Egyptian Informatics Journal (2013) **14**, 195–204

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Improving readability through extractive summarization for learners with reading difficulties

K. Nandhini *, S.R. Balasundaram

Department of Computer Applications, National Institute of Technology, Tirucirappalli, India

Received 8 April 2013; revised 3 July 2013; accepted 22 September 2013
Available online 19 October 2013

KEYWORDS

Text summarization;
Extractive summarization;
Machine learning;
Naive Bayes classifier;
Assistive reading

Abstract In this paper, we describe the design and evaluation of extractive summarization approach to assist the learners with reading difficulties. As existing summarization approaches inherently assign more weights to the important sentences, our approach predicts the summary sentences that are important as well as readable to the target audience with good accuracy. We used supervised machine learning technique for summary extraction of science and social subjects in the educational text. Various independent features from the existing literature for predicting important sentences and proposed learner dependent features for predicting readable sentences are extracted from texts and are used for automatic classification. We performed both extrinsic and intrinsic evaluation on this approach and the intrinsic evaluation is carried out using F-measure and readability analysis. The extrinsic evaluation comprises of learner feedback using likert scale and the effect of assistive summary on improving readability for learners' with reading difficulty using ANOVA. The results show significant improvement in readability for the target audience using assistive summary.

1. Introduction

Over the past five decades, research on text summarization is widely seen in numerous applications related to information retrieval, intelligence gathering, information extraction, text mining, and indexing [1–3]. Automatic document summarization is the process of reducing the size of document preserving the important semantic content. Its purpose is to identify a summary of a document without reading the entire document. The main goal of a summary is to present the main ideas in a document, in less space. For general audience, the summarization technique has got variety of application like overview, gist and concept mapping. When focusing on a specific group of audience, the need and the purpose vary accordingly. The need for summarization can be thought of in varieties of applications including e-Learning. Recently, there is a tremendous growth in the use of electronic materials in most of the academic sectors. Many institutions prepare lecture notes and reading materials for courses and make them available

* Corresponding author.
E-mail address: nandhinikumaresh@yahoo.co.in (K. Nandhini).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

ELSEVIER Production and hosting by Elsevier

1110-8665 © 2013 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.
http://dx.doi.org/10.1016/j.eij.2013.09.001

in varieties of forms – computer based, web based, etc. One of the major issues in these kinds of materials is that they must be simple, readable and understandable by all. The generalized deficits in reading comprehension of many students with learning difficulties suggest the importance of systematic instruction in learning strategies.

The effect of summarization in comprehending text is significant for learners with reading difficulties [4]. Students with learning difficulties are diverse in nature and the difficulties include reading difficulties, math difficulties, writing difficulties or any other problems associated with learning process. As reading is an essential component of life, our focus is mainly on the learners with reading difficulties. Students with reading difficulties face significant challenges in comprehending and organizing information from text. Summarizing is taught either alone [5] or as one of several strategies [6] and is shown to improve comprehension and recall whatever is read [7]. As a strategy performed either during or after reading, summarizing helps readers to focus on main ideas or other key skill concepts that have been taught and to disregard less relevant ones [8]. The act of comprehending entails three essential elements: (1) the reader, (2) the text, and (3) the activity [9]. For learners with reading difficulty, the activity (summarization) can be used to assist them in reading comprehension. However, the problem of the learner cannot be cured but can be treated successfully [10] either through remedial or through supportive tools. There exist various tools for remedying or improving the skill like Word recognition, reciprocal teaching [6], trigger word teaching [11], graphic organizer [12], etc. Not all can be benefited by remedying instruction due to ability, age, and suitability. In such case providing support is next alternate solution by means of bypassing their disabilities. When the learner does not have significant effect under remedial course, the alternate option is to support the learner through assistive tool. There exist numerous assistive tools to support, to name a few are BrowseAloud, SpeakOut, etc. Aforementioned works focus on aiding toward reading process. But the ultimate goal of reading is to comprehend the text, which is less explored. To comprehend the given text, summarization is one of the important strategies that the learner has to apply [13]. If a learner is having problem in applying summarization strategy then he/she will be benefited by alternate solution namely assistive tool for extracting the important content from the text.

The design and evaluation of summarizing systems have to be related to the three classes of context factor namely input factors, purpose factors and output factors [9]. Most of the summarization systems available are designed for generic. There exist various specialized versions of summarization for the disabled, such as blind [14] deaf [15] and so on. The targeted audience and the purpose mainly determine the system design and evaluation [16]. Our targeted audiences are learners with reading difficulties those capable of decoding but find hard in understanding the content better. Many of these students have difficulty in finding main ideas and important supporting details. Failure to employ appropriate learning strategies is often a critical component of learning disabilities [17]. The purpose of assistive summary is to aid the reading difficulties in improving text readability which in turn understanding the content better. The major reason for difficulty in understanding the text for the target audience is due to working memory span. Considering the features like unfamiliar vocabulary [18], trigger words [11], polysyllabic words [19] and noun entity [20] for reducing the working memory difficulty. Our approach delves into extracting sentences that are important as well as readable to the target audience. Our objective is to extract summary that contains important, readable sentences. To solve this problem, we proposed a set of learner dependent readability related features in combination with set of existing features for extracting important sentences. Using Naive Bayes classifier the sentences are classified into summary or not a summary and the sentences are extracted based on the weighted score of individual features according to rate. We performed both type of evaluation to evaluate the performance of the summary and in particular extrinsic evaluation, the task of comprehension for reading difficulties.

The rest of the paper is organized as follows. In Section 2, we discuss the related works in text summarization. Section 3 describes proposed methodology for summary extraction. In Section 4, the results from the proposed approach are analyzed using both intrinsic and extrinsic evaluation.

