq 1$ **then**
 A[i].append(senten[k])
 B[i].append(tmp[k]+loc[k])
 else if $tmp[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(tmp[k])
 else if $loc[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(loc[k])
 else if $ne[k] \geq 1$ and $a1[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(ne[k]+a1[k])
 else if $ne[k] \geq 1$ and $a0[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(ne[k]+a0[k])
 else if $ne[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(ne[k])
 else if $a0[k] \geq 1$ and $a1[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(a0[k]+a1[k])
 else if $a1[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(a1[k])
 else if $waprepos[k] \geq 1$ and $a0[k] \geq 1$ **then**
 A[i].append (senten[k])
 B[i].append(waprepos[k]+a0[k])
 end if
 else
 f \leftarrow f + ct
 ct \leftarrow 1
 f \leftarrow f + ct
 ct \leftarrow 1
 end if
 for i in range (0,8) **do**
 S[i] \leftarrow sorted(B[i], A[i])
 allsent \leftarrow first + S[i]
 end for
 for j in range (0,len(allsent)) **do**
 if $len(sentall) \leftarrow 100$ **then**
 sentall.append(all[j])
 end if
 end for
end for

Table 3. Comparison of Rouge-1 and Rouge-2 performance evaluation results of summarization methods on DUC 2001 and DUC 2002.

METHODS	DUC 2001		DUC 2002	
	R-1	R-2	R-1	R-2
EBDS (proposed)	0.451	0.202	0.481	0.232
COSUM [25]	0.472	0.201	0.490	0.230
WA-CombA [44]	0.449	0.201	0.477	0.223
MA-SingleDocSum [24]	0.446	0.201	0.481	0.226
RI-full [61]	0.458	0.199	0.486	0.225
Linear-NN-T [26]	0.446	0.196	0.472	0.217
Sup-NN-T [26]	0.444	0.195	0.472	0.217
Entail-BSM [24]	0.439	0.190	0.469	0.217
STEIN [17]	0.438	0.183	0.475	0.218
LSA [16]	0.435	0.180	0.477	0.216
Coh-SingleDocSum [55]	0.450	0.179	0.473	0.209
OZ [60]	0.424	0.171	0.462	0.206
Baseline	0.433	0.169	0.456	0.193
NMF [22]	0.425	0.162	0.458	0.194
TexRank [10]	0.434	0.157	0.462	0.194
LexRank [13]	0.408	0.151	0.438	0.179
LDA [20]	0.388	0.130	0.432	0.172
HP-UFPE [14]	0.357	0.118	0.456	0.205

4.1. Datasets

Since news articles explain daily events and happenings, the proposed method is tested on the benchmark single-document summarizations datasets of the DUCs: DUC 2001⁴ and DUC 2002⁵. The DUC 2001 dataset contains 30 clusters of news articles collected from the Text Retrieval Conference (TREC-9) with a total of 309 documents. The dataset also includes 100-word summaries of 3 evaluators (926 model summaries). On the other hand, the DUC 2002 dataset consists of 567 news articles derived from 59 news clusters. Out of 567 news articles, since 34 articles of the dataset are repeated, 533 articles are used in the single-document summarization. The dataset at the same time has 1047 model summaries that consist of 100 words.

4.2. Evaluation metrics

Over the years, different evaluation metrics [53,59] have been proposed to evaluate the quality of the generated summaries. Among them, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [53] is found to be the most correlated to human evaluation [34]. ROUGE includes several measures such as ROUGE-N, ROUGE-L, and ROUGE-SU. Liu [34] evaluated the Pearson correlation between the 17 ROUGE measures versus human judgments on DUC 2001 and DUC 2002 datasets. Among the evaluated 17 ROUGE measures, with original summaries (not stemmed/stopped) the ROUGE-2 measure shared the best performances. Therefore, in the experiments we have used ROUGE-N, which is defined as:

$$ROUGE(N) = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} count(gram_n)}, \quad (2)$$

⁴DUC 2001- Document understanding conference 2001 [online]. Website <http://www.nlp.ir.nist.gov/projects/duc/> [Accessed 22 September 2018].

⁵DUC 2002- Document understanding conference 2002 [online]. Website <http://www.nlp.ir.nist.gov/projects/duc/> [Accessed 22 September 2018].

where N is the length of an N -gram, $count_{match}(gram_n)$ refers to the maximum length of n -grams appearing in both a candidate summary and reference summaries, and $count(gram_n)$ in the denominator Equation (2) refers to the n -grams appearing in the reference summary. Based on the similarity to human judgment and the popularity, the performance of the proposed method in this paper is compared with ROUGE-1 and ROUGE-2 evaluation metrics. The system evaluation is done between the generated summaries and the total number of model summaries of each dataset by using the ROUGE 2.0 Java Package.

