A Multi Document Summarization of Learning Materials using Bigram Embedding Technique and Integer Linear Programming

Sakkaravarthy Iyyappan K[1]*, Balasundaram SR[2]

Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamilnadu, India. kschakra@gmail.com

Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamilnadu, India. bblsundar@nitt.edu

Abstract: In the present era of the Internet, teachers and learners are heavily inclined to use e-learning systems for an efficient learning process. Due to the proliferation of educational text contents in these e-learning systems, the need for incorporating advanced text analysis tools and techniques are becoming inevitable. Multi Document Summarization (MDS) is a technique for producing concise summaries from a collection of related text documents. The usage of MDS in the context of e-learning is more promising for providing summaries for learning materials which helps students and teachers to focus on key concepts of the learning materials. In this paper a semantic approach towards the learning material summarization is proposed based on bigram embedding and ILP technique. This approach considers bigram as the basic meaningful semantic unit of the sentences to understand and summarize documents. Embedding techniques are employed to learn the vector representation of phrases to semantically identify similar phrases to reduce the redundancy and improving coherence. Using ILP technique, the summaries were generated by selecting important sentences while reducing the redundancy using phrase vectors. Experimental results on newly created educational dataset (EduSumm) shows better performance compared to the baseline systems.

Keywords: E- learning, multi document summarization, bigram embedding, ILP technique

1. INTRODUCTION

In the present age of digital information, there is an overwhelming amount of text, and it's growing at an unprecedented rate. Extracting crucial facts and insights from this vast pool of text poses a significant challenge. The process of condensing lengthy texts into brief yet meaningful summaries has the potential to bring about significant cost and time savings in industries spanning news, finance, and healthcare. However, manually crafting summaries from the original text is a laborious and costly endeavor. Automatic summarization addresses this issue by generating summaries from input text, all while maintaining the integrity of the original content. Automatic Text Summarization (ATS) aims to extract relevant information from one or more input documents and present it as a summary. The key objectives of an effective summarization system are salience and coverage (non-redundancy) [1]. Salience is achieved by evaluating the significance of sentences or their sub elements like words, bigrams, or phrases. Coverage, or the avoidance of redundancy, is crucial in summary creation as it minimizes repetitive information, thereby enhancing the overall informational content. Multi Document Summarization (MDS) is an extension of Single Document Summarization (SDS) in which information from multiple documents is joined and then summarized, presenting additional challenges. Due to the lexical diversity of source documents, the same semantic concept may be expressed through different lexical items, leading to a higher potential for redundancy in the summary [2].

In recent years, the advent of e-learning technologies has revolutionized the methods of teaching and learning. Educators and students can now engage and exchange learning materials through web or mobile platforms. To effectively manage this abundance of textual content, e-Learning systems should integrate advanced text analysis tools [3]. Notably, approximately 60% of students express a preference for summarized preview materials in addition to lecture slides, adding an extra responsibility for teachers to generate this supplementary material [4]. This work focuses on the automatic summarization of textual data, which represent the most widespread learning material. The proposed approach for extractive multi-document summarization in learning materials adopts a semantic framework, focusing on two crucial aspects: salience of information and reduced redundancy among summary sentences. The performance of the proposed approach was assessed using a manually annotated Education text dataset (EduSumm), employing the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric. Various ROUGE scores, including ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4, were computed to compare the system-generated summaries with the reference summaries from baseline systems and state-of-the-art methods existing in the literature. This analysis aimed to evaluate the effectiveness of the proposed approach and measure its improvement over existing techniques.

2. RELATED WORK

In the text summarization domain, lot of extractive summarization methods [5, 6, 7] have been developed using number of statistical, machine learning, graph based and optimization techniques. Most of the extractive approaches rank the sentences on the basis of importance score derived from the features such as term frequency, topic signature, cue words and sentence position [8]. Various supervised [9] and unsupervised machine learning methods [10] were used to identify the important sentences and extract them as summaries. Well known graph-based approaches such as LexRank [11] and Text Rank [12] compute sentence importance by modelling the relationships among the sentences as a graph in the document summarization process. But in general, most of the graph-based summarization approaches do not consider the semantic relatedness in the text . Integer linear programming (ILP) is one of the important optimization techniques used for summary generation by selecting sentences using objective function of maximizing importance, minimizing redundancy with set of constraints like summary length. McDonald [13] proposed an ILP formulation for summary generation by optimizing importance and redundancy whereas later the authors [14] added machine learning algorithm to infer the importance of the sentences. Gillick [15] performs sentence selection by considering bigrams as the concepts and optimizes weighted concepts using ILP without explicitly considering the redundancy. Yu [16] extended the ILP model to use phrases as concepts instead of bigrams to summarize product reviews. Luo and Litman [17] used the bigram-based approach to summarize students' feedback with matrix factorization method to semantically identifying same bigrams expressed as different lexical items. Later they focused on extracting phrases and summarizing students' feedback as a phrase summarization approach using clustering algorithm. Some approaches preferred to use phrase as basic processing unit for processing compared with words and sentences.

Attempts to apply text summarization for the learning materials could be seen in the literature. One such earlier approach is to enhance E-Learning platform to answer students' questions without using any knowledge base [18]. Shimada [3] summarized the contents of Lecture slides and used that as a preview material to enhance the understanding of the contents. Personalized summaries were generated and given to students based on their performance in the subjects after the lectures. Also [19] considered providing personalized summaries in mobile environment considering the screen size of the device as a factor. Extractive summarization techniques are used to generate personalized summaries which helps mobile learning based on the interests of the learners. Some works like [20] involves summarization in the multimedia contents of the e-learning systems with transferability between the videos and their corresponding text.

Many current summarization methods for learning materials lack a semantic analysis of the text content. This study introduces a novel approach that addresses redundancy and coherence concerns by employing a semantic-based method, using bigrams as the fundamental units. The research focuses on integrating a bigram embedding-based semantic approach for multi-document summarization within an e-learning setting.

3. PROPOSED METHODOLOGY

In the realm of text analytics, an n-gram refers to a sequence of consecutive n words. A single word is termed a unigram, two consecutive words form a bigram, and so forth. Many concept-based ILP (Integer Linear Programming) summarization systems adopt bigrams as proxies for concepts due to their simplicity and performance. But one major drawback of existing concept-based systems is lacking semantic checking ability to identify the similar bigrams. This approach (fig 1) uses embedding techniques to identify the semantic similarity among the bigrams.

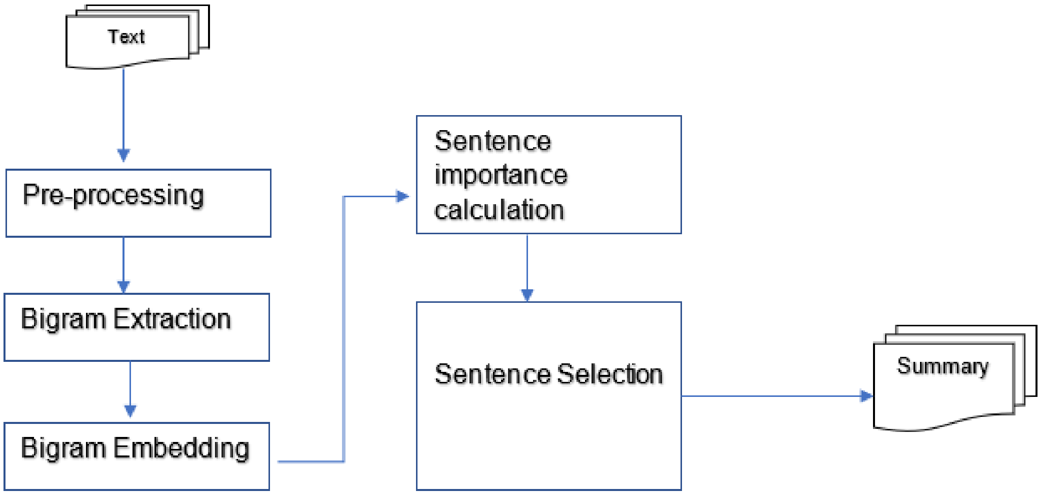

Figure 1: Proposed Model

Pre-processing:

Input documents are pre-processed. In this work the non-textual contents such as images, references, videos, audios are disregarded and textual data is focused

Bigram extraction:

While various methods have been previously suggested for bigram extraction, they generally align with either data-driven or linguistic approaches. However, it is out of the scope to study and compare them in this work. Instead, a straightforward and effective data-driven technique [21] is employed here for bigram extraction. This technique entails the formation of bigrams based on the scores assigned to n-grams, with the highest-scoring n-grams being chosen to construct the phrases. The score of the n-grams is computed using Equation 1.

$$Score(w_i, w_j) = \frac{Count(w_i, w_j) - \delta}{Count(w_i) * Count(w_j)}$$

Here wi and wj are word pairs captured from corpus. Count (wi, wj) is the co-occurrence frequency of wi and wj. count (wi) is the single word frequency of wi, and count (wj) is the word frequency of wj. δ is used as a threshold that filters low occurring bigrams. The n-grams with score above the chosen limit are then used as phrases. The phrases are identified, those word unigrams are tagged and combined with an underscore sign (_) in between them. For example, the words “classification” and “algorithm” will be combined as “classification_algorithm” and tagged as a phrase.

Bigram Embedding

In document analysis, the foundational element is the word and there has been considerable focus on transforming text representations into word embeddings, a process that entails mapping individual words to continuous vector spaces. This technique holds immense significance in Natural Language Processing (NLP) and encompasses various methods, including network model training and co-occurrence matrix reduction. A notable example is word2vec [21], a neural architecture-based model designed for learning distributional representations of words from text data. While both Skip-Gram and Continuous Bag of Words (CBOW) implementations are widely used, embedding techniques have been extended to encompass phrases, sentences, and even paragraphs. One common approach involves adding the vectors of individual words to produce bigram vectors. However, this simplistic addition may not capture the precise meaning of bigrams, potentially resulting in a loss of their accurate representation in the vector space. For example, simply adding the word vectors for "motor" and "cycle" does not effectively represent the bigram "motor cycle". Although bigrams in the context of NLP applications often convey the semantics of a sentence, there has been limited exploration of bigram embedding. In this approach, these issues are addressed by training an embedding model that directly treats a bigram as a single unit, thereby preserving its meaning in the vector space. Following the extraction of phrases from the text, a neural network-based approach is employed to acquire vector representations for these bigrams. This characteristic enables bigrams with similar meaning in the corpus to cluster together in the vector space, allowing for similarity computations through vector arithmetic. While the process of learning embeddings can take two specific forms, namely Skip-Gram and Continuous Bag of Words (CBOW), this work adopts the Skip-Gram model due to its superior performance compared to the CBOW model.

Skip Gram Model: The goal of Skip-Gram (fig 2) model training is to learn word representations that best predict the existence of the contextual words (surrounding words).

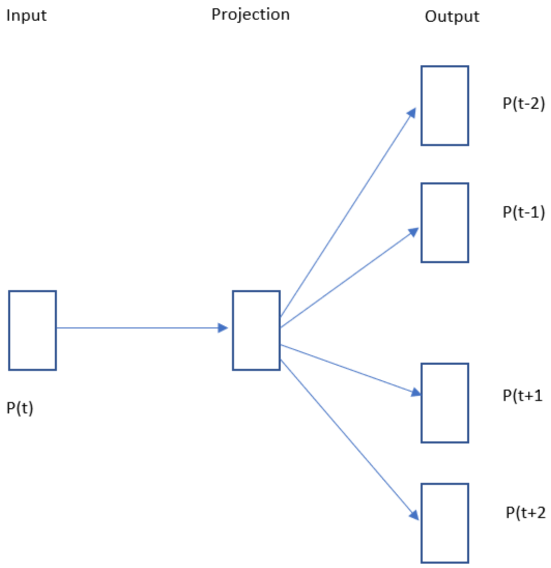

Figure 2: Skip Gram Model

In this case learning bigram representation is done using contextual bigrams(surrounding bigrams). Formally, given a training corpus of vocabulary T with bigrams p1,p2,p3….pT, the objective of the Skip-Gram is to maximize the average log probability:

$$\frac{1}{T}\sum_{i=1}^{T} \quad \sum_{-n<j<n,j\neq 0} \quad p\left(p_{i+j} \mid p_i\right) \tag{2}$$

Where n is the context window size, p_i is the target bigram, and p_{i+j} is the surrounding context bigram. The softmax function is used to estimate the probability $p\left(p_{i+j} \mid p_i\right)$ as given below.

$$p\left(p_{i+j} \mid p_i\right) = \frac{exp\ (V_{p_{i+j}}^{T} V_{p_i})}{\sum_{p=1}^{P} \quad exp\ (V_{p}^{T} V_{p_i})} \tag{3}$$

Where $V_{p_{i+j}}^{T}$ and $V_{p_i}$are vectors representations of target bigram and context bigram and P is the vocabulary size.

Bigram Features

In the task of summarization, a critical aspect is discerning the importance of sentences. In this study, the significance of sentences is determined by assessing the importance of the bigram contained within them. The literature offers a range of statistical and heuristic methods for weighting sentences. Some approaches concentrate on the constituent parts of sentences, for instance, prioritizing 'weighting bigrams' where a bigram is treated as a concept. Despite its simplicity, Document Frequency (DF) stands out as one of the most effective methods for gauging the importance of bigrams.

This study employs an unsupervised bigram weighting model, originally introduced by [13], to assign weights to phrases. This model combines factors such as document frequency, sentence position within documents, and the occurrence of phrases in document titles. The underlying hypothesis is that a phrase appearing in the majority of documents within the set, occurring initially at the beginning of documents, and being associated with document titles, is deemed significant. In cases where a title is not explicitly present, the first sentence of the document is considered as the title.

Summary Generation:

The fundamental idea behind this work is that bigram vectors effectively capture the semantic relationships within sentences and documents. The importance of sentences, as well as the similarity between them, is evaluated based on the presence of specific bigram within the sentences. This approach enables a nuanced understanding of both the individual sentences and their interconnections.

Sentence selection by MMR:

MMR (Maximal Marginal Relevance) employs a greedy algorithm designed to balance relevance and redundancy, utilizing similarity metrics. In many existing MMR approaches, sentences are treated as the fundamental units, and their importance is determined accordingly. Similarity is gauged with respect to both the query and other sentences, employing various subunits of sentences (such as words) through a range of similarity measures. A sentence is deemed highly significant if it incorporates more crucial bigrams. Additionally, two sentences are considered similar if the bigrams within them exhibit similarity. This similarity among bigrams is assessed using cosine similarity metrics applied to the bigram vectors.

Sentence selection by ILP:

Integer Linear Programming (ILP) is an optimization technique used to find the exact solution for the given objective with set of constraints. ILP method defined in this work focuses to select sentences that maximize the importance score of the summary while reducing the redundancy among the sentences. The objective function and constraints for the ILP formulation are given below.

$$(x,y) \quad \sum_{i=1}^{n} \quad imp(\,s_i).x_i - \sum_{i=1}^{n} \quad \sum_{j=i+1}^{n} \quad sim(s_i\,,s_j).\; y_{i,j} \tag{10}$$

Subject to:

$$\sum_{i=1}^{n} \quad l_i\,.x_i \leq L \tag{10a}$$

and for all i,j :

$$y_{i,j} - x_i \leq 0 \tag{10b}$$

$$y_{i,j} - x_j \leq 0 \tag{10c}$$

$$x_i + x_j - y_{i,j} \leq 1 \tag{10d}$$

Here n is the total number of the sentences in the source documents and $imp(\,s_i)$ is the importance score of the sentence s_i. The similarity between the sentences $s_i$and s_j is expressed as $sim(s_i\,,s_j)$ $ and the maximum length of the summary is L.