2. Related work

Various researches on text summarization methods have been proposed and evaluated. There are two main approaches in the task of summarization namely extraction and abstraction [3]. Extraction involves concatenating extracts taken from the corpus into a summary, whereas abstraction involves generating novel sentences from information extracted from the corpus.

The extractive summarization techniques can be further classified into two groups: the supervised techniques that rely on pre-existing document summary pairs, and the unsupervised techniques, based on properties and heuristics derived from the text. Supervised extractive summarization techniques treat the summarization task as a two-class classification problem at the sentence level [21,22]. Many unsupervised methods have been developed for document summarization by exploiting different features and relationships of the sentences [23,24]. Early research on text summarization exploits various features such as word frequency [1], sentence position [25], cue phrases [26], sentence length and upper case letter [21], TF–IDF [27] and so on. Nowadays, corpus-based approaches play an important role in text summarization [21,28]. By exploiting technologies of machine learning, it becomes possible to learn rules from a corpus of documents and their corresponding summaries. The major advantage is that corpus-based approaches are easy to implement.

Much of the summarizing work done so far has not referred to summary use, which mainly decides the system design. The major weakness of extractive summarization is its poor readability. As readability is an essential component in text comprehension [29], extractive summarization lacks cohesion and sentence ordering. To achieve readable summary, extracted sentences must be appropriately ordered [29,30]. Yeh et al. proposed a modified corpus-based approach using Naive Bayes classifier [22]. Our idea is to combine the learner dependent features with existing features and the feature weights calculated using GA. Text summarization can be evaluated in two ways namely intrinsic and extrinsic evaluation. The

quality of the summary can be evaluated directly based on various norms like the fluency of the summary [32,33], coverage of stipulated "key/essential ideas" in the source [34,33], or similarity to an "ideal" summary [26,21]. Extrinsic evaluation measures the quality of the summarization based on how it affects the completion of some other task. To name a few are question–answering, comprehension tasks [35] and so on. We evaluated the proposed application using likert scale feedback and comprehension questions extrinsically.

2.1. Corpus-based approach

Kupiec et al. proposed a supervised machine learning algorithm to train a summarizer using Naive Bayes classifier [21]. For each sentence S, the probability that it will be classified as a summary sentence, given the N features. The Naive Bayes rule is as follows:

$$P(s \in S \mid F1, F2, \ldots Fn) = \frac{P(F1, F2, \ldots Fn \mid s \in S)P(s \in S)}{P(F1, F2, \ldots Fn)} \quad (1)$$

Assuming statistical independence of the features:

$$P(s \in S) \mid F1, F2, \ldots Fn) = \frac{\prod_{j=1}^{n} P(Fj \mid s \in S)P(s \in S)}{\prod_{j=1}^{n} P(Fj)} \quad (2)$$

$P(s \notin S)$ is a constant and $P(F_j|s \notin S)$ and $P(F_j)$ can be estimated directly from the training set by counting occurrences. This produces a simple Bayesian classification function that assigns for each S, a score which can be used to select sentences for inclusion in a generated summary.

Yeh et al. proposed the modified corpus-based approach using trainable summarizer that considers features such as sentence position, keywords, centrality and title similarity and the feature weights are calculated using genetic algorithm [22]. Fattah and Ren investigated the effect of each sentence featured on summarization task and used all features in combination to train genetic algorithm (GA) and mathematical regression (MR) models to obtain a suitable combination of feature weights.

3. Proposed methodology

Using trainable summarizer, the existing features such as sentence position, keywords, title similarity and centrality along with proposed features such as sentence length, trigger words, polysyllabic words and noun occurrences are extracted. The sentence score is calculated using the weighted combination of features, where the weights are calculated using genetic algorithm. The main objective of this work is to summarize the given text. Each sentence is represented by a set of predefined features $(F_1, F_2, \ldots, F_n)$ then a supervised learning algorithm is used to train the summarizer to extract important and readable sentences, based on feature vectors. In order to perform this extraction, potential features, classifier and a training corpus of document summary pair are required. The flow diagram of the summary extraction is given in Fig. 1.

The following features are considered, as they consider both important and readability for extracting summary sentences.

3.1. Feature extraction

The feature extraction depends mainly on the purpose and the target audience of the text summarization. The features have to represent the text importance, learner difficulty factor and the motivation of the application. To identify the important sentences from the text, features like sentence position [25], centrality [36], title similarity [21] and keyword occurrences [26] are considered. The use of summarization is to aid the reading difficulties.

As readability is one of the prime factors to be considered, we used features used by the Flesh-Kincaid [37] and FOG [38] metrics such as average words per sentence, average syllables per word, and percentage of words in the document with 3+ syllables. Sentences with more entities are more difficult for users to semantically encode due to working memory limitations [20]. So, the feature noun occurrences measures the average number of entities per sentence the reader must keep in mind to understand the sentence. As the learner may have problem in mapping trigger words [11], average trigger words per sentences is considered as another feature. The significance of the learner dependent features is discussed in detail in [39]. The objective is to extract the important and readable sentences as summary. Documents from educational text books are used for evaluation. The contents of documents are segmented into sentences. Then we extract words from each sentences and represent each sentence as vector of content words, words tagged with noun or verb tag are considered as content words.

3.1.1. Informative features

3.1.1.1. F1-sentence position. Sentence position is an important factor to decide sentence importance in the document [25,26]. The significance of position of sentence plays a vital role in various domains accordingly. The important sentences to be included in the summary are usually located in a particular position. In the case of educational text, boundaries carry important information.

3.1.1.2. F2-title to sentence similarity. Titles contain the group of words that give important clues about concepts of the text [21,26]. If the sentence has higher intersection with the title words then the score for the sentence is calculated using the formula given in Table 1.

3.1.1.3. F3-centrality. The centrality of the sentence implies its similarity to other sentences [36]. If a sentence has higher centrality, then those sentences are the best candidates to be included in the summary.

3.1.1.4. F4-cue phrases. The sentences contain cue phrases like defined as, called, significant, means, important, etc., contain important definitions that are to be added in the important class category [26].

3.1.1.5. F5-keywords. The importance of the word is determined by its term frequency which is calculated as follows: TF–ISF is a statistical technique used to evaluate how important a word is to a document [40,27].

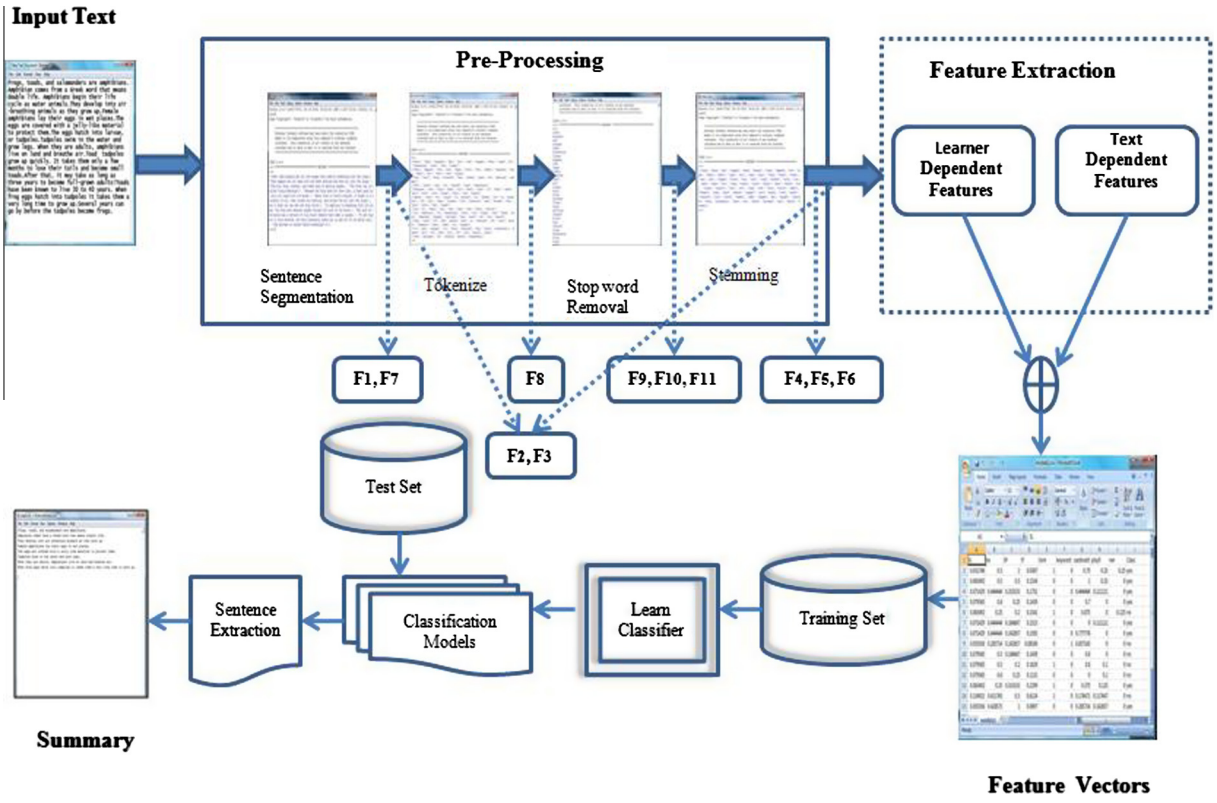

Figure 1 Summarization process.

Table 1 Informative features.

S. no.	Features	Formula
1	F1: Sentence position	$SP = Max\left(\frac{1}{i}, \frac{1}{n-i+1}\right)$
2	F2: Title to sentence similarity	$\mathrm{TS} = \frac{\sum_{i=1}^{n} AiXBi}{\sqrt{\sum_{i=1}^{n}(\mathrm{Ai})^2} X \sqrt{\sum_{i=1}^{n}(\mathrm{Bi})^2}}$
3	F3: Centrality	$Cen = \frac{\lvert Words \text{ in Si } \cap \text{Words in other S}\rvert}{\lvert \text{Words in Si } \cup \text{Words in other S}\rvert}$
4	F4: Cue phrases	As in database
5	F5: Keywords	$TFi = \frac{Ti}{\sum_{k=1}^{n} Tk}, ISF = \log\frac{N}{ni}$
		$Score = TF * ISF$

In order to show that these features are contributing and are independent of each other. The correlation of individual feature in predicting the summary sentences is analyzed. In readability research, analysis focuses much on correlation. A correlation coefficient ($r =$) is a descriptive statistics that ranges from $+1.00$ to 0.0 or 0.0 to -1.00. Both $+1.00$ and -1.00 represent prefect correlation depending on whether the elements are positively or negatively correlated. To study the relation of the features, we analyzed the text in original and summary version. The correlation of individual feature (informative features) in summary extraction is explained in Section 4.2.

3.1.2. Readability features

As the targeted audiences are having reading difficulties, features that extract easy to understand sentences helps them in reading and in turn understanding the text better.

3.1.2.1. F6-sentence length. The major reason for having difficulty in text comprehension is due to the individuals working memory problem. The reader was unable to integrate the semantic of words present in the sentences. When sentence length is more, the learner may fail to integrate it completely. Due to short term memory and sequencing difficulties,

Table 2 Readability features.

