2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579748

Bert-based Braille Summarization of Long Documents

Yamuna K.*, Shriamrut V.†, Drishti Singh‡, Vaishnavi Gopalasamy§ and Vivek Menon¶
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri
Kollam 690 525, India
Email: *yamunak1709@gmail.com, †shriamrutvenkatesh@gmail.com, ‡singhdrishti1001@gmail.com,
§vaishnavigopalasamy@gmail.com, ¶vivekmenon@am.amrita.edu

***Abstract*—Availability of accurate document summaries is of paramount importance in our lives, though there is a dearth of such options for the visually challenged people. With a large volume of existing text summarization work dealing with short documents, this work aims to tackle the problem of large document summarization in braille. We propose a novel domain and genre agnostic approach to document summarization in braille using a BERT extractive summarizer. Experimental evaluation of our approach involved a comparison of the generated and reference summaries using ROUGE metrics. Given the lack of robust convertors from English to Braille, another key contribution of our work is the development of a novel English & Braille interconversion library to facilitate further research in this domain.**

***Index Terms*—Braille Summarization, Text Summarization, BERT, Long Documents, NLP**

I. Introduction

In today's world, time is a valuable investment. In our busy lives, most of us hardly have time to read a complete book, let alone spending time on finding one that interests us. Books are packed with insights into knowledge and life, blurring the boundaries of age restrictions and encouraging us in the hardest of times. The first rule of choosing books is to read what you like! With so many options, any reader can find a book that suits their preferences. Usually, a reader picks a book of their preferred genre by reading the inside flap. If the plot description seems interesting, then that's the book. The situation is not the same with people who are visually impaired or have lost their ability to see. They have to read braille with their fingertips or special equipment, which complicates the process a bit more.

Braille is a unique system of reading and writing developed for the visually impaired, with a set of symbols composed of small rectangular braille cells that contain tiny three-dimensional tactile bumps called raised dots traditionally written with embossed paper as thick as a file folder or index card. Technological advancements let all of us walk together regardless of the inabilities people face; for example, one being braille e-readers which highly stress upon the improvement of cognitive skills, and boosting literacy and numeracy of the visually impaired community [1]. Speech synthesizers are also commonly used for the task, but the method of reading through braille remains popular in the daily lives of visually impaired people, especially the deaf-blind [2].

Despite witnessing a steadily increasing number of books becoming available electronically as soft copies, the number of books in braille, available for specially-abled people is relatively less, and it is necessary to first translate the available English books into braille. Only then can the visually challenged community read the books with their fingertips or electronic equipment. However, manual translation and summarization are time-consuming and prone to errors. Hence, the requirement for different language processing techniques that are able to handle larger documents such as books is becoming increasingly more significant.

As we know, document summarization is a very useful means for people to quickly read and browse for books of their interest [3]. Single-document summarization is the process of automatically creates a shorter version of a document automatically. This task has received a lot of interest in the Natural Language Processing (NLP) community due to its potential for various applications for information access. Existing summarization systems are primarily concerned with the content quality and fluency of summaries, and they typically extract informative and diverse sentences in the input text to form a summary of a specific length or construct an abstractive summary by rewording and rebuilding sentences. Despite abstractive summarization being highly desirable for multiple reasons, as well as being the focus of numerous research papers, this method is difficult to automatically generate, either demanding many Graphics Processing Units(GPUs) that need to be trained over time via deep learning approaches or complex algorithms and rules with a narrow range of applicability for conventional NLP approaches [4] [5]. Further, it necessitates a more in-depth examination of the text specific to the domain. With this obvious hurdle in mind, this paper employs extractive summarization.

Although many neural models have been proposed for extractive summarization recently, a staggering amount of implemented solutions utilize outdated natural language processing algorithms requiring regular maintenance due to poor generalization. As a result, many of the summary outputs obtained from the mentioned tools may appear irregular in their creation of content. Currently, a few attempts have been made for end-to-end training of documents for summarization tasks but are yet to achieve a significant improvement in performance. Moreover, the summaries are usually produced

Authorized licensed use limited to: UNIVERSITY OF ESSEX. Downloaded on July 10,2024 at 18:43:44 UTC from IEEE Xplore. Restrictions apply.

for sighted people, but not for visually impaired people. A text summary can be translated into a braille summary for the reading of the visually impaired, and the length of a braille summary is defined as the number of the braille cells in the summary. Here, automatic translation plays a major role, as manual translation is a tedious and time consuming task.