Binary variable x_i denotes that the sentences_i is included and binary variable $y_{i,j}$ denotes that both sentence s_i and s_j are included in the summary. Constraint 10a verifies that the length of the summary does not exceed the maximum summary limit. Last three constraints ensure that the values of x_i, x_j and $y_{i,j}$ are consistent that is if $y_{i,j}$is 1 then x_i=1 and x_j=1 (or if $y_{i,j}$=0 then x_i and x_j must be 0).

Dataset & Evaluation Metric:

This work primarily focuses on educational text data. The proposed system's performance is assessed using a manually annotated dataset specifically tailored for summarizing learning materials. This dataset is created by four subject experts, including the authors, who annotated text documents in four major engineering subjects: operating systems, software engineering, computer architecture, and artificial intelligence. This dataset serves as a valuable resource for evaluating the performance of our proposed system in the context of educational material summarization.

The summaries generated by the proposed system were assessed using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric [22]. This metric computes the n-gram overlap between the system-generated summary and manually crafted reference summaries. Specifically, ROUGE-1, ROUGE-2, and ROUGE-4 are employed in this study for evaluating the summaries, as these measures exhibit a stronger correlation with human assessments.

Table 2: Basic statistics about EduSumm dataset.

Number of Clusters	
Number of Documents	500
Total No.of Sentences	13,135
Total No.of Words	336,073
Total No.of Phrases	273,316

4. RESULTS & DISCUSSION

The baseline methods used in this work are Textrank,Lexrank and ICSI. The Rouge scores for the educational dataset is reported in Table 2. It demonstrates that our models perform better with the existing baseline methods. The MMR and ILP produces better results compared to all other mentioned works in which ILP model slightly performs well. Proposed ILP outperforms existing approaches by 8% in R1 score, 14% in R2 scores and 13% in R4 scores.

Table 2: ROUGE comparison on EduSumm dataset.

Systems	**R1**	**R2**	**R4**
Textrank	41.01	9.11	1.24
Lexrank	40.34	9.01	1.01
ICSI	40.4	9.01	1.2
MMR	42.72	9.37	1.30
ILP	**43.11**	**9.91**	**1.38**

User evaluation on EduSumm for usability:

In order to evaluate the usefulness of learning material summaries, 25 students were asked to assign a score from 1 to 5 for the summaries to answer the question "How much the given summary is useful to grasp the main contents?" The average of the scores is reported in the Table 3. In order to measure the agreement among the human evaluators in ranking the three system Kendall concordance coefficient (W) is used. The correlation among the human experts were relatively good with the value W=7.2. It shows that the W is significant at the 95% level by applying the Chi-Square test. These statistical observations support that the ranking provided by the experts are reliable.

Table 3: User Evaluation on DUC and Edusumm

Systems	**Score on EduSumm Dataset (higher is better)**
Textrank	2.1
MMR	3.1
ILP	**4.2**

5. CONCLUSION

The presented approach introduces a multi-document summarization method for learning materials, employing bigram embeddings as the fundamental processing units. Leveraging a distributional semantic model, this technique prioritizes semantic analysis to reduce redundancy and enhance coverage in summary generation. The experimental results show that the proposed method produces better performance on newly created educational EduSumm dataset regarding the ROUGE metrics. Human expert's evaluation shows that the proposed approach is better in terms of coherence also. To demonstrate the usability of the proposed in e-learning environment a student evaluation for the generated summaries was conducted.

6. REFERENCES

1. Christensen J, Soderland S, Etzioni O. Towards coherent multi-document summarization. In: Proc. of the 2013 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Atlanta; 2013. p. 1163-1173.
2. Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. Artif. Intell. Rev. 2017;47:1-66.
3. Baralis E, Cagliero L, Farinetti L. Generation and Evaluation of Summaries of Academic Teaching Material. In: Proc. of the IEEE 39th Annual Computer Software and Applications Conference; 2015. p. 881-886.
4. Shimada A, Okubo F, Yin C, Ogata H. Automatic Summarization of Lecture Slides for Enhanced Student Preview. IEEE Trans. Learn. Technol. 2018;11:165-178.
5. Nasr Azadani M, Ghadiri N, Davoodijam E. Graph-based biomedical text summarization: An itemset mining and sentence clustering approach. J. Biomed. Inform. 2018;84:42-58.
6. Ouyang Y, Li W, Li S, Lu Q. Applying regression models to query-focused multi-document summarization. Inf. Process. Manage. 2011;47:227-237.
7. Tabak FS, Evrim V. Event-based summarization of news articles. Turk. J. Elect. Eng. Comp. Sci. 2020;28:850-864.
8. Mei JP, Chen L. SumCR: a new subtopic-based extractive approach for text summarization. Knowl. Inf. Syst. 2012;31:527-545.
9. Shen D, Sun JT, Li H, Yang Q, Chen Z. Document summarization using conditional random fields. In: Proc. of the 20th International Joint Conference on Artif. Intell. 2007. p. 2862-2867.
10. Nomoto T, Matsumoto Y. A new approach to unsupervised text summarization. In: Proc. of the 24th Annual Int. ACM SIGIR Conference on Res. Dev. in Inf. Retrieval. New York, NY, USA; 2004. p. 26-34.
11. Erkan G, Radev DR. Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. 2004;22:457-479.
12. Mihalcea R, Tarau P. TextRank: Bringing order into texts. In: Proc. of the Conference on Empirical Methods in Nat. Lang. Proc. Barcelona, Spain; 2004. p. 8-15.
13. McDonald R. A study of global inference algorithms in multi-document summarization. In: Proc. of 29th European conf. on IR research. Rome, Italy; 2007. p. 557-564.
14. Galanis D, Lampouras G, Androutsopoulos I. Extractive multi-document summarization with integer linear programming and support vector regression. In: Proc. of COLING. 2012; p. 911-926.
15. Gillick D, Favre B, Tür DH, Bohnet B, Liu Y, et al. The icsi/utd summarization system at TAC 2009. In: Proc. of the Text Analysis Conf. Workshop. Gaithersburg, MD USA; 2009. p. 1-20.
16. Yu N, Huang M, Shi Y, Zhu X. Product review summarization by exploiting phrase properties. In: Proc. of COLING 2016, the 26th Int. Conf. on Comput. Linguistics. 2016; p. 1113-1124.
17. Luo W, Litman D. Summarizing student responses to reflection prompts. In: Proc. of the 2015 Conf. on Empirical Methods in Nat. Lang. Proc. 2015; p. 1955-1960.
18. Saraswathi S, Hemamalini M, Janani S, Priyadharshini V. Multidocument text summarization in e-learning system for operating system domain. In: Adv. in Comput. and Commun. Springer Berlin Heidelberg; 2011. p. 175-186.
19. Yang G, Chen NS, Kinshuk, Sutinen E, Anderson T, Wen D. The effectiveness of automatic text summarization in mobile learning contexts. Comput. & Educ. 2011;68:233-243.
20. Chang W, Yang J, and Wu Y. A keyword-based video summarization learning platform with multimodal surrogates. In: Proc. of the 11th IEEE Int. Conf. on Adv. Learn. Tech. 2011; p. 37-41.
21. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the 26th Int. Conf. on Neural Inf. Process. Syst. 2013; p. 3111-3119.
22. Lin CY. ROUGE: A Package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proc. of the ACL-04 Workshop. Barcelona, Spain; 2004. p. 74-81.

DOI: https://doi.org/10.15379/ijmst.v10i2.3149