S. no.	Features	Formula
6	F6: Sentence length	$ASL = \frac{\#(words\ in\ Si)}{Length(Si)}$
7	F7: Trigger words	$PTW = \frac{\#(Trigger\ word\ in\ Si)}{Length(Si)}$
8	F8: Polysyllabic words	$PPSY = \frac{\#polysyllabic\ words\ in\ Si}{Length(Si)}$
9	F9: Noun occurrences	$PNE = \frac{\#propernouns\ in\ Si}{LengthofSi}$

students can lose their way in comprehending texts because they are not retaining important facts and therefore do not fully understand the text.

3.1.2.2. F7-trigger words. Occurrences of more number of trigger words create distraction to the dyslexic reader [41]. Sentences having higher number of trigger words are considered as difficult sentences. The percentage of trigger words in the text can be calculated using the formula given in Table 2.

3.1.2.3. F8-polysyllabic words. Sentences containing longer and in frequent (hard) words are hard to follow. Sentences having higher number of hard words are considered as difficult sentences. Words that are harder to decode will increase the difficulty level of comprehension. The average numbers of syllables and the average sentence length have been used to determine the reading difficult of the text as in SMOG [19]. Hence the number of syllables in a word is also one of the measures of word difficulty. The percentage of polysyllabic words in the text can be calculated as shown in Table 2.

3.1.2.4. F9-noun occurrences. The number of nouns introduced in a text, relates to the working memory burden on their targeted readers such as dyslexic [20]. The noun counts in individual sentences also decide its easiness or difficult nature. Usually the sentence that contains more proper nouns is an important one and it is most probably included in the summary. But for the readers with difficulty, more nouns create confusion.

3.2. Summary generation

The sentence score can be calculated by linear weighted combination of all features. In this work, Features are extracted from the text. The summary generation can be seen as a two-class classification problem using Naive Bayes classifier [21], where a sentence is labeled as summary if it belongs to the extractive reference summary or as not a summary otherwise.

4. Evaluation

4.1. Dataset

A collection of hundred articles from educational text of grade four to grade seven are used for evaluation. The texts are mainly from science and social subjects because the magnitude of concepts in social studies and science full of abstract concepts are the challenges faced by the reading difficulties. The text basically belongs to two levels of language one is from the text books [42] and another is of grade level text [43]. The statistics of the corpus is shown in Table 3.

The reference summaries of this text are created manually by three independent human annotators. One of them rank the sentences based on its importance and another rank the sentences based on its difficulty and the third scores whether it is easy, hard or average. The relative score is considered for classifying the reference summary. The significance of each sentences to be included in summary are ranked by the annotators. The sentences that score maximum rank are included in summary according to the.

4.2. Feature correlation

In order to prove the effect of individual features, each feature has to be independent and the correlation between the feature and the class to be good. To predict the feature correlation, the gold summary along with the annotator rank is considered. For informative features, there exists positive correlation between the features and the gold summary. The correlation coefficient of various features is shown in Table 4.

In the case of readability features, There exists negative correlation between the features and the class. This shows that when the score for the features increase, the possibility of being in the summary class is less. The correlation coefficient of various readability features are shown in Table 5.

Negative correlation shows that if these feature value increases, the difficulty of the sentences increases. However these features show either positive correlation or negative correlation with the class and minimal or no correlation between the features.

4.3. Feature significance

In our experiments, the training and testing are done using 70% and 30% of datasets. The following Table 3 shows the effect of individual features in sample dataset. The effect of individual feature's average F-measure is tabulated. The symbols and its explanation are given in Table 6.

From the Table 7, it is evident that the features like positive keyword, percentage of trigger words, percentage of polysyllabic words and sentence position have higher score when compared to other features in the case of 10% compression rate as in Table 8.

As the dataset basically belongs to educational text, the significance of keywords is more in the case of definitions and explanations. The occurrence of trigger words is inversely proportional to the content word in a sentence. The percentage of trigger words plays a significant role in classification. Percentage of polysyllabic has next higher score where the occurrence of polysyllabic words is about 20% of the text. The sentence position has next higher F-measure due to the nature of the text.

Table 3 Statistics of dataset.

Number of docs	100
Average # of sent. per doc.	18
Maximum # of sent. per doc.	38
Minimum # of sent. per doc.	11
Summary as% of doc. length	30%
Average summary size (in # of sent)	13
Maximum # of sent. per summary	15
Minimum # of sent. per summary	8

Table 4 Correlation of informative features.

S. no.	Features	Correlation coefficient
1	Sentence position	.52
2	Title similarity	.471
3	Cardinality	.35
4	Cue phrases	.37
5	Keywords	.41

Table 5 Correlation of readability related features.

S. no.	Features	Correlation coefficient
1	Average sentence length	−.37
2	Percentage of trigger words	−.35
3	Percentage of polysyllabic words	−.31
4	Percentage of noun entity	−.39

Table 6 Notations.

S. no.	Symbols	Features
1	ASL	Average sentence length
2	PTW	Percentage of trigger words
3	SP	Sentence position
4	TF	Term frequency
5	TS	Title similarity
6	PKey	Positive keyword
7	Nkey	Negative keyword
8	Cen	Centrality
9	PPSY	Percentage of polysyllabic word
10	PNE	Percentage of noun entity
11	PMSY	Percentage of monosyllabic words

Table 7 F-measure of individual features.

S. no.	Features	10%	20%	30%
1	ASL	0.167	0.454	0.590
2	PTW	0.234	0.419	0.469
3	SP	0.205	0.363	0.450
4	TF	0.174	0.438	0.548
5	TS	0.170	0.294	0.471
6	PKey	0.288	0.466	0.502
7	Nkey	0.147	0.248	0.359
8	Cen	0.177	0.392	0.450
9	PPSY	0.227	0.446	0.570
10	PNE	0.174	0.305	0.425
11	PMSY	0.155	0.230	0.349

Table 8 Leave one out features statistics.

S. no.	Features	10%	20%	30%
1	ASL	0.257(+)	0.513(+)	0.738(+)
2	PTW	0.259(+)	0.532(+)	0.738(+)
3	SP	0.237(+)	0.499(+)	0.686(+)
4	TF	0.275(+)	0.527(+)	0.713(+)
5	TS	0.257(+)	0.497(+)	0.655(+)
6	PKey	0.297(−)	0.527(+)	0.707(+)
7	Nkey	0.299(−)	0.572(−)	0.745(−)
8	Cen	0.285()	0.561(−)	0.703(+)
9	PPSY	0.241(+)	0.483(+)	0.661(+)
10	PNE	0.289()	0.567(−)	0.722(+)
11	PMSY	0.295(−)	0.573(−)	0.741(−)

As the increases the effect of individual feature varies. For 30%, average sentence length, percentage of polysyllabic words, term frequency and positive keyword scores higher individual F-measure. In the case of 10% compression rate, the F-measure is 0.284 when all features are considered. However, when we used all features except positive keyword, or Negative keyword or percentage of monosyllabic count the F-measure is higher when compared with others. These features have negative effect in that corpus. On the other hand, when all features except centrality or percentage of noun entity are considered, the F-measure remains same. These features have neutral effect in the corpus. When all features except, Average sentence length or percentage of polysyllabic words or sentence position or term frequency or title similarity are considered.

The F-measure gets decreased which shows that these features have positive effect in the dataset. When all features except sentence length is considered, the F-measure is very low and hence significant feature for 10% compression rate. The same centrality feature has negative and positive effect respectively. When the compression rate is 20% and 30% and when the Compression rate is high, centrality feature has its effect due to more number of sentences to be considered. The best combination features for 30% compression rate are average sentence length, percentage of trigger words, sentence position, term frequency, title similarity, positive keyword, centrality and percentage of noun occurrences. The feature like negative keyword and percentage of monosyllabic count are ignored due to its negative effect in the dataset. The effect of informative features, readability features and combined features are evaluated using Precision, Recall and F-Measure. The comparison can be made with an extract of a system with the gold summary.

Table 9 shows the average precision, average recall and average accuracy for a sample set of documents. The text based features focus on extracting important sentences whereas learner based features focus on extracting understandability related sentences. It is evident that the accuracy of the classifier increases when the combined features are used. The idea of extracting the easy to understand sentence, does not dilute the accuracy of extracting important sentences.

4.4. Intrinsic evaluation

The quality of the summary based on the coverage between it and the manual summary. The combined feature model is compared with baseline approach Lead method (Top N sentences of the text as summary according to the compression rate) and with MCBA [31] for various compression rates. The results for precision are tabulated in Table 10.

When compression rate is 10%, the baseline method and MCBA produces almost nearer results. But when ratio increases the difference between them are significant. The proposed model predicts better than lead approach by 4.6%, 18.3% and 27.2% for 10%, 20% and 30% respectively. But when compared to MCBA it produces 3.2%, 4.2% and 8.2% respectively. The reason is that the proposed model differs from MCBA only by a set of features for extracting readable sentences and to reduce the difficulty of the learners. The aim is not to improve the performance through precision or recall but to improve the readability of the extracted sentences.

Table 9 Category based features.

Features	Avg. precision	Avg. recall	Avg. accuracy
Readability features	0.75	0.80	0.76
Informative features	0.84	0.80	0.78
Combined features	0.82	0.82	0.79

Table 10 Performance evaluation (precision) when compared to MCBA and Lead method.

	10%			20%			30%		
	Lead	MCBA	Proposed	Lead	MCBA	Proposed	Lead	MCBA	Proposed
Set1	0.33	0.33	0.33	0.423	0.5	0.5	0.5	0.66	0.83
Set2	0.33	0.25	0.33	0.33	0.5	0.5	0.53	0.5	0.66
Set3	0.4	0.5	0.5	0.5	0.56	0.66	0.4	0.83	0.83
Set4	0.2	0.25	0.4	0.5	0.65	0.75	0.66	0.75	0.8
Set5	0.4	0.4	0.33	0.4	0.6	0.66	0.5	0.8	0.83
Average	0.332	0.346	0.378	0.4306	0.482	0.614	0.518	0.708	0.79

Table 11 Performance evaluation (recall) when compared to MCBA and lead method.

	10%			20%			30%		
	Lead	MCBA	Proposed	Lead	MCBA	Proposed	Lead	MCBA	Proposed
Set1	0.3	0.25	0.25	0.4	0.375	0.475	0.375	0.5	0.625
Set2	0.3	0.125	0.25	0.5	0.35	0.375	0.375	0.25	0.5
Set3	0.2	0.5	0.5	0.5	0.55	0.666	0.4	0.833	0.833
Set4	0.3	0.142	0.285	0.4	0.428	0.428	0.5	0.428	0.571
Set5	0.4	0.285	0.285	0.4	0.428	0.571	0.5	0.571	0.714
Average	0.301	0.260	0.314	0.44	0.406	0.503	0.43	0.516	0.648

Table 12 Performance evaluation (F-measure) when compared to MCBA and lead method.

	10%			20%			30%		
	Lead	MCBA	Proposed	Lead	MCBA	Proposed	Lead	MCBA	Proposed
Set1	0.314	0.284	0.284	0.411	0.428	0.487	0.428	0.568	0.713
Set2	0.314	0.166	0.284	0.397	0.333	0.428	0.439	0.333	0.568
Set3	0.266	0.5	0.5	0.5	0.554	0.662	0.4	0.831	0.831
Set4	0.25	0.181	0.332	0.444	0.544	0.544	0.568	0.544	0.666
Set5	0.4	0.332	0.305	0.4	0.499	0.612	0.5	0.666	0.767
Average	0.309	0.293	0.341	0.430	0.472	0.547	0.467	0.589	0.709

Table 13 Characteristics of various readability metrics.

Readability metric	Readability formulae	Lexical feature	Grammatical feature
FOG	3.0680 + 0.877 * ASL + 0.984 * % MSW	Percentage of Mono Syllable Words (MSW)	Average sentence length
SMOG	3 + Sqrt (Number of PSW).	Poly Syllable Word count (PSW)	ASL (implicit)
Flesch Kincaid	0.39 * Avg #Words/Sentence + 11.80 * Avg # Syllables/Word −15.59	Syllables per word	ASL

The performance of proposed model can be compared to existing methods with respect to recall are shown in Table 11. The relative improvement over baseline method in average recall of proposed model is 1.2%, 6.3% and 21% for various compression rates. In the case of MCBA the relative improvement is 5%, 9.6% and 13.2%. In the terms of misclassification, the proposed model performs better when compares to baseline method and MCBA up to 21% and 13% respectively.

The relative improvement in F-measure of proposed over Lead method is 3.2%, 11.6% and 24.2% for 10%, 20% and 30% compression rate respectively. Moreover when compared to MCBA, the proposed model has 4.8%, 7.4% and 12.04% improvement in the case of 10%, 20% and 30% respectively and is shown in Table 12. It is evident that the performance of the model as got improvement in the case of precision, recall and F-measure for various level of compression rate. As the objective is to the improve readability, the readability of the extracted summary can be measured using various existing metrics.

4.5. Readability

Dale and Chall proposed the following definition of readability [44]: Readability is the sum total (including the interactions) of

Table 14 Readability of summary in various metrics.

	Flesch Kincaid			Gunning fox index			SMOG index		
	Baseline	MCBA	Proposed	Baseline	MCBA	Proposed	Baseline	MCBA	Proposed
Set1	2.8	3.2	2.6	5.7	5.8	5.2	4.4	4.4	4.4
Set2	6.2	5.4	5.4	8.7	8	8.1	6.3	6	6
Set3	3.1	3.1	2.5	4.9	4.9	4.8	1.8	1.8	1.8
Set4	6.6	6.4	5.8	10.6	10.4	9.6	7.8	7.7	7.1
Set5	4.8	4.3	4.2	7.6	7.5	7.5	5.4	5.4	5

all those elements within a given piece of printed material that affects the success, a group of readers have with it. The success is the extent to which they understand it, read it at an optimum speed, and find it interesting. The term readability is used in this paper refer to the ease with which a reader can read and understand text. To date, there exist numerous readability metrics that use simple linear functions with combination of two or three shallow language features. These shallow language features are lexical and syntactic features. Basically the lexical features focuses on measuring the difficulty of the word such as number of syllables in a word, number of characters in a word and word frequency. The syntactic feature focuses on measuring the difficulty of the sentence by the average sentence length. For predicting the readability of the text the metrics we considered are FOG [38], SMOG [19] that use average sentence length and the percentage of words with at least three syllable as parameters, Flesch Kincaid [37] use average sentence length and average syllables per word to calculate text difficulty.

Different readability formulae are likely to give different scores on the same piece of text, as they give different weights to various aspects of the text is shown in Table 13. The idea of comparing the readability of the original text and the summary by various readability metrics will help in analyzing whether the extractive summarization helps in comprehending the text better. After summarizing the text, increase in readability will project the lexical and syntactic complexity of the text. On the whole, from these comparisons the impact of summarization in readability can be easily drawn.

The readability of the proposed model is relatively better than existing model and is tabulated in Table 14. Though the difference is less, the predicted grade level is lower than other models. The difference in readability is very minimum in the case of SMOG as it only considers polysyllabic word. The score is somewhat higher in the case of Flesch Kincaid, that includes ASL as well as monosyllabic words and it considers all words including trigger words to be monosyllabic, but these learners have problem in trigger words when compared to others. None of these metric considers features like average sentence length, Polysyllabic words, trigger words and grade level vocabulary all into account.

4.6. Extrinsic evaluation

Fifteen reading difficulties, of grade four to seven were participated in the study. A comparison group of fifteen participants without reading difficulties also participated in the experiment. The experiment was composed of two parts namely

1. Learner self-perception using likert scale feedback.
2. Objective type comprehension questions to measure the impact of the summary.

Through the user survey, we infer how the participants were influenced by assistive summary in text reading and comprehension.

4.6.1. Learner self-perception using likert scale feedback readability

The readability of the text can be measured by asking subjective question to the participants that how much easy it is to understand a given text. Likert scale measure can be used to measure the meta-comprehension of text in educational studies. The survey questions consist of

1. Whether assistive summary helps in improving the readability of the text?
2. Which of this summary is useful/easy?

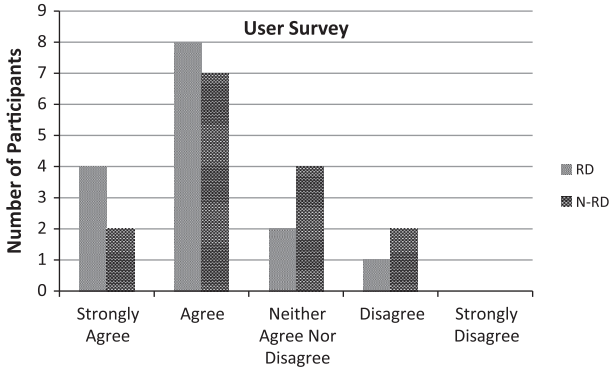

Figure 2 Likert scale feedback.

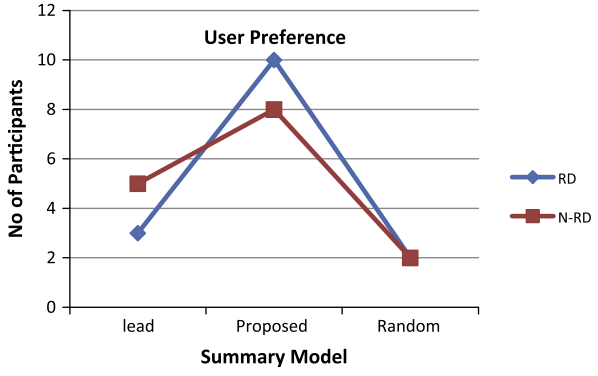

Figure 3 User preference in summary.

Table 15 Effect of assistive summary on learner for research comprehension.

Learner	Summary approach	Mean	SD	Count
Reading difficulties	Random	1.93	0.63	15
	Proposed	3.33	0.66	15
	Total	2.63	1.136	30
Non-reading difficulties	Random	1.93	0.209	15
	Proposed	2.26	0.63	15
	Total	2.1	0.43	30

The survey question one helps in predicting the summary use in readability. Fig. 2 shows the user agreement in assistive summary benefit (Readability). In case of Readability, Reading difficulties agree in assistive summary for improving text readability (0.8). Whereas for non-difficult learners about (0.4), finds it less useful. Few reported that it misleads them for not to focus on other sentences in the text. Few strongly agree that it is very useful. In case of survey question two, participants are given three texts as mentioned in reading test. And which summary helps them in reading the text better was analyzed. The learner feedback in summary preference is shown in Fig. 3.

The summary generated by proposed method is mostly the preference for both reading difficulties and non-difficult readers. However few of them also opted for lead due to its cohesive nature and random due to confusion in selecting the choice. Another type of extrinsic evaluation for summarization is the reading comprehension task [45,35]. In such an evaluation, a user is given a either full source or full text supplemented with summary, along with multiple-choice questions relating to the full source information. A system can then calculate how well they perform on the test given the condition. This evaluation framework relies on the idea that truly informative summaries should be able to act as assistive summary to the full source. We are considering this summary as assistive and not the replacement of the original text.

4.6.2. Objective type questionnaire

The direct way to measure comprehension of information from a text is through objective comprehension questions such as multiple choices or yes/no used in physiological studies of comprehension. Each text is evaluated using six questions, four of them are factual questions and two of them are inferential questions. The effects of assistive summary on reading difficulties in answering comprehension questions are measured. The learner (Reading Difficult and Non-Difficult Learner) response on with and without assistive summary (lead and proposed) are analyzed. The major questions underlined are

1. Whether there was any significant difference in reading comprehension between students who were assisted using summary and conventional reading strategy.
2. Whether there was any significance difference in reading comprehension between learner who has assisted using proposed approach or baseline approach.

To examine the effect of assistive summary on reading difficulties for reading comprehension, descriptive statistics were calculated for this factor. Table 15. List the mean and standard deviation, as well as overall value for each group.

Though there was significant effect of assistive summary in answering questions among reading difficulties, the number of factual questions answered correctly is more when compared to inferential questions. From the analysis, it was clear that the assistive summary improves readability and recall in the reading difficulties than understanding it in a better way. A comparison of the mean across the groups show that the proposed ($M = 3.33$) performed generally better than the baseline ($M = 1.93$). Regarding the effect on learner there is a difference between RD ($M = 3.33$) and NRD ($M = 2.26$) among the learners. In order to compute statistical significance, the data were submitted to a two way ANOVA with Assistive summary approaches and learner as between group variables. From the results of two way ANOVA, $F(1,14) = 10.36$, $P = .00^{*}$, there is significant effect of assistive summary on reading comprehension. Moreover there is significant effect of learner $F(1,14) = 1.52$, $P = .004$. This means summary have significant effect on reading difficulties than non-reading difficulties. It is clear that the proposed summary has significant effect on reading difficulties than existing methods due to the consideration of readability related features. When the text is hard, summary extracted from the text is also hard. So, the comprehension needs to improve through various simplification strategies.

5. Conclusion

This article proposes a novel application for extractive summarization to the learners' with reading difficulties. The major contribution in this work is proposing a set of learner dependent features for extracting easy to understand summary. Experimental results on educational text from text books and grade level texts demonstrate the effectiveness of the summarization approach and the combined effect of informative features and readability features. The extrinsic evaluation result shows the significant effect of assistive summary in improving readability for the target audience. As lack of cohesion is a major drawback of extractive summarization, improving cohesion in extractive summarization and extracting individualized summary is our future work.

References

[1] Luhn H. The automatic creation of literature abstracts. IBM J Res Develop 2002;2(2):159–65.
[2] Mani I, Maybury M. Advances in automatic text summarization. MIT Press; 1999.

[3] Hahn U, Mani I. The challenges of automatic summarization. Computer 2000;33(11):29–36.
[4] Gajria M, Jitendra A, Sood S, Sacks G. Improving comprehension of expository text in students with ld a research synthesis. J Learn Disabil 2007;40(3):210–25.
[5] Armbruster B, Anderson T, Ostertag J. Does text structure/summarization instruction facilitate learning from expository text? Read Res Quart 1987:331–46.
[6] Palinscar A, Brown A. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognit Instruct 1984;1(2):117–75.
[7] National Reading Panel. Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction, National Institute of Child Health and Human Development, National Institutes of Health; 2000.
[8] Kamil M. Vocabulary and comprehension instruction: summary and implications of the national reading panel findings. Voice Evid Read Res 2004:213–34.
[9] Reutzel D, Smith J, Fawson P. An evaluation of two approaches for teaching reading comprehension strategies in the primary years using science information texts. Early Child Res Quart 2005;20(3):276–305.
[10] Hung L, Huang K. Two different approaches to radical-based remedial Chinese reading for low-achieving beginning readers in primary school. Bull Special Educ 2006;31:43–71.
[11] Davis R, Braun E. The gift of dyslexia: why some of the smartest people can't read... and how they can learn, penguin, group; 2010.
[12] Rello L, Saggion H, Baeza-Yates R, Graells E. Graphical schemes may improve readability but not understandability for people with dyslexia. NAACL-HLT; 2012.
[13] Gajria M, Salvia J. The effect of summarization instruction on text comprehension of students with learning disabilities. Except Child 1992;58(6):508–16.
[14] Grefenstette G. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In: Working notes of the AAAI spring symposium on intelligent, text summarization; 1998. p. 111–8.
[15] Vandeghinste V, Pan Y. Sentence compression for automated subtitling: A hybrid approach. In: Proceedings of the ACL workshop on, text summarization; 2004. p. 89–95.
[16] Sparck Jones K. Automatic summarising: the state of the art. Inform Process Manage 2007;43(6):1449–81.
[17] Graves A. Effects of direct instruction and meta comprehension training on finding main ideas. Learn Disabil Res 1986.
[18] Hsueh-Chao MH, Nation P. Unknown vocabulary density and reading comprehension. Read Foreign Lang 2000;13(1):403–30.
[19] Mc Laughlin G. Smog grading-a new readability formula. J Read 1969:639–46.
[20] Feng L, Elhadad N, Huenerfauth M. Cognitively motivated features for readability assessment. In: Proceedings of the 12th conference of the European chapter of the association for, computational linguistics; 2009. p. 229–37.
[21] Kupiec J, Pedersen J, Chen F. A trainable document summarizer. In: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval. ACM; 1995. p. 68–73.
[22] Yeh J, Ke H, Yang W, Meng I. Text summarization using a trainable summarizer and latent semantic analysis. Inform Process Manage 2005;41(1):75–95.
[23] Mihalcea R, Ceylan H. Explorations in automatic book summarization. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL); 2007. p. 380–9.
[24] Aliguliyev R. Automatic document summarization by sentence extraction. Comput Technol 2007;12(5):5–15.
[25] Baxendale P. Machine-made index for technical literature an experiment. IBM J Res Develop 1958;2(4):354–61.
[26] Edmundson H. New methods in automatic extracting. J ACM 1969;16(2):264–85.
[27] Aone C, Okurowski M, Gorlinsky J. Trainable, scalable summarization using robust nlp and machine learning. In: Proceedings of the 17th international conference on computational linguistics, Association for computational linguistics, vol. 1; 1998. p. 62–6.
[28] Lin C, Hovy E. Identifying topics by position. In: Proceedings of the fifth conference on applied natural language processing. Association for computational linguistics; 1997. p. 283–90.
[29] Barzilay R, Elhadad N, McKeown K. Inferring strategies for sentence ordering in multidocument news summarization. J Art Intell Res 2002;17:35–55.
[30] Lapata M. Probabilistic text structuring: experiments with sentence ordering. In: Proceedings of the 41st annual meeting on association for computational linguistics, Association for computational linguistics, vol. 1; 2003. p. 545–52.
[31] Fattah M, Ren F. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Comput Speech Lang 2009;23(1):126–44.
[32] Minel JL, Nugier S, Piat G. How to appreciate the quality of automatic text summarization? In: Proc of the ACL/EACL'97 workshop on intelligent scalable text summarization, Citeseer; 1997. p. 25–30.
[33] Brandow R, Mitze K, Rau LF. Automatic condensation of electronic publications by sentence selection. Inform Process Manage 1995;31(5):675–85.
[34] Paice CD. Constructing literature abstracts by computer: techniques and prospects. Inform Process Manage 1990;26(1):171–86.
[35] Morris A, Kasper G, Adams D. The effects and limitations of automated text condensing on reading comprehension performance. Inform Syst Res 1992;3(1):17–35.
[36] Erkan G, Radev D. Lexrank: graph-based lexical centrality as salience in text summarization. J Art Intell Res 2004;22:457–79.
[37] Kincaid J, Fishburne R, Chissom Rl. Derivation of new readability formulas (automated readability index fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report; 1975. p. 8–75.
[38] Gunning R. The technique of clear writing. New York: McGraw-Hill; 1952.
[39] Nandhini K, Balasundaram S. Significance of learner dependent features for improving text readability using extractive summarization. In: Intelligent human computer interaction (IHCI); 2012. p. 1–5.
[40] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inform Process Manage 1988;24(5):513–23.
[41] Kano A. Evaluating phrase sets for use with text entry method evaluation with dyslexic participants. In: Workshop on improving and assessing pen-based input techniques at british HCI; 2005.
[42] http://www.samacheerkalvi.in.
[43] http://www.impact-information.com/scales.pdf.
[44] Dale E, Chall JS. A formula for predicting readability. Educational research bulletin; 1948. p. 11–28.
[45] Hirschman L, Light M, Breck E, Burger J. Deep read: a reading comprehension system. In: Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics, Association for computational linguistics; 1999. p. 325–32.