In this work, we address the difficulty of long document summarization in braille. While there exists a considerable amount of research in the field of text summarization, most of it has focused on the summarization of short documents, with a particular focus on news articles. Research pertaining to long documents, such as books is very minimal. Further, books vary in length as well as in genre, necessitating the use of various summarization methods.

In this paper, we elucidate the details of a new benchmark by compiling a dataset consisting of books with human-generated summaries obtained online used as a baseline, designed specifically for the generation of braille summaries. Due to the need for better encoding and various levels of attention on both words and sentences along with potential external memory units for storing farther but rather more significant information, we have used the bert-extractive-summarizer [6], a simple variant of BERT as an encoder for extractive summarization. Further, we have developed a novel, English & Braille interconversion library to generate braille summaries.

II. Related Work

Automatic summarization has received significant attention from the natural language processing community, ever since the initial approaches to automatic abstraction that laid the groundwork for today's text summarization techniques [7] [8] [9]. The extraction process, which is possibly the most important aspect of an effective summarization algorithm, has received the most attention to date, and it is also the subject of our current work.

In the summarization literature, there are two key trends: supervised machine learning algorithms that are trained using pre-existing document-summary pairs, and unsupervised approaches which rely on the properties of the text and heuristics.

Among the unsupervised approaches, classic summarization methods account for both frequency and centrality. Frequency-driven approaches are based on the notion that more important information will appear in documents more frequently than less important, comprehensive explanations. For example, the SUMBASIC system [10] was motivated by word likelihood estimation, which assigned each sentence a weight equal to the average probability of the sentence's subject terms. Sentences that are more analogous to other sentences are regarded as central, since they are assumed to carry the original documents' most central ideas. This assumption is the foundation of graph-based summarization frameworks. A graph-based approach, where a graph correlating to each sentence in the document is generated and integrated to form a graph of the document based on semantic analysis, and summaries are generated by ranking sentences using a Page Rank Algorithm is discussed in [11]. A centroid-based summarization approach, with a pseudo sentence of the document called a centroid obtained by normalizing the document, consisting of words with tf-idf scores above a predefined threshold, with sentences later sorted in descending order to generate the summary is discussed in [12]. All of the above-mentioned approaches concentrate on extracting the most commonly recurring information from a document. However, extracting the most common or central parts of noisy documents containing a serious amount of unimportant, superfluous content may not be the most ideal approach.

In addition to unsupervised methods, various machine learning techniques have been developed for extractive summarization [13], [14]. These approaches rely on labeled training data and also require the creation of features that are informative enough to allow the learning algorithm to discriminate keyphrases from non-keyphrases.

A good summary should highlight the most significant information from the original document while simultaneously being cohesive, non-superfluous, and legible. Recently, the state-of-the-art BERTSUM [15], a simple variant of BERT as an encoder with its pre-training on a huge dataset, and the use of flat architecture with inter-sentence transform layers introduced an advanced approach for single document summarization outperforming the existing traditional approaches. A Python-based RESTful service using the BERT model for text embeddings and K-Means clustering for summary generation on the Lectures dataset is discussed in [6]. We can see that all the existing summarization frameworks are designed for sighted people, but not for individuals who are blind or visually disabled.

It has been a long road to develop methods to help blind and visually impaired people browse information as conveniently as ordinary people. Some of the existing work for blind and visually impaired people include Integer Linear Programming (ILP)-based summarization [16] to generate braille summary of news articles for visually impaired people. Moreover, few existing libraries supporting braille conversion are restricted to monolithic conversion and while dealing with complex text containing undefined characters runs into errors, prohibiting further conversion. After the popularity of braille, a lot of work focused on how to improve the accessibility of web information for blind people [1]. Unfortunately, there is still a long way to go in terms of adapting such architectures specifically pertaining to the automatic summarization of books. There are currently inadequate solutions for encoding books, which often comprise many paragraphs or a collection of interrelated documents. Furthermore, a significant and increasing body of work focuses on summarizing short documents specific to domains and genres, with evaluations typically focusing on news articles that exist.

Different from the existing research, our work is based on recent advances in terms of the use of the state-of-the-art BERT Extractive Summarizer, which we successfully extend to the braille-summarization of books for visually impaired people. Our work, albeit on a smaller scale, is an effort

Authorized licensed use limited to: UNIVERSITY OF ESSEX. Downloaded on July 10,2024 at 18:43:44 UTC from IEEE Xplore. Restrictions apply.

to model extractive summarization tasks for long documents irrespective of domain and genres, with no dependence on handcrafting features and effectively avoiding secondary or redundant information. Further, our work has also resulted in a novel English & Braille interconversion library which facilitates the conversion of English text to braille code and vice-versa. We also aim to address the data scarcity that is becoming increasingly apparent in areas other than news, as well as in languages other than English, such as braille, by building datasets of books that will largely enhance the research in this relatively unexplored field.

III. Methodology

The project architecture depicted in Fig. 1 starts from building a dataset consisting of English books and their summaries, obtained from standard websites that make human-generated summaries available online. The dataset thus obtained is analyzed using Exploratory Data Analysis (EDA). The EDA results are processed, studied, and interpreted to discover additional insights in the generated dataset of various lengths, domains, and genres. The explored data is then passed through a pre-trained Bert Extractive Summarizer with necessary parameters initialized, to generate summaries of the English books. This summary is then passed through a novel English & Braille interconverter designed and developed to generate a braille summary. To alleviate the insufficiency of data pertaining to availability of books in braille, some of the select English books are transformed into braille using the interconverter we had developed and thus increase the scope of research in this area. Finally, the summary generated is evaluated with the ideal summary using ROUGE metrics. In this research, we use the Google Colaboratory tool (Google Colab), with Python as a programming language.

A. Dataset

The initial challenge we faced when we first began to work on automatic book summarization for visually impaired people was the lack of an appropriate standard dataset, developed especially for the braille summarization of long documents. Currently, the existing standard datasets for document summarization tasks like CNN/Daily Mail, Gigaword, and others, benefit from the summarization of short documents. The shortage of such datasets is perhaps unsurprising since manual summarization of long documents is a tedious task compared to the summarization of short documents. Furthermore, books are often available in hard copy and are usually protected by copyright laws that prevent them from being converted to electronic format, limiting their use in the public domain for different language processing tasks. We generated a dataset based on the finding that some English literature courses involve books that are often accessible as abstracts — intended to make the content of the books more accessible to students. CliffsNotes is one of the major content providers that make summaries of series of guides in various domains available online. Fortunately, we were able to locate online digital

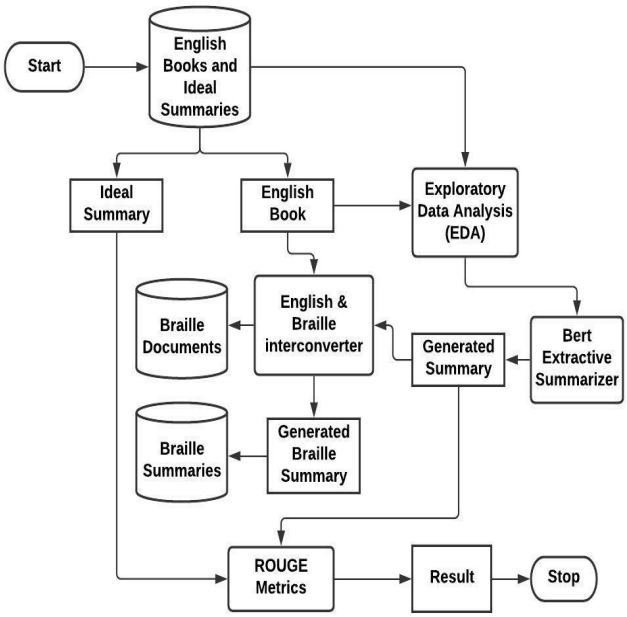

Fig. 1. Architecture of BERT-based Braille Summarization

versions of most of these books in the public domain sych as Gutenberg.

For example, the following is an excerpt from CliffsNotes summary of Pride and Prejudice by Jane Austen.

> *"When Charles Bingley, a rich single man, moves to the Netherfield estate, the neighborhood residents are thrilled, especially Mrs. Bennet, who hopes to marry one of her five daughters to him. When the Bennet daughters meet him at a local ball, they are impressed by his outgoing personality and friendly disposition. They are less impressed, however, by Bingley's friend Fitzwilliam Darcy, a landowning aristocrat who is....."*

We started with the books that had a CliffsNotes summary and eliminated all of the books where an electronic copy was unavailable. This gave us a 'gold standard' dataset of 20 books, each with manually created summaries.

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach where the researchers take a bird's eye view of the data and attempt to make some sense of it. It's usually the first step in data processing, carried out before any formal statistical techniques are applied. It aids in determining patterns in data collection using statistics and probability. Scatter plots, histograms, and other graphical techniques are often used to visualize EDA performance. EDA is commonly used in pre-processing to visualize, check for associations between data or variables, and

Authorized licensed use limited to: UNIVERSITY OF ESSEX. Downloaded on July 10,2024 at 18:43:44 UTC from IEEE Xplore. Restrictions apply.

also provide a summary of numerical data such as minimum value, maximum value, average, etc.

C. BERT-based Braille Summarization

The BERT-based braille summarization architecture has leveraged an existing bert-extractive-summarizer library [6] that generates summaries of the books in the dataset and then passes the generated summaries to the English & Braille interconverter library for obtaining the braille summaries.

1) Bert Extractive Summarizer for Summarization: BERT is a language model that learns the context between words and sentences in a given text and is regarded as one of the best language models. The training methods used in BERT are highly systematic and facilitate transfer learning. Transfer learning works on the concept of storing knowledge gained from one task, to solve another different task. For instance, knowing how to ride a bicycle and applying the knowledge gained to ride a motorcycle. Similarly, BERT learns the language of the text after pre-training, and by fine-tuning, it can be further used for solving various NLP-related problems [17]. Furthermore, BERT is advanced enough to extract meaning from all complexities of language, and hence, steps such as stop word elimination, stemming, and lower-case transformations are purposefully avoided. Due to its outperformance over other extractive summarization algorithms, we adopted the state-of-the-art, BERT extractive summarization framework [6] for addressing the new task of braille summarization. The BERT extractive summarizer makes use of a pipeline that tokenizes the input paragraphs' text into clean sentences, and passes them to the BERT model to produce embeddings that contain contextual information about the sentences, and further K-means clustering is performed on these embeddings and the sentences that are centroids of the cluster are then finalized for the summary along with considering the length of the summary to be generated.

2) English & Braille Interconversion library: After the successful generation of English summaries, they were further translated using the novel English & Braille interconverter, generated to convert text to 6-dot braille encodings(Grade 1). In Grade-1 braille, capital letters of the English alphabet and numbers are represented using two braille codes. Braille does not have a unique alphabet for upper case letters. A dot 6 is placed in front of the letter to be capitalized to indicate capitalization. Similarly, a special number sign (dots 3, 4, 5, and 6) placed before the first 10 alphabets a-j is used to create braille numerals. Braille symbols are available in the range U+2800 to U+28FF in the Unicode system. The library makes use of braille pattern representations in Unicode to create a mapping between English and braille characters. For the benefit of the research community, this library capable of converting English to braille and vice-versa is made available on Github. [18].

Since braille is the key means of literacy and communication for blind and visually challenged people irrespective of the age group, our framework also provides the conversion of braille back to English to facilitate the exchange of knowledge between sighted and visually challenged people.

Algorithm 1: English to Braille conversion

```
Result: Braille Patterns
braille_patterns := null;
english_words := tokenize (english_text);
for english_word in english_words do
    braille_word := empty string;
    for i:= 0 to english_word.length do
        braille_character :=
          get_braille_character(english_word[i]);
        braille_word := braille_word |braille_character;
    end
    braille_patterns := braille_patterns |braille_word;
end
```

Algorithm 2: Braille to English conversion

```
Result: English Text
english_text := null;
braille_words := tokenize (braille_patterns);
for braille_word in braille_words do
    english_word := empty string;
    for i:=0 to braille_word.length do
        if is_english_available(braille_word[i]) then
            english_char :=
              get_english_char(braille_word[i]);
            english_word := english_word
              |english_char;
        else if
          is_english_available(braille_word[i:i+2])
          then
            english_char :=
              get_english_char(braille_word[i:i+2]);
            english_word := english_word
              |english_char;
        else
            continue;
        end
        english_text := english_text |english_word;
    end
end
```

IV. Evaluation Metrics

An acceptable summary should be simple to read and provide a decent summary for the provided text content. Since manual evaluation for determining the quality of summaries of various documents is tedious and complex, a number of suggestions have been made to make most of the process less tedious. Recently, ROUGE (Recall-Oriented Understudy for Gisty Evaluation) metrics [19] are used in determining the standard of summaries. The ROUGE metrics evaluate the quality of the generated summary in comparison to manually

Authorized licensed use limited to: UNIVERSITY OF ESSEX. Downloaded on July 10,2024 at 18:43:44 UTC from IEEE Xplore. Restrictions apply.

written reference summaries. Various forms of ROUGE are available. We have used ROUGE-1, ROUGE-2, and ROUGE-L to evaluate our model as a means of assessing informativeness and calculated the domain specific measures - recall, precision, and F1-Score in order to acquire a good quantitative value.

$$Recall = \frac{Count\ of\ overlapping\ words}{Total\ words\ in\ generated\ summary} \quad (1)$$

$$Precision = \frac{Count\ of\ overlapping\ words}{Total\ words\ in\ reference\ summary} \quad (2)$$

$$F1\ Score = \frac{2\ *\ precision\ *\ recall}{precision\ +\ recall} \quad (3)$$

V. Results

Successful data collection was obtained for a total of 20 books. The dataset of books was collected based on variable length and genres for testing on a model which worked accurately well for different genres and long documents. The books in the given dataset have a mean length of 475258 words, including the summaries from Cliffnotes with a mean length of 7,476 words. Fig. 2 plots the summary length versus the length of the book dataset. Most books are of length between 1,08,049 to 8,93,246 words, and reference summaries are 2,837 to 15,236 words. For very long books, with more than 5,00,000 words, the summaries tend to become correspondingly longer.

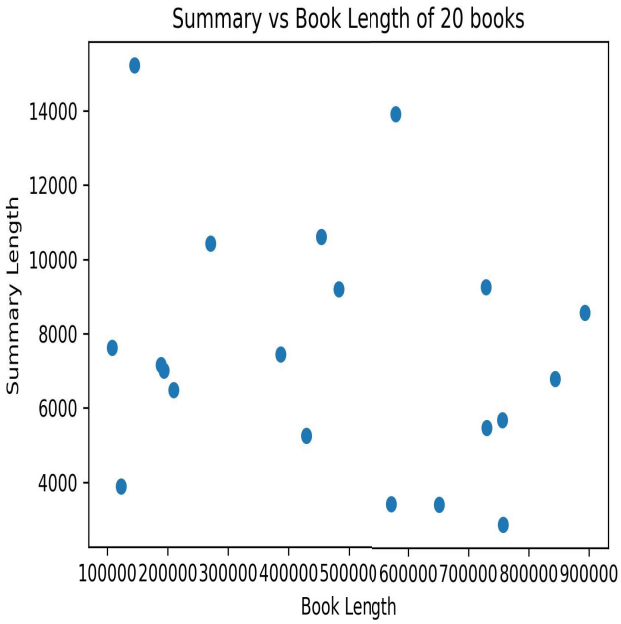

Fig. 2. Book length vs Summary Length of 20 books

The result of EDA can be used as a basis for a better understanding of the dataset used for braille summarization. Fig. 3 shows the word cloud formed using the frequency of the words obtained from the summary and Fig. 4 illustrates the top 10 frequent words in our English books dataset.

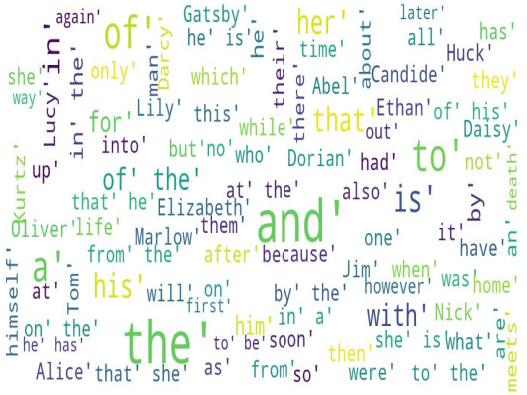

Fig. 3. Word cloud

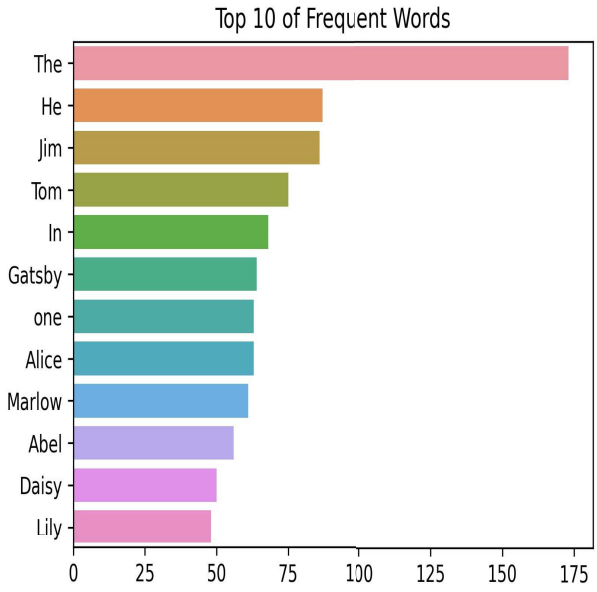

Fig. 4. Top 10 frequent words of the dataset

The Flesch Score measures the ease of readability of a given text. Fig. plots the Flesch reading score of all the books in the dataset generated. If the score is greater than 60, it is considered appropriate for all age groups above 11. As a result, children could share the pleasure of reading these books as well.

Table I illustrates the figures collected using ROUGE-1,

Authorized licensed use limited to: UNIVERSITY OF ESSEX. Downloaded on July 10,2024 at 18:43:44 UTC from IEEE Xplore. Restrictions apply.

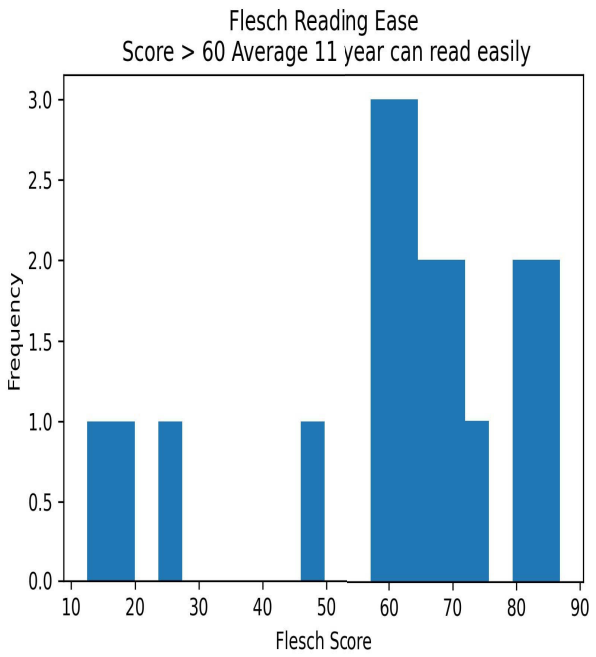

Fig. 5. Flesch Score

ROUGE-2, and ROUGE-L for our final summarizer, given the dataset generated.

TABLE I
SUMMARY EVALUATION

	ROUGE-1	ROUGE-2	ROUGE-L
Recall	0.485	0.092	0.173
Precision	0.363	0.069	0.129
F-measure	0.412	0.078	0.147

The comparison with existing benchmark datasets is of less relevance due to the large disparity between the type of data in the text summarised in this project as compared to that of the benchmark datasets.

VI. Conclusion

While there has been a significant amount of work done in the domain of text summarization, the majority of research is concerned with short document datasets such as news articles. In this paper, we have attempted to overcome the existing gaps by focusing on long text document summarization, irrespective of domain and genres, independent of handcrafted features and thereby avoiding secondary or redundant information. We believe this paper has made major contributions in this relatively unexplored arena, by making an initial progress in collecting large document book data for future research in this field, in addition to setting a new benchmark for long document summarization using our novel English & Braille interconversion library. In our summarizer, the ROUGE Measure is quite accurate for our dataset with relatively lower error rates against the existing benchmark implementations. Provided with a good number of books available in English texts and now in braille as well, we hope that the concept of automation in braille book summarization will play a vital role in enabling informed choices of reading amongst the visually challenged. This work could be extended to design an end-to-end model by directly generating braille summarization from braille books. We believe that our work will ease up and enhance research concