Expert Systems With Applications 134 (2019) 167–177

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique

Darshna Patel[a,*], Saurabh Shah[a], Hitesh Chhinkaniwala[b]

[a] C.U. Shah University-Wadhwan, Surendranagar, Gujarat, India
[b] Adani Institute of Infrastructure Engineering, Ahmedabad, Gujarat, India

ARTICLE INFO

Article history:
Received 14 July 2018
Revised 3 May 2019
Accepted 28 May 2019
Available online 29 May 2019

Keywords:
Feature extraction
Fuzzy logic
Text summarization
Sentence scoring
Cosine similarity
Rouge
Machine learning
Extrinsic evaluation

ABSTRACT

Nowadays abundant amount of information is available on Internet which makes it difficult for the users to locate desired information. Automatic methods are needed to efficiently sieve and scavenge useful information from the Internet. Text summarization is identified and accepted as one of the solutions to find desired contents from one or more documents. The objective of proposed multi-document summarization is to gain good content coverage with information diversity. The proposed statistical feature based model utilizes the fuzzy model to deal with the imprecise and uncertainty of feature weight. Redundancy removal using cosine similarity is presented as enrichment to proposed work. The proposed approach is compared with DUC (Document Understanding Conference) participant systems and other summarization systems such as TexLexAn, ItemSum, Yago Summarizer, MSSF and PatSum using ROUGE measure on dataset DUC 2004. The experimental results show that our proposed work achieves a significant performance improvement over the other summarizers.

1. Introduction

The popularity of the Internet is increasing drastically. As plentiful information is available on the Internet, it becomes a highly time consuming and tedious task to read entire text and documents and obtain the relevant information on specific topics. Text summarization is acknowledged as a solution for this issue as it creates automatic briefing of the information (Sanchez-Gomez, Vega-Rodríguez, & Pérez, 2017). Text summarization can be defined as a shortened version of text generated from one or more documents without losing main contents or idea of the original document(s) and it is no longer than half of the original text (Aliguliyev, 2009). Automatic text summarization has become a promising approach to overcome the information overload and automatically generates the summary from all the textual information of the topic (Moradi & Ghadiri, 2017). There are other disciplines too which are related to text summarization such as automatic text classification (Liu et al., 2017) and text clustering (Wei, Lu, Chang, Zhou, & Bao, 2015), information retrieval and extraction, query answering (Yulianti, Chen, Scholer, Croft, & Sanderson, 2017), sentence ordering (Bollegala, Okazaki, & Ishizuka, 2012) etc. The aim of summarization system is to produce concise and fluent summary of the given text by covering most important part of the contents and with minimum redundancy from different input sources.

There exists a variety of taxonomies for text summarization based on frequency of input sources, the way of summary generated, purpose of summary, language of input sources and genre is shown in Fig. 1. There are two kinds of algorithms exist about which various works have been published around text summarization. They are Extraction based summarization and Abstraction based summarization. Extraction based summarization is based on the extraction of sentences from text document verbatim. In this approach, there is no condensing happening in any format. It is just picking up sentences to form a shorter summary. Abstraction based summarization works differently. Apart from picking up most relevant sentences, it changes the manner in which a document is represented. It regenerates the extracted text. As per the size of input sources considered for generating summary, it can be classified as single document or multi-document summarization. When a single document is given as input for text summarization, it is called single document summarization whereas in multi-document summarization, a set of documents are given as input to generate the summary. Domain specific summarization utilizes the domain specific knowledge to generate the summary, whereas domain independent summarization (generic) applies general features to get the important segments from the text. Modern researchers have drifted to domain specific

[*] Corresponding author.
E-mail addresses: darshnabpatel@gmail.com, darshna.it@ksv.ac.in (D. Patel).

https://doi.org/10.1016/j.eswa.2019.05.045
0957-4174/© 2019 Published by Elsevier Ltd.

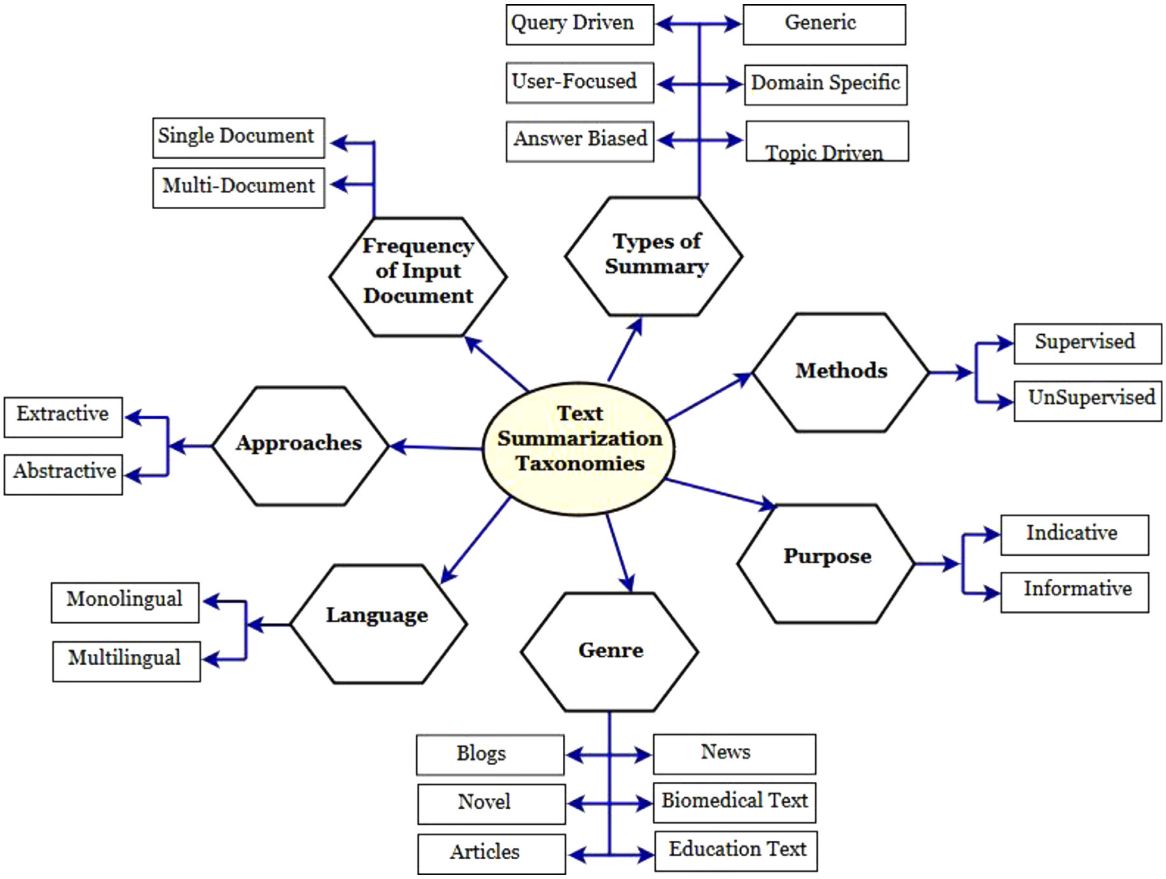

Fig. 1. Text summarization taxonomies.

summarization techniques. As per target audience, summary can be query specific (Malviya & Tiwary, 2016; Yousefi-Azar & Hamey, 2017), user focused (Liu, He, Chen, Peng, & Fang, 2013), answer biased (Yulianti et al., 2017) or topic focused. Query specific summary is dependent on audience for whom the summary is intended. It responds to the query triggered by users. User focused technique generates summaries to adapt the interest of particular users while topic specific summary gives special attention to the particular topic of document. In recent years, text summarization field has improved in terms of language processing and machine learning technologies for sentence scoring tasks (Fang, Mu, Deng, & Wu, 2017). Summarization systems can be supervised or unsupervised depending on whether the training data is required or not. A supervised system gets trained from labeled data to select crucial concepts from documents. Unsupervised systems are predominant epitome and generate summary from documents. They do not rely on any training instances labeled by humans. With regard to language of input documents, it may be monolingual or multilingual. There are many types of summaries used in everyday life such as News articles, Digest (Summary of stories on same topic), Summary of long text like Novel, book or magazine (Wu et al., 2017), highlights (Summary of some events), online debate forum (Cai & Li, 2011), opinion mining, abstract of research paper, online blogs, biomedical text, education Text, feedbacks of any product, book, music, movies, plays etc.

In this paper, we propose fuzzy logic based multi-document summarization system to extract important sentences to generate non-redundant summary. The proposed approach is an extractive based generic summarization system and summary in the context of this proposed work is textual summary produced from one or several news related documents.

The rest of this paper is organized as follows. Section 2 presents the past researches on text summarization. Section 3 explains proposed method with preprocessing and extracting important features. Sections 4 and 5 describe fuzzy logic system, followed by Cosine similarity measure used in proposed method to remove redundancy from final summary. Section 6 describes the experimental setup and result analysis and finally, there is conclusion and suggestions for future work that can be carried out in Section 7.

2. Related work

Several efforts have been made in the progression of text summarization systems with the use of different approaches, technologies and tools. The three most common approaches used in literature are the statistical approach, the linguistics approach and combination of both the approaches called hybrid approach (Tayal, Raghuwanshi, & Malik, 2017). Statistical based approach extracts most salient sentences based on the shallow features of text such as title of document, sentence position, cue words, thematic words, keywords etc. It derives the weight of keywords and calculates the significance of sentences and generates the summary. It (Ferreira, Freitas et al., 2014) evaluates the hypothesis that quality of summary can be improved by using different combination of various shallow text features depending upon the context. Linguistics based approach looks into terms semantics and identifies the relationship between key terms by Part Of Speech (POS) Tagging, thesaurus usage, grammar analysis etc. Statistical methods are more efficient in computation but linguistics based approach looks into the semantic so generates better summary (Tayal et al., 2017). Hybrid method combines both the approaches to generate the summary.

Conventional methods rely on statistics to create summaries. However some systems are developed which take advantages of both statistics and linguistics approaches. Tayal et al. (2017) presented a method of text summarization based on soft computing which takes advantage of statistical and linguistic approach. It uses title and semantic similarity to create summary. An unsupervised generic approach uses graph model based statistical similarity and also makes linguistic treatment on text by performing co-reference resolution and discourse analysis (Ferreira, De Souza Cabral et al., 2014). Emotion plays the important role in communication to convey any message effectively, based on that hybrid method for single document summarization is proposed which uses statistical features and sentiment analysis as semantic feature (Yadav & Sharan, 2015).

Most of the researches follow single objective optimization model. Nature inspired optimization approaches are suitable choice to address the optimization problem. Sanchez-Gomez et al. (2017) presented multi-objective optimization model based on Artificial Bee Colony (ABC). It simultaneously maximizes the functions of content coverage and redundancy removal of document collection while generating summary. Being inspired by other optimization model Rautray and Balabantaray (2017) presented meta-heuristics Cat Swarm Optimization (CSO) based multi document summarization. Memenic algorithm (Mendoza, Bonilla, Noguera, Cobos, & León, 2014) is proposed for extractive based single document summarization as binary optimization problem. They use individual statistical features of sentences to score the sentences. Alguliev, Aliguliyev, and Mehdiyev (2011) presented text summarization task as an optimization problem and extracted the salient sentences from set of documents. It also reduces the redundant information from summary. They have used adaptive DE (Differential Evolution) as an optimization solution. Recently some researchers have used cellular learning automata (CLA) and fuzzy logic with some evolutionary and soft computing techniques for text summarization. Abbasi-ghalehtaki, Khotanlou, and Esmaeilpour (2016) have constructed hybrid model of fuzzy logic, Cellular learning automata, Artificial Bee Colony (ABC) and PSO-GA for text summarization. To calculate the sentence similarity, artificial bee colony and cellular learning automata are used. Likewise, to assign the best weight to the extracted features, PSO and Genetic algorithm are used followed by fuzzy logic is used for final scoring.

As the extractive summarization techniques are easier to use, these techniques have received more attention in the field of text summarization. General graph based methods do not consider the significance of words in sentence scoring. Fang et al. (2017) proposed unsupervised word sentence co-ranking algorithm which combines the word-sentence relationship with graph based ranking algorithm in such a way that mutual influence can able to deliver the principal status of words and sentences more accurately. Another graph based heterogeneous ranking algorithm (Yan & Wan, 2014) is proposed which also uses the semantic role information to improve the performance of multi document summarization. Text Clustering is also heavily researched and varieties of techniques have been used in the fields of text summarization. The main idea of clustering is to group the documents which are similar in contents. Wei et al. (2015) proposed clustering based approach which also use WordNet and lexical chain to capture the main theme of the documents. Most of the previous research focuses on determining only relevance/coverage of input document and takes redundancy removal as post processing step. Different from previous researches, Chen, Liu, Chen, and Wang (2018) proposed novel density peaks clustering summarization framework which takes both relevance and redundancy removal through one pass process. Cai and Li (2011) proposed novel approach that simultaneously clusters and rank the sentences based on spectral analysis. Fattah (2014) used trainable summarizer which takes into account some statistical features. These features are then used with the combination of Maximum entropy, a naive-bayes classifier and support vector machine to construct the text summarization model. Nandhini and Balasundaram (2013) proposed supervised machine learning approach especially for learners to solve the learning difficulties. They combined the existing features with learner dependent features and use genetic algorithm to weigh the features. Text classification has also received significant attention due to increasing amount of text information. Several classification methods are used in recent years. Liu et al. (2017) proposed model to address the problem of imbalanced classes in supervised document classification. They used heuristics to optimize the classification accuracy by expanding the regions over which they cover. Bollegala et al. (2012) has focused specially on sentence ordering in multi document summarization. To capture the sentence order preference against another sentence, they have defined five preference experts such as chronological expert, probabilistic expert, topical closeness expert, precedence expert and succession expert; and used hedge regression algorithm for weighting linear combination of those experts.

There are very few research work carried out on abstractive summarization; most of them are based on syntactic and semantic based approaches. Many researches use ontologies in text summarization to recognize the semantic concepts from documents. The problem with term-based method is that it cannot work with the problem of synonymy and polysemy. Ontology based methods address this issue by using semantic information of document. Semantic based approach mostly depends on the human expert to construct domain ontology and rules. This limitation is treated in a framework (Khan, Salim, & Jaya Kumar, 2015). The researchers have used semantic role labeling (SRL) to build semantic representation from source text documents. Yago based summarizer (Baralis, Cagliero, Jabeen, Fiori, & Shah, 2013) is multi-document summarizer that integrates the ontological knowledge to generate the accurate summary. The limitation of ontology based method is construction of ontology; it requires lot of man power. To overcome this problem, a pattern based model (Qiang, Chen, Ding, Xie, & Wu, 2016) is proposed which makes use of closed patterns to extract the significant sentences from document. Another itemset-based summarizers ELSA proposed by Cagliero, Garza, Baralis, & Torino (2019) to overcome the limitation of LSA and itemset based summarizers by taking best of two both. A topic based abstractive summarization system (Chowanda, Sanyoto, Suhartono, & Setiadi, 2017) is developed to generate the summary from online debate forums. They have included the stance of statement to improve the performance of summary. Pal and Saha (2014) proposed unsupervised approach based on semantic information of input text and evaluated the importance of sentence with the help of Simplified Lesk Algorithm and WordNet. Some of the approaches use word graph for abstractive summarization. The generation algorithm finds the path among words on graph, so, many sentences with incorrect meaning can be generated. To overcome this issue, Le (2013) used rhetorical structure and word graph to generate the abstractive summary.

Since 1980s, there have been some research works on biased summarization in terms of query biased summarization, user biased summarization or topic focused summarization. Yousefi-Azar and Hamey (2017) introduced unsupervised approach using deep neural network for query based email summarization. They used Auto-Encoder (AE) for automatic feature learning process which is completely independent from human generated features. Malviya and Tiwary (2016) also developed a system which generates summarized research article based on query generated by user. The key step in topic biased summarization is acquisition of topics. Liu et al. (2013) proposed topic oriented characteristics database of word-occurrence with sample documents by analyzing user interest based on existing studies. They use this expandable semantic database to get the biased topic and guide their study of summarization. A novel approach TAOS (Fang et al., 2015) is proposed to generate topic aspect oriented multi document summary with the use of latent variable. Mei and Chen (2012) introduced novel extractive based approach SumCR for multi-document summarization. They used subtopic level 'exemplar' feature and document level 'position' feature to extract the important sentences. With the hypothesis that most of titles of scientific and technological literature reflect the topic, Liu et al. (2014) has proposed Titled-LDA model that simultaneously generate the document summary and title of document. Wu et al. (2017) proposed Novel document (long text) summarization based on LDA topic modeling algorithm. Yulianti et al. (2017) formulated idea of document summarization for answering non-factoid queries to extract the answer biased summary from retrieved CQA (Community Question Answering) documents. Most of the researches available for text summarization are designed for English language only. Cabral et al. (2014) has proposed a language independent summarizer which used different techniques for language classifier and language translation. They used Microsoft API for language translation.

In recent years, in biomedical fields, varieties of summarization approaches have been proposed. Moradi and Ghadiri (2017) uses six different feature selection approaches to identify the important concepts and classifies the sentences as summary or non-summary sentences of biomedical articles. Lloret, Romá-Ferri, and Palomar (2013) also analyses the appropriateness of two text summarization approaches in the automatic generation of abstract from specific biomedical domain. They have analyzed that since both the extractive and abstractive approaches are useful for producing summary from biomedical research paper, abstractive approach is better from a human perspective. In literature, very few researches are found on Arabic language summarization due to intrinsic complexity of language itself. Al-Abdallah and Al-Taani, (2017) used the statistical and semantic based sentence features to extract most relevant sentences from the single Arabic document. They used Particle Swarm Optimization algorithm to extract the most relevant sentences from Arabic single document. Oufaida, Nouali, and Blache (2014) introduced novel Arabic language summarization system for single and multi-documents based on sentence clustering and adapted discriminate analysis methods.

As seen in literature review, over the last fifty years, tremendous research has been performed in the field of text summarization. Various novel methods and approaches such as statistical based, linguistic based, graph based, topic based, discourse based, based on machine learning, Evolutionary & optimization algorithms and abstractive & mathematical approaches have been developed to improve the quality of summary. By analyzing the previous research through literature survey, we found that recent research approaches outperformed the past one but still there are some open issues exists such as maximum content coverage, redundancy, sentence ordering and cohesiveness in contents in the field of extractive and abstractive summarization.

With the encouragement to enhance the task of Text summarization, this research work aims to address the challenging task of Automatic Text Summarization of news documents of English language with the help of classical statistical text processing principles. The statistical techniques were found to be simple and faster in implementation. They worked efficiently with larger documents also. In all the sentence scoring based approaches like statistical, machine learning and graph based, score of sentences are based on extracted features. The weight of these features can be uncertain and undetermined. Some of the features may have more importance and some of them may have very less importance. The features are not weighted legitimately. In order to provide balanced weight for the features vector, fuzzy logic system is utilized in the proposed research work. The key characteristic of fuzzy logic is it is based on natural language and can handle approximate and uncertain data which makes it attractive for automatic text summarization. In our approach, redundancy is considered as negative factor which can affect the quality of summary. As a post-processing, redundant sentences (which are identified by cosine similarity measure) are removed from the summary to improve the quality of summary.

3. The proposed multi-document summarization

Multi-document Summarization system has always extra subtasks as compared to single document summarization such as sentence extraction from multiple document, topic detection, sentence ordering and redundancy reduction. The proposed multi-document summarization system is augmented version of our single document summarization system (Patel & Chhinkaniwala, 2018). It performs all the tasks of SDS with extra subtask that minimizes redundant contents from final summary as shown in Fig. 2. We have proposed our approach for three aspects of text summarization.

- Information richness that is main content coverage
- Information diversity that is minimum similarity within contents
- Compression ratio that is expected length of summary

3.1. Input pre-processing

Prior to input multi-documents to the proposed method, some pre-processing steps are required to the set of raw documents.

- Removing Stop words: The most commonly used words such as 'a', 'an', the' etc. do not have any semantic information with respects to the document are removed. All the stop words are predefined and placed in independent file.
- Stemming: Stemming is the process of converting each words into their root form by removing its prefix and suffix. We have used porter stemming algorithm for stemming process.
- Special character removal: All the special characters like punctuation, interrogation, exclamation etc. are removed by space character from set of input documents.
- Segmentation: It is the process of extracting each sentence separately from documents. All the sentences from documents are extracted and stored in order.
- Tokenization: After the sentence segmentation, tokenization takes place on all the sentences. It is the process of separating words from each sentence. It is used to identify the character structure such as date time, punctuation mark, number etc.

3.2. Feature extraction

In feature extraction step, the preprocessed data in word form is used to determine sentence score. The efficacy of different sentence scoring methods depends upon the kind of text, genre of text, structure and language of input text (Ferreira, De Souza Cabral et al., 2014). The core perception is different topics can prefer different aspects; different aspects can be represented by different combination of features (Fang et al., 2015). All the text features are classified as word level and sentence level features. We have performed experiments on different combination of shallow/statistical text features on different dataset and choose best combination of features which can give best result in terms of coverage and relevancy for news domain. The features used in proposed approach are explained below.

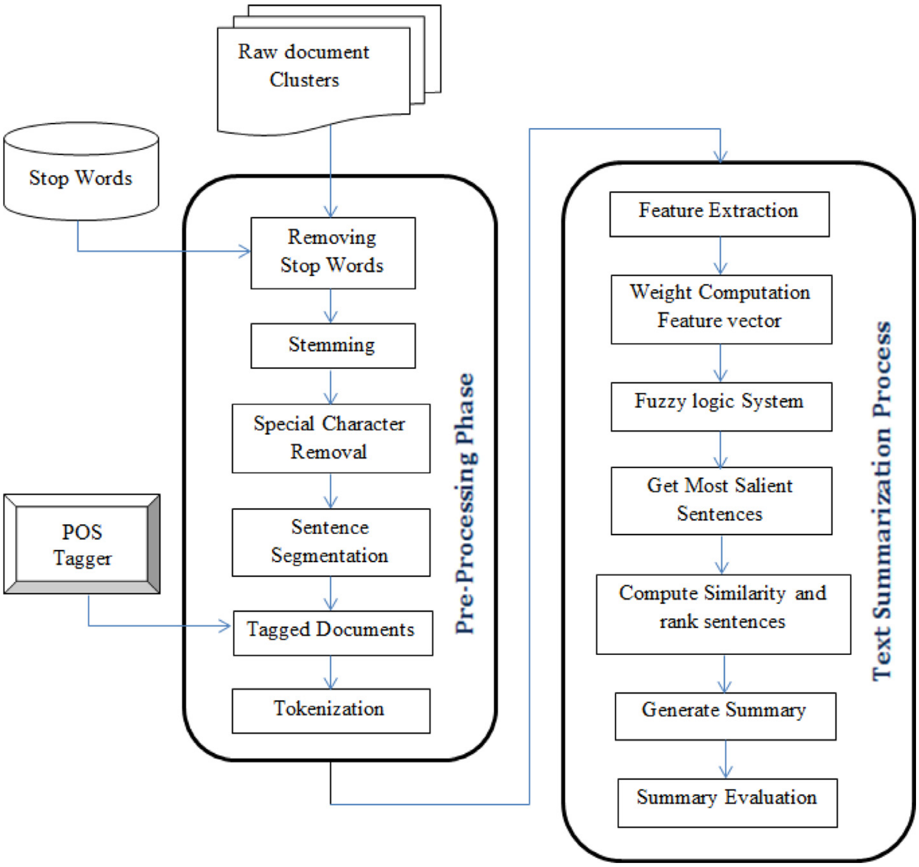

Fig. 2. Proposed multi-document summarization processing flow.

3.2.1. *Word features*

Each sentence is collection of words so individual word score is playing important roles to calculate the sentence score. Here each keyword receives individual score and total of all sentence constituent words score is sentence score. Different word features are calculated as below:

- Title-Word: The sentences comprise of words which are also found in title of document, those sentence are considered as relevant content or theme of document and considered as most significant to be included in summary. The score of title word i in sentence j is calculated per Eq. (1) (Abbasi-ghalehtaki et al., 2016).

$$\text{Title\_Word Score}(w_{ij}) = 5/10 \tag{1}$$

- Thematic Word: Thematic words are list of domain-specific most frequent keywords in the documents. A sentence having maximum number of thematic words should be selected for summary. In proposed work, top 5 frequent keywords are used for consideration as thematic word.The score of thematic word i in sentence j is calculated as per Eq. (2) (Abbasi-ghalehtaki et al., 2016).

$$\text{Thematic\_Word Score}\left(w_{ij}\right) = {}^{4}\!/_{10} \tag{2}$$

- Proper-Noun Word: Sentences containing proper-nouns or named entities such as person, organization or location is considered as an important sentence and should be included in summary.The score of proper-noun word i in sentence j is calculated as per Eq. (3) (Abbasi-ghalehtaki et al., 2016)

$$\text{Proper\_Noun\_Score}\left(w_{ij}\right) = {}^{2}\!/_{10} \tag{3}$$

- Keywords: If the word is repeated and containing only in limited sentences are important and have high probability to be included in summary. Keywords are generally nouns. It calculates the Term Frequency (TF) and Inverse Document Frequency (IDF) for every word from each document as shown in Eqs. (4) and (5).

$$\mathrm{TF-IDF}(t_i) = \mathrm{TF}(t_i) \times \mathrm{IDF}(t_i) \tag{4}$$

$$\mathrm{TF-IDF}(D_i) = \sum_{t_i \in T}^{D} \mathrm{TF} \times \mathrm{IDF} \tag{5}$$

TF is frequency of term (t_i) in document / total number of terms (T) in document

$\mathrm{IDF} = \log(\frac{D}{D_{t_i}})$ where D is the total number of documents and D_{t_i} is the total of document in which term t_i occurs.

- Numerical data: Sentences containing numerical data are considered as significant and have to be incorporated in summary. If word i in sentence j is a numerical data, then score is given as per Eq. (6) (Abbasi-ghalehtaki et al., 2016).

$$\text{Numerical \_Score }\left(w_{ij}\right) = {}^{1}\!/_{10} \tag{6}$$

3.2.2. *Sentence features*

These features investigate the features of sentence itself. We have identified two features of sentences such as sentence position and sentence length which are calculated (Abbasi-ghalehtaki et al., 2016) as below:

- Sentence Position: Location of the sentence always shows importance of sentences. Leading sentences of documents are always important and should be included in summary. The score

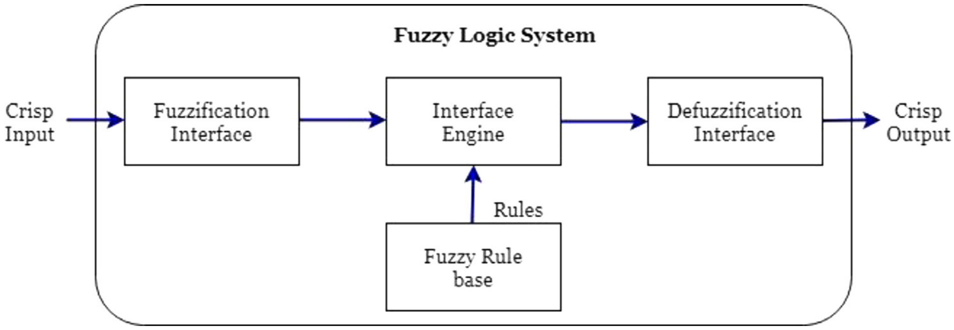

Fig. 3. General architecture of fuzzy logic system.

of sentence position feature is computed by Eq. (7)

$$\text{Position Score}(\mathrm{S_i}) = 1 - \frac{\mathrm{i}-1}{\mathrm{N}} \tag{7}$$

where S_i is i^{th} sentence in document and N is total number of sentences in document

- Sentence Length: Short sentences do not cover any crucial information so we are not considering short sentences as essential one. The score is computed as shown in Eq. (8).

$$\text{Sentence\_Length}(\mathrm{S_i}) = \frac{\text{No. of word occuring in S}}{\text{No. Word Occuring in longest Sentence}} \tag{8}$$

3.2.3. Sentence scoring

The proposed research work focuses on the extraction of salient sentences based on sentence score. In statistical and unsupervised machine learning approach, sentence score is total weight of sentence features. The feature score of each sentence that we have calculated in previous section are used to sort the salient and most important sentences. The limitation in general statistical method is weight of the features can be estimated and uncertain. We have used Fuzzy Logic System (FLS) to solve this issue. The fuzzy logic system is described Section 4.

3.3. Summary generation process

To generate the summary, all the sentences are arranged according to their highest to lowest score obtained from fuzzy logic system. Numbers of sentences are selected based on given compression ratio and similarity measure with other sentences included in summary. To calculate sentence similarity, we used cosine similarity measure which is explained in Section 5.

Following steps are followed:

Step 1: Calculate similarity index

We add next sentence in the summary if similarity is less than threshold θ.

The Parameters used are as below:

(1) Final_Summary= "" // initially Final_Summary is empty,
(2) Similarity Threshold "θ",
(3) SLen is length of summary as per given compression ratio

```
For $k=1$ to total number of sentences
if (Similarity_Compare(Final_Summary, $k^{th}$ sentence) $<\theta$) AND
    (Length (Final_Summary) <SLen)
Then
Final_Summary= Final_Summary+ $k^{th}$ sentence
```

Step 2: Arrangement of sentences

Sentences are arranged in the final summary as per their original position in source documents in such a way toretain cohesiveness. For that we have defined some rules as below:

R1: Sentences are first arranged in descending order according to their score obtained from FLS.
R2: If a particular document contains two or more similar sentences with identical scores, the former should be given priority over the later.
R3: If two sentences from different document shows same score and at same position in their respective documents then, sentence appearing in earlier document is given preference over the other sentence.

4. Fuzzy logic system

There are varieties of applications developed with the help of fuzzy logic. Fuzzy logic is flexible, easy to understand and can be considered as tolerant to imprecise data (What Is Fuzzy Logic, 2019).

The Fuzzy Logic System (FLS) consists of four main components as shown in Fig. 3

- **Fuzzifiers**: In this component, the text features are given as crisps input and converted into linguistic values by using membership function. The proposed work used Triangular Membership Function (TMF) for each feature and divided into three fuzzy set: Less, medium and more. The Triangular Curve is specified by three parameters a, b and c as shown in Eq. (9)

$$f(x,a,b,c) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a \le x \le b \\ \frac{c-x}{c-b}, & b \le x \le c \\ 0, & c \le x \end{cases} \tag{9}$$

Where a & c locate the feet of triangle and b locates the peak.

- **Fuzzy Inference Engine:** This is the main component of fuzzy logic system. It performs formulating outputs based on obtained membership functions and fuzzy rules. It takes the fuzzy input from the fuzzifiers with rule based and takes decision. Mamdani fuzzy inference system (FIS) is the most commonly used system in many applications, due to its straightforward structure of min-max operations. The Mamdani method is increasingly appropriate for text summarization system as it captures expert knowledge that enables us to depict the abilityin more insightful and more human-like way. We used MATLAB tool for implementing and editing fuzzy inference system.
- **Rule base:** The process of rule designing is an important phase in the fuzzy classification algorithm. All the rules have been built manually and for that human knowledge is used to design the if-then rules. In inference process, all the possible (almost all combination) rules are defined by three human experts; we got different number of rules from three experts and eventually finalized total 273 rules. All the Fuzzy rules are designed keeping in mind that high priority is given to features such as title word feature, sentence position, thematic words, inclusion

of numerical data and proper noun as compare to the features such as sentence length. This consideration is done based on following criteria.

- Leading sentences always contains important information.
- Sentences having title words and frequent words considered as related to topic and should be included in summary.
- In news documents numbers or figures are important and those sentences containing numerical data are considered as salient sentences.
- Short length sentences such as date, author name and very big sentences without any information are considered as not important.

Some rules are shown below:

- IF (title_word is more) and (TF-IDF IS more) and (thematic_score is more) and (pronoun is more) and (Num_data is more) and (Sentence_Len is medium)THEN (sentence is VeryImporant)
- IF (title_word is more) and (TF-IDF IS more) and (thematic_score is more) and (pronoun is average) and (Num_data is less) and (Sentence_Len is medium)THEN (sentence is Imporant)
- IF (title_word is less) and (TF-IDF IS less) and (thematic_score is less) and (pronoun is more) and (Num_data is less) and (Sentence_Len is short)THEN (sentence is unImporant)

- **Defuzzification**: The final step of fuzzy model is defuzzification of the fuzzy sets generated by the Fuzzification. Defuzzification process converts the linguistic inference results back into crisp output. The built-in method used for defuzzification is the centroid method which returns the center of area under the curve. The generalized triangular membership function is used as output membership function as given in Eq. (10).

$$c(x,y) = \left(\frac{a+b+c}{3}, \frac{l+m+n}{3}\right) \tag{10}$$

where a, b and c are the standard values of low, medium and high respectively and l, m and n are the calculated values of unimportant, average, Important and veryImportant respectively.

5. Redundancy removal

In extending the single document summarization to multi-document summarization, the first step we have chosen is minimizing redundant information from final summary because it is greater chances to get multiple sentences having similar contents from multiple documents. The significant step in removing redundant information is to identify the similar contents from summary using appropriate similarity measures. Vector based methods use statistical information on words and most commonly used in text summarization context (Aliguliyev, 2009). In this method, each sentence is represented as word vector and similarity between sentences are calculated pair wise using similarity measures. Cosine similarity is one of the most used similarity measure in text summarization (Alguliev et al., 2011; Sanchez-Gomez et al., 2017) which is explain below.

The Cosine similarity measures the resemblance between pair of sentences $\mathbf{S_i} = \{\mathbf{w_{i1}}, \mathbf{w_{i2}}, \mathbf{w_{i3}}, \ldots, \ \mathbf{w_{im}}, \}$ and $\mathbf{S_j} = \{\mathbf{w_{j1}}, \mathbf{w_{j2}}, \mathbf{w_{j3}}, \ldots, \ \mathbf{w_{jm}}, \}$ is computed using Eq. (11)

$$\mathbf{Sim}(\mathbf{S_i}, \mathbf{S_j}) = \frac{\sum_{\mathbf{k}=1}^{\mathbf{m}} \mathbf{w_{ik}} \mathbf{w_{jk}}}{\sqrt{\sum_{\mathbf{k}=1}^{\mathbf{m}} \mathbf{w_{ik}^2} \cdot \ \sum_{\mathbf{k}=1}^{\mathbf{m}} \mathbf{w_{jk}^2}}}, \tag{11}$$

$$\boldsymbol{where\, i\, and\, j} = 1, 2, 3, \ldots, \ \boldsymbol{n\, and\, n\, is\, total\, number\, of\, sentences}$$

Here, $\mathbf{w_{ik}}$ and $\mathbf{w_{jk}}$ is the weight of corresponding terms $\mathbf{t_k}$ in sentences $\mathbf{S_i}$ **and** $\mathbf{S_j}$.

In our work, the weight $\mathbf{w_{ik}}$ associated with term $\mathbf{t_k}$ in sentence $\mathbf{S_i}$ is calculated using term-frequency inverse sentence frequency (TF-ISF) schema. It is the combination of TF and ISF where TF is how many times term t appears in sentence and ISF is how many sentences of the summary collection contains the term t. The weights are calculated using Eq. (12)

$$\mathbf{w_{ik}} = \mathbf{TF_{ik}} \ . \ \log {}^{\mathbf{n}}/_{\mathbf{n_k}} \tag{12}$$

Where $\mathbf{TF_{ik}}$ counts how many times the terms $\mathbf{t_k}$ appears in the sentence $\mathbf{s_i}$ and $\mathbf{n_k}$ denotes number of sentences containing the term $\mathbf{t_k}$

In our experiment, we have applied cosine similarity measure to the high scoring sentences extracted by previous section. At beginning, we select the first sentence from high scoring sentences ranking list. Then we observe the next sentence and compare it with sentence(s) already included in summary sentences list. Those sentences which are too similar to the already included sentences (cosine similarity value is greater than threshold value 0.8) are considered as redundant sentences and are not to be included in summary. This process is continued until summary reach to the desired length. We have used python tool to calculate the cosine similarity between two sentences (scikit-learn, 2019).

6. Experimental setup and result analysis

This section illustrates the experimental setup for proposed multi-document summarization system and performance analysis. All the experiments and implementation are done in window system with Intel 2.50 GHz CPU and 4GB memory.

6.1. Dataset

The proposed framework is evaluated on DUC 2004 Dataset which is benchmark content for English written multi-document summarization. DUC 2004 contains large varieties of articles of different subjects. All the articles are clustered according to their subjects and each cluster is given as input to the proposed summarization system. DUC 2004 also provides four human generated summary as reference or gold standard summary. Participants of the contest compare their summary with gold standard summary for qualitative evaluation (Baralis et al., 2013; Qiang et al., 2016). To carry out the comparison with the DUC'04 participants, one widely used open source text summarization system TexLexAn (TexLexAn, 2019), and other summarizer systems like ItemSum, Yago Summarizer, MSSF and PatSum systems (Baralis et al., 2013), Qiang et al. (2016), we have generated 665 bytes of summary.

6.2. Experimental settings

There are various methods available to evaluate the summarization output such as pyramid evaluation, content coverage score, responsiveness evaluation, relative utility and ROUGE score (Louis and Nenkova, 2013). However, ROUGE is still most popular evaluation method for automatic summarization system and has been adopted as official tool for performance evaluation. ROUGE automatically compares the system generated summary and human generated summary by n-gram co-occurrence so remove the need for manual judgment (Louis and Nenkova, 2013). We are using rouge 2.0_0.2 java based toolkit for evaluation. We have carried out several evaluation score such as ROUGE-1, ROUGE-2, ROUGE-4, ROUGE-L and ROUGE-SU4. As previously done by Baralis et al. (2013) and Qiang et al. (2016), we have reported only ROUGE-2 and ROUGE-4 scores in paper as shown Table 1

Table 1
Comparison between proposed MDS and other approaches on DUC 2004 data.

Summarizers		ROUGE-2			ROUGE-4		
		Recall	Precision	*F*-measure	Recall	Precision	*F*-measure
Top Ranked DUC 2004 Peers	Peer 124	0.083	0.081	0.082	0.012	0.012	0.012
	Peer 65	0.092	0.091	0.091	0.015	0.015	0.015
	Peer 19	0.080	0.081	0.080	0.010	0.010	0.010
	Peer 44	0.076	0.079	0.077	0.012	0.013	0.012
	Peer 81	0.081	0.079	0.080	0.013	0.013	0.013
	Peer 104	0.086	0.084	0.085	0.011	0.010	0.011
TexLexAn		0.067	0.067	0.067	0.007	0.007	0.007
ItemSum		0.083	0.085	0.084	0.012	0.014	0.014
Baseline		0.092	0.091	0.092	0.014	0.014	0.014
Yago Summarizer		0.095	0.094	0.095	0.017	0.017	0.017
MSSF		0.098	**0.098**	0.098	0.017	0.017	0.017
PatSum		**0.102**	**0.102**	**0.102**	**0.020**	**0.020**	**0.020**
Proposed MDS		**0.155**	0.073	**0.099**	**0.068**	0.026	**0.038**

Table 2
The associated *p*-values of the paired *t*-test (95% significance level).

	ROUGE-1 Recall	ROUGE-2 Recall	ROUGE-4 Recall
Proposed Vs. SYSTEM 65	1.6e−2	2.1e−2	1e−2
Proposed Vs. SYSTEM 104	4e−3	1.1e−2	4e−3
Proposed Vs. SYSTEM 19	2e−3	3e−3	4.3e−2
Proposed Vs. SYSTEM 44	4e−3	1e−3	6e−3
Proposed Vs. SYSTEM 81	5e−3	7e−3	4e−2
Proposed Vs. SYSTEM 124	1.4e−2	1.3e−2	5e−3

[1] Null hypothesis (H0): There is no difference between the two models.
[2] Alternative hypothesis (H1): The first model outperforms the second model.

6.3. Result analysis

The proposed MDS is compared with the DUC'04 participants, commonly used open source text summarizer TexLexAn (TexLexAn, 2019) and other summarizer systems like ItemSum, Yago Summarizer, MSSF and PatSum systems by Baralis et al. (2013) and Qiang et al. (2016), For DUC systems, we have taken only six systems having best scores. Table 1 shows all the results of all systems with proposed MDS in terms of Recall, Precision and *F*-measure. The two best performing methods are marked as bold entries. From Table 1, we can observe that performance of proposed MDS system is improved than all other methods regarding ROUGE-2 Recall metric. This shows that proposed system has a high content coverage in final summary. The proposed MDS outperforms in result of ROUGE-4 concerning all Recall, Precision and *F*-measure than other systems.

In order to check the statistical significance of proposed approach performance improvement against other DUC'04 participant systems and other systems, a t-tests was performed on recall parameter for statistical significance to verify whether the improvement on summary of proposed approach over other approaches are statistically significant or not. T-test can be used when comparing simple means, when sample standard derivation is known. For the DUC 2004 participated systems, the summaries that were provided by the DUC 2004 were considered. Paired *t*-test ($p < 0.05$) was performed for DUC participated systems with significant level is 5% and number of sample is 20.

As shown in Table 2 the value of 1.6e-2 in row two and column two shows the associated p-value of the paired *t*-test between System65 and proposed MDS. As can be seen, all the tested p values are less than 5e-2. So at 95% confidence interval, the test demonstrates that our proposed MDS obtains significant results compared to the others.

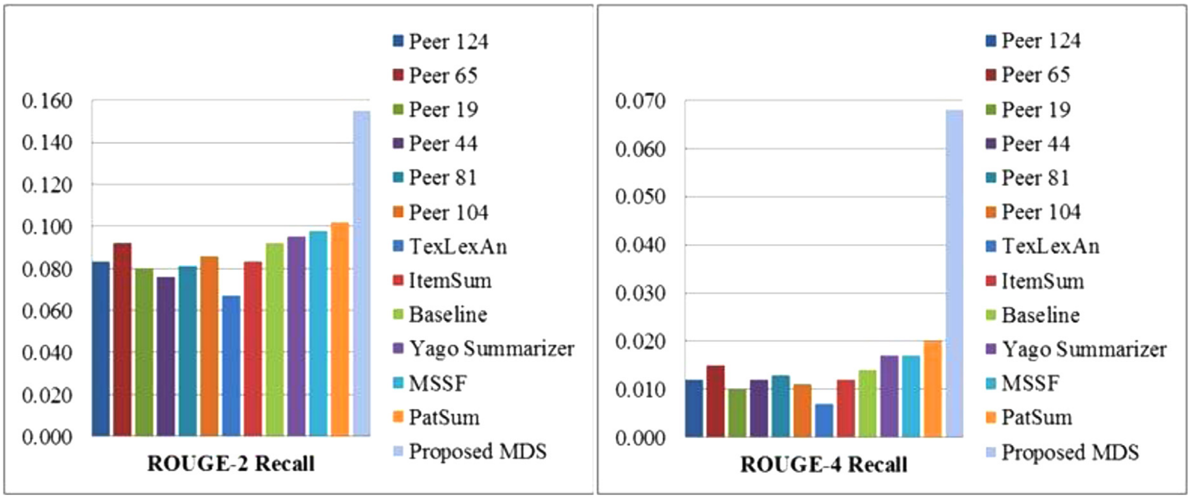

Fig. 4. Comparative chart for recall for ROUGE-2 and ROUGE-4.

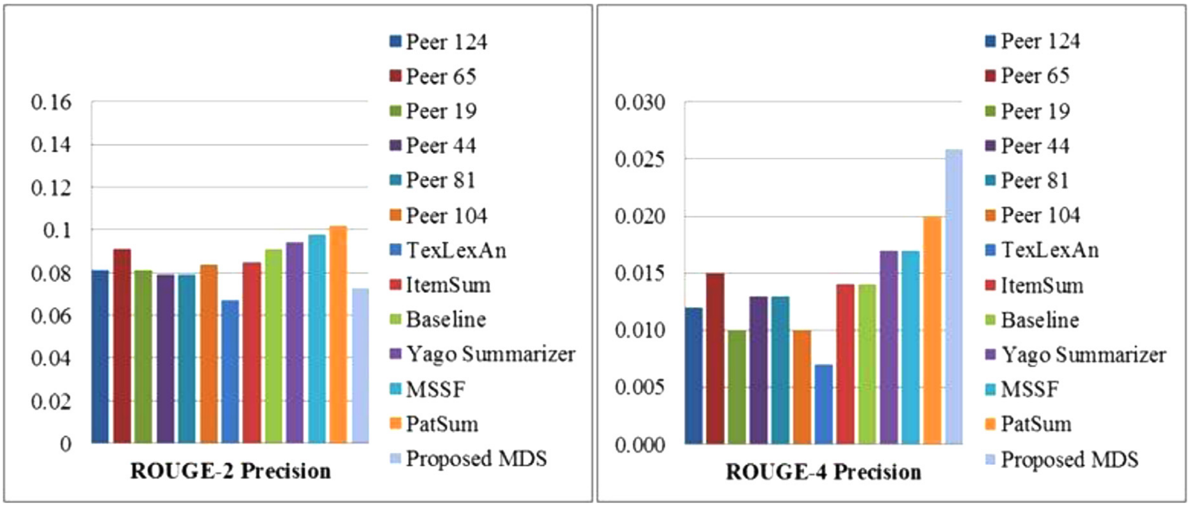

Fig. 5. Comparative chart for precision for ROUGE-2 and ROUGE-4.

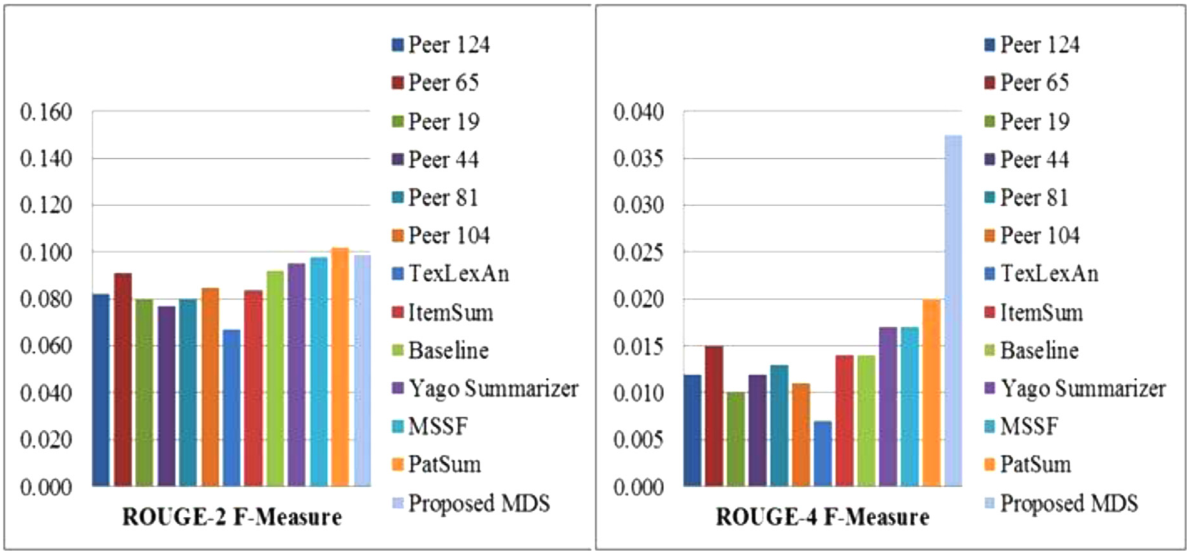

Fig. 6. Comparative chart for F-measure for ROUGE-2 and ROUGE-4.

Table 3
The associated *p*-values of the one sample *t*-test (95% significance level).

	ROUGE-2 Recall	ROUGE-4 Recall
Proposed Vs. TexLexAn	2e−3	5e−3
Proposed Vs. ItemSUM	7e−3	1.2e−2
Proposed Vs. Baseline	1.7e−2	1.2e−2
Proposed Vs. Yago	2.3e−2	1.9e−2
Proposed Vs. MSSF	3.2e−2	1.9e−2
Proposed Vs. PatSum	4.9e−2	2.8e−2

Table 3 shows associated p values for one sample T-test. For one sample *t*-test, we have taken 10 samples (Recall values for 10 randomly selected clusters) for proposed approach and for other system such as TexlexAn, ItemSUM, Baseline, Yago, MSSF and PatSum, we have considered their average best values reported by authors in their research paper.

Figs. 4–6 shows comparative chart for ROUGE-2 and ROUGE-4 score obtained by different summarization systems in terms of Recall, Precision and *F*-measure respectively.

7. Conclusion

This paper proposes and implements most important summarization approaches used in multi-document summarization. It focuses on text features based multi-document summarizer to create generic extractive summary. Rule based fuzzy logic is used for final sentences scoring. After accomplishment of the sentences scoring of all sentences, sentences are arranged in descending order based on their score from fuzzy inference system. It should be noted that tackling redundant information was the main issue in multi-document summarization. Cosine similarity measure is used to remove sentences which are having similar content from extracted salient sentences to generate final summary. All experiments are evaluated on DUC 2004 dataset using ROUGE 2 and ROUGE 4. Results show that proposed approach outperformed other systems.

There are several other directions to extend the present current work. The proposed approach is evaluated on News dataset, the strategies used in the sentence scoring are in better tuning process but depend on text corpus. Efforts in such direction can be performed and applied on other genre also. All the features used in proposed approach are language independent so by adding a language detection component, it can be possible to produce a multilingual, multi-document TS system. The proposed multi-document summarization system can be improved to generate the quality summary by adding some semantic and linguistics features and use of WordNet lexical data dictionary to the current version. As compare to single document summarization, multi-document summarization has more probability of ambiguity. That results in assignment of higher score values to some words improperly by the sentence scoring algorithm. Lexical chain may help to solve this problem. In future, we can improve our system with the help of morphological analyzer, a lexical database and semantic tools with statistical methods.

Sentence Ordering is difficult but very important task in document summarization and question answering system. In future we also want to work on sentence ordering issue to generate the coherent summary.

Author contribution

Prof. Darshna Patel, Dr. Saurabh Shah and Dr. Hitesh Chhinkaniwala, all the authors have contributed in Conceptualization, Investigation, Methodology, designing and implementation of the research work. Prof. Darshna Patel has contributed for Original Draft writing and both co-authors have reviewed and editing. All co-authors have provided critical feedback and helped shape the research, analysis and manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Darshna Patel: Conceptualization, Investigation, Methodology, Writing - review & editing, Writing - original draft. **Saurabh Shah:** Conceptualization, Investigation, Methodology, Writing - review & editing. **Hitesh Chhinkaniwala:** Conceptualization, Investigation, Methodology, Writing - review & editing.

References

Abbasi-ghalehtaki, R., Khotanlou, H., & Esmaeilpour, M. (2016). Fuzzy evolutionary cellular learning automata model for text summarization. *Swarm and Evolutionary Computation, 30*, 11–26. https://doi.org/10.1016/j.swevo.2016.03.004.

Al-Abdallah, R. Z., & Al-Taani, A. T. (2017). Arabic single-document text summarization using particle swarm optimization algorithm. *Procedia Computer Science, 117*, 30–37. https://doi.org/10.1016/j.procs.2017.10.091.

Alguliev, R. M., Aliguliyev, R. M., & Mehdiyev, C. A. (2011). Sentence selection for generic document summarization using an adaptive differential evolution algorithm. *Swarm and Evolutionary Computation, 1*(4), 213–222. https://doi.org/10.1016/j.swevo.2011.06.006.

Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications, 36*(4), 7764–7772. https://doi.org/10.1016/j.eswa.2008.11.022.

Baralis, E., Cagliero, L., Jabeen, S., Fiori, A., & Shah, S. (2013). Multi-document summarization based on the Yago ontology. *Expert Systems with Applications, 40*(17), 6976–6984. https://doi.org/10.1016/j.eswa.2013.06.047.

Bollegala, D., Okazaki, N., & Ishizuka, M. (2012). A preference learning approach to sentence ordering for multi-document summarization. *Information Sciences, 217*, 78–95. https://doi.org/10.1016/j.ins.2012.06.015.

Cabral, L., de, S., Lins, R. D., Mello, R. F., Freitas, F., Ávila, B., et al. (2014). A platform for language independent summarization. In *Proceedings of the 2014 ACM Symposium on Document Engineering - DocEng '14* (pp. 203–206). https://doi.org/10.1145/2644866.2644890.

Cagliero, L., Garza, P., Baralis, E., & Torino, P. (2019). ELSA : a A multilingual document summarization algorithm based on frequent itemsets and latent semantic analysis, 37(2), 1–33.

Cai, X., & Li, W. (2011). A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously. *Information Sciences, 181*(18), 3816–3827. https://doi.org/10.1016/j.ins.2011.04.052.

Chen, K. Y., Liu, S. H., Chen, B., & Wang, H. M. (2018). An information distillation framework for extractive summarization. *IEEE/ACM Transactions on Audio Speech and Language Processing, 26*(1), 161–170. https://doi.org/10.1109/TASLP.2017.2764545.

Chowanda, A. D., Sanyoto, A. R., Suhartono, D., & Setiadi, C. J. (2017). Automatic debate text summarization in online debate forum. *Procedia Computer Science, 116*, 11–19. https://doi.org/10.1016/j.procs.2017.10.003.

Fang, C., Mu, D., Deng, Z., & Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications, 72*, 189–195. https://doi.org/10.1016/j.eswa.2016.12.021.

Fang, H., Lu, W., Wu, F., Zhang, Y., Shang, X., Shao, J., et al. (2015). Topic aspect-oriented summarization via group selection. *Neurocomputing, 149*(PC), 1613–1619. https://doi.org/10.1016/j.neucom.2014.08.031.

Fattah, M. A. (2014). A hybrid machine learning model for multi-document summarization. *Applied intelligence, 40*(4), 592–600.

Ferreira, R., De Souza Cabral, L., Freitas, F., Lins, R. D., De França Silva, G., Simske, S. J., et al. (2014a). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications, 41*(13), 5780–5787. https://doi.org/10.1016/j.eswa.2014.03.023.

Ferreira, R., Freitas, F., Cabral, L., de, S., Lins, R. D., Lima, R., et al. (2014b). A context based text summarization system. In *2014 11th IAPR international workshop on document analysis systems* (pp. 66–70). https://doi.org/10.1109/DAS.2014.19.

Khan, A., Salim, N., & Jaya Kumar, Y. (2015). A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing Journal, 30*, 737–747. https://doi.org/10.1016/j.asoc.2015.01.070.

Le, H. T. (2013). An approach to abstractive text summarization, (January), 371–376. https://doi.org/10.1109/SOCPAR.2013.7054161

Liu, C., Wang, W., Tu, G., Xiang, Y., Wang, S., & Lv, F. (2017). A new centroid-based classification model for text categorization. *Knowledge-Based Systems, 136*, 15–26. https://doi.org/10.1016/j.knosys.2017.08.020.

Liu, Na, Li, M. X., Lu, Y., Tang, X. J., Wang, H. W., & Xiao, P. (2014). Mixture of topic model for multi-document summarization. In *26th Chinese control and decision conference, CCDC 2014* (pp. 5168–5172). https://doi.org/10.1109/CCDC.2014.6853102.

Liu, Nan, He, Y., Chen, Q., Peng, M., & Fang, W. (2013). Multi-document biased summarization based on topic-oriented characteristic database of term-pair Co-occurrence. In *2013 IEEE 3rd international conference on information science and technology, ICIST 2013* (pp. 832–837). https://doi.org/10.1109/ICIST.2013.6747670.

Lloret, E., Romá-Ferri, M. T., & Palomar, M. (2013). COMPENDIUM: A text summarization system for generating abstracts of research papers. *Data and Knowledge Engineering, 88*, 164–175. https://doi.org/10.1016/j.datak.2013.08.005.

Louis, A., & Nenkova, A. (2013). Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics, 39*, 267–300. doi:10.1162/coli_a_00123.

Malviya, S., & Tiwary, U. S. (2016). Knowledge based summarization and document generation using Bayesian network. *Procedia Computer Science, 89*, 333–340. https://doi.org/10.1016/j.procs.2016.06.080.

Mei, J. P., & Chen, L. (2012). SumCR: A new subtopic-based extractive approach for text summarization. *Knowledge and Information Systems, 31*(3), 527–545. https://doi.org/10.1007/s10115-011-0437-x.

Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., & León, E. (2014). Extractive single-document summarization based on genetic operators and guided local search. *Expert Systems with Applications, 41*(9), 4158–4169. https://doi.org/10.1016/j.eswa.2013.12.042.

Moradi, M., & Ghadiri, N. (2017). Different approaches for identifying important concepts in probabilistic biomedical text summarization. *Artificial Intelligence in Medicine, 84*, 101–116. https://doi.org/10.1016/j.artmed.2017.11.004.

Nandhini, K., & Balasundaram, S. R. (2013). Improving readability through extractive summarization for learners with reading difficulties. *Egyptian Informatics Journal, 14*(3), 195–204. https://doi.org/10.1016/j.eij.2013.09.001.

Open Source Text Analyzer Classifier Summarizer [WWW Document], 2019. [WWW Document]. Texlexan.sourceforge.net. URL http://texlexan.sourceforge.net/ (accessed 5. 30. 18).

Oufaida, H., Nouali, O., & Blache, P. (2014). Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. *Journal of King Saud University - Computer and Information Sciences, 26*(4), 450–461. https://doi.org/10.1016/j.jksuci.2014.06.008.

Pal, A. R., & Saha, D. (2014). An approach to automatic text summarization using WordNet. In *Advance computing conference (IACC), 2014 IEEE international. IEEE, 2014* (pp. 1169–1173). https://doi.org/10.1109/IAdCC.2014.6779492.

Patel, D., & Chhinkaniwala, H. (2018). Fuzzy logic-based single document summarisation with improved sentence scoring technique. *International Journal of Knowledge Engineering and Data Mining, 5*(1/2), 125–138.

Qiang, J. P., Chen, P., Ding, W., Xie, F., & Wu, X. (2016). Multi-document summarization using closed patterns. *Knowledge-Based Systems, 99*, 28–38. https://doi.org/10.1016/j.knosys.2016.01.030.

Rautray, R., & Balabantaray, R. C. (2017). Cat swarm optimization based evolutionary framework for multi document summarization. *Physica A: Statistical Mechanics and Its Applications, 477*, 174–186. https://doi.org/10.1016/j.physa.2017.02.056.

Sanchez-Gomez, J. M., Vega-Rodríguez, M. A., & Pérez, C. J. (2017). Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2017.11.029.

Tayal, M. A., Raghuwanshi, M. M., & Malik, L. G. (2017). ATSSC: Development of an approach based on soft computing for text summarization. *Computer Speech and Language, 41*, 214–235. https://doi.org/10.1016/j.csl.2016.07.002.

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications, 42*(4), 2264–2275. https://doi.org/10.1016/j.eswa.2014.10.023.

What Is Fuzzy Logic?- MATLAB & Simulink- MathWorks India [WWW Document], 2019. [WWW Document]. In.mathworks.com. URL https://in.mathworks.com/help/fuzzy/what-is-fuzzy-logic.html (accessed 4. 30. 18).

Wu, Z., Lei, L., Li, G., Huang, H., Zheng, C., Chen, E., et al. (2017). A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications, 84*, 12–23. https://doi.org/10.1016/j.eswa.2017.04.054.

Yadav, C. S., & Sharan, A. (2015). Hybrid approach for single text document summarization using statistical and sentiment features. *International Journal of Information Retrieval Research, 5*(4), 46–70. https://doi.org/10.4018/IJIRR.2015100104.

Yan, S., & Wan, X. (2014). SRRank: Leveraging semantic roles for extractive multi-document summarization. *IEEE/ACM Transactions on Audio Speech and Language Processing, 22*(12), 2048–2058. https://doi.org/10.1109/TASLP.2014.2360461.

Yousefi-Azar, M., & Hamey, L. (2017). Text summarization using unsupervised deep learning. *Expert Systems with Applications, 68*, 93–105. https://doi.org/10.1016/j.eswa.2016.10.017.

Yulianti, E., Chen, R. C., Scholer, F., Croft, B., & Sanderson, M. (2017). Document Summarization for answering non-factoid queries. *IEEE Transactions on Knowledge and Data Engineering, 4347*(c), 1–14. https://doi.org/10.1109/TKDE.2017.2754373.

5.3. Preprocessing data — scikit-learn 0.21.2 documentation [WWW Document], 2019. [WWW Document]. Scikit-learn.org. URL https://scikit-learn.org/stable/modules/preprocessing.html (accessed 5. 10. 18).