4.3. Performance evaluation

The proposed EBDS system's performance is compared with 17 well-known or recent single-document summarization methods that are explained in Section 2. Comparison of the proposed method is made against the following methods: (a) well known graph-based methods of TextRank [10], LexRank [13], and Garcia's [15] coherence scoring method (Coh-SingleDocSum⁶); (b) state of the art dimension reduction methods including LSA [16], LDA [20], NMF [22], STEIN [17], and OZ [60]; (c) the best results of Long et al. [24] (Entail-BSM⁶); (d) Oliveira et al.'s [44] best-performing combination "weighted avg. CombA" (WA-CombArefno1); (e) Alguliyev et al.'s [25] proposed method (COSUM) and Mendoza et al.'s [23] proposed method (MA-SingleDocSum); (f) Long et al.'s [24] best-performing entailment-based scoring method (Entail-BSM⁶); (g) Mao et al.'s [26] best-performing combination method (Sup-NN-T and Lin-NN-T); (h) reinforcement learning with full features (RL-Full) [61]; (i) Ferreira et al.'s [14] best sentence-scoring combination for news articles (HP-UFPE FS); and (j) the baseline, which is the first n sentences that do not exceed 100 words.

Table 3 present the results obtained by EBDS and other methods in terms of the recall values of ROUGE-1 (R-1) and ROUGE-2 (R-2) metrics on the DUC 2001 and DUC 2002 datasets. In addition to the proposed method, the benchmark algorithms TextRank and LexRank, dimension reduction methods, and baseline are reimplemented in this study. The provided results of the recent methods are used for comparison.

As the results demonstrate, the proposed EBDS system outperformed the other methods in the R-2 measure while falling behind COSUM and RL-Full in the R-1 measure on both datasets. The top 5 results of the experiments are shared by the recent methods, namely COSUM, EBDS, RL-Full, MA-SingleDocSum, and WA-CombA. The success of the top algorithms differs with small R-1 value ranges of [0.446–0.472] and [0.477–0.490] in the DUC 2001 and DUC 2002 datasets. Even smaller value ranges of [0.199–0.202] and [0.223–0.232] are observed for R-2 in the DUC 2001 and DUC 2002 datasets, proving the competitiveness of the EBDS and recent methods in obtaining the summary sentences. In the tests, summaries obtained by dimension reduction methods fall behind many of the machine learning and recent methods while performing better than the graph-based methods. As a related method to EBDS, WA-CombA is able to get into the list of the top best-performing algorithms while the other related method, HP-UFPE FS, did not achieve comparable results.

5. Conclusion

In this study, a summarization method that takes into account the tight relationship between news and events is used. Although most of the datasets used in summarization include news articles, the relation between the main idea of the news and the event context has not been investigated in much detail. By the introduced EBDS framework the most effective event extraction and summarization features are combined together to extract summaries from single documents. Among the various event features, six features, TMP, LOC, NER,

⁶Named assigned in this paper to ease the comparison.

A0, A1, and WAPREPOS, are found to be effective for producing coherent summaries. In particular, sentences containing TMP, LOC, A0, and A1 features together with the first sentence have been proven to be the most effective in representing the relevant summaries.

The proposed EBDS method is evaluated by R-1 and R-2 performance measures on the DUC 2001 and DUC 2002 datasets. In the experiments, the presented EBDS method outperformed all the compared methods in the R-2 measure and obtained comparable results in the R-1 measure by achieving the third highest score in the ranking. The analysis of the results indicates that the features relating to the context of events has a close relation to the main idea in the text, which can further be analyzed in more detail to better improve the quality of summaries.

References

- [1] Nenkova A. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In: Proceedings AAAI; Pittsburgh, PA, USA; 2005. pp. 1436-1441.
- [2] Salim N, Khanam A. A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology* 2014; 59 (1): 64-72.
- [3] Liu F, Flanigan J, Thomson S, Sadeh N, Smith NA. Toward abstractive summarization using semantic representations. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Denver, CO, USA; 2005. pp. 1077–1086.
- [4] Eidheim HL. Temporal summarization of time critical events: A system for summarizing events over time in a continuous stream of documents. MSc, Norwegian University of Science and Technology, Trondheim, Norway, 2015.
- [5] Narayan S, Cohen SB, Lapata M. Ranking sentences for extractive summarization with reinforcement learning. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; New Orleans, LA, USA; 2018. pp. 1747–1759.
- [6] Chasin R, Rumshisky A, Uzuner O, Szolovits P. Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge poor unsupervised methods. *Journal of American Medical Informatics Association* 2014; 21 (5): 842–849.
- [7] Sizov G. Extraction-based automatic summarization. MSc, Norwegian University of Science and Technology, Trondheim, Norway, 2010.
- [8] Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 2016; 47 (1): 1–66.
- [9] Nenkova A, McKeown K. Automatic summarization. *Foundations and Trends in Information Retrieval* 2011; 5 (3): 103-233. doi: 10.1561/15000000015.
- [10] Mihalcea R, Tarau P. TextRank: Bringing order into texts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; Barcelona, Spain; 2004. pp. 8–15.
- [11] Kleinberg J. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 1999; 46 (5): 604-632.
- [12] Page L, Brin S. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 1998; 30 (1): 107-117.
- [13] Erkan G, Radev DR. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 2004; 22: 457–479.
- [14] Ferreira R, Freitas R, Cabral LS, Lins RD, Lima R et al. A context based text summarization system. In: IEEE 11th IAPR International Workshop on Document Analysis Systems; Tours, France; 2014. pp. 66-70.

- [15] Garcia CRC, Lima R, Espinasse B, Oliveira H. Towards coherent single-document summarization: an integer linear programming-based approach. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing; New York, NY, USA; 2018. pp. 712-719.
- [16] Gong Y, Liu X. Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; New Orleans, LA, USA; 2001. pp. 19-25.
- [17] Steinberger J, Karel J. Sentence compression for the LSA-based summarizer. In: Proceedings of the 7th International Conference on Information Systems Implementation and Modelling; Prerov, Czech Republic; 2006. pp. 141-148.
- [18] Steinberger J, Karel J. Text summarization and singular value decomposition. *Lecture Notes in Computer Science* 2004; 2457: 245-254.
- [19] Hidayata EH, Firdausillah F, Hastuti K, Dewi IN, Azhari A. Automatic text summarization using latent Dirichlet allocation (lda) for document clustering. *International Journal of Advances in Intelligent Informatics* 2015; 1 (3): 132-139.
- [20] Blei DM, Ng AY, Jordan M. Latent Dirichlet allocation. *Journal of Machine Learning Research* 2003; 3: 993-1022.
- [21] Lee JH, Park S, Ahn CM, Kim D. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing and Management* 2009; 45 (1): 20-34.
- [22] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999; 401 (6755): 788.
- [23] Mendoza M, Bonilla S, Noguera C, Cobos C, León E. Extractive single-document summarization based on genetic operators and guided local search. *Expert Systems with Applications* 2014; 41 (9): 4158-4169.
- [24] Long DH, Nguyen MT, Bach NX, Nguyen LM, Phuong TM. An entailment-based scoring method for content selection in document summarization. In: Proceedings of the Ninth International Symposium on Information and Communication Technology; Danang City, Vietnam; 2018. pp. 122-129.
- [25] Alguliyev RM, Aliguliyev RM, Isazade NR, Abdi A, Idris N. COSUM: Text summarization based on clustering and optimization. *Expert Systems* 2019; 36 (1): e12340.
- [26] Mao X, Yang H, Huang S, Liu Y, Li R. Extractive summarization using supervised and unsupervised learning. *Expert Systems with Applications* 2019; 133: 173-181.
- [27] Liu Y, Titov I, Lapata M. Single document summarization as tree induction. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Minneapolis, MN, USA; 2019. pp. 1745-1755.
- [28] Ramaswamy G, Navratil J, Chaudhari U, Zilca R, Pelecanos J. The IBM systems for the NIST 2003, speaker recognition evaluation. In: NIST 2003 Speaker Recognition Workshop; College Park, MD, USA; 2003. pp. 1-20.
- [29] Li SJ, Lu Q, Li Q. Experiments of ontology construction with formal concept analysis. In: Proceedings of IJCNLP 05 Workshop on Ontologies and Lexical Resources; Jeju Island, South Korea; 2005. pp. 67-75.
- [30] Kolya AK, Ekbal A, Bandyopadhyay S. A hybrid approach for event extraction. *Polibitis* 2012; 46: 55-59.
- [31] Punyakok V, Roth D, Yih W. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* 2008; 34: 257-287.
- [32] Miller AG, Beckwith R, Fellbaum C. Introduction to Wordnet: An on-line lexical database. *International Journal of Lexicography* 1990; 3 (4): 235-244.
- [33] Judicibus D. Method, system, and program for merging query search results. US Patent, 2003.
- [34] Lin CY, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; Edmonton, Canada; 2003. pp. 150-157.

- [35] Pai MY, Chen MY, Chu CH, Chen YM. Development of a semantic-based content mapping mechanism for information retrieval. *Expert Systems with Applications* 2013; 40 (7): 2447–2461.
- [36] Zhang S, Lv Q. Hybrid EGU-based group event participation prediction in event-based social networks. *Knowledge-Based Systems* 2018; 143: 19-29.
- [37] Filatova E, Hatzivassiloglou V. Event-based extractive summarization. In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*; Barcelona, Spain; 2004. pp. 104-111.
- [38] Wu M. Investigations on event-based summarization. In: *Proceedings of the COLING/ACL 2006 Student Research Workshop*; Sydney, Australia; 2006. pp. 37–42.
- [39] Xu W, Li W, Wu M, Yuan C. Deriving event relevance from the ontology constructed with formal concept analysis. *Lecture Notes in Computer Science* 2006; 3878: 480–489.
- [40] Liu M, Li W, Wu M, Hu H. Event-based extractive summarization using event semantic relevance from external linguistic resource. In: *Sixth International Conference on Advanced Language Processing and Web Information Technology*; Luoyang, China; 2007. pp. 117-122.
- [41] Aksoy C, Bugdayci A, Gur T, Uysal I, Can F. Semantic argument frequency-based multi-document summarization. In: *Proceedings of Computer and Information Sciences*; Northern Cyprus; 2009. pp. 460-464.
- [42] Payne T. *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge, UK: Cambridge University Press, 1997.
- [43] Glavas G, Snajder J. Event graphs for information retrieval and multi-document. *Expert Systems with Applications* 2014; 41: 6904–6916.
- [44] Oliveira H, Ferreira R, Lima R, Lins RD, Freitas F et al. Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications* 2016; 65: 68-86.
- [45] Huang D, Hu S, Min YCH. Discovering event evolution graphs based on news articles relationships. In: *Proceedings of the IEEE 11th International Conference on e-Business Engineering*; Guangzhou, China; 2014. pp. 246-251.
- [46] Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S et al. The Stanford corenlp natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*; Baltimore, MD, USA; 2014. pp. 55-60.
- [47] Goyal A, Kumar M, Gupta V. Named entity recognition: applications, approaches and challenges. *International Journal of Advance Research in Science and Engineering* 2017; 35 (5): 482-489.
- [48] Palmer M, Xue N. Calibrating features for semantic role labeling. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language*; Barcelona, Spain; 2004. pp. 88–94.
- [49] Grimshaw J. *Argument Structure*. Cambridge, MA, USA: MIT Press, 1990.
- [50] Clark D. *Comprehensive English Grammar*. Birmingham, UK: The Educational Trading Company, 1871.
- [51] Shroff KP, Maheta HH. A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy. In: *Computer Communication and Informatics (ICCCI) 2015 International Conference*; Coimbatore, India; 2015. pp. 1-6.
- [52] Liu H, Motoda H. *Computational Methods of Feature Selection*. New York, NY, USA: Chapman and Hall/CRC Press, 2007.
- [53] Lin CY. ROUGE: A Package for automatic evaluation of summaries. In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*; Barcelona, Spain; 2004. pp. 74–81.
- [54] Tomar AS. An implementation of Pearson correlation method for predicting items to user in e-commerce. *International Journal of Engineering Sciences and Research Technology* 2016; 5 (7): 873-887.
- [55] Mani KB. Text summarization using deep learning and ridge regression. 2016 arXiv preprint, arXiv: 1612.08333.
- [56] Gillick D, Favre B, Tür DH, Bohnet B, Liu Y et al. The icsi/utd summarization system at TAC 2009. In: *Proceedings of the Text Analysis Conference Workshop*; Gaithersburg, MD USA; 2009. pp. 1-20.

- [57] Suanmali L, Salim N, Binwahlan MS. SRL-GSM: A hybrid approach based on semantic role labeling and general statistic method for text summarization. *Journal of Applied Sciences* 2010; 10 (3): 166-173.
- [58] Meena YK, Gopalani D. Feature priority based sentence filtering method for extractive automatic text summarization. *Procedia Computer Science* 2015; 48: 728-734.
- [59] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*; Philadelphia, PA, USA; 2002. pp. 311-318.
- [60] Ozsoy MG, Alpaslan NF. Text summarization using latent semantic analysis. *Journal of Information Science* 2011; 37 (4): 405-417.
- [61] Henß S, Mieskes M, Gurevych I. A reinforcement learning approach for adaptive single and multi-document summarization. In: *Proceedings of International Conference of the German Society for Computational Linguistics and Language Technology*; Duisburg-Essen, Germany; 2015. pp. 3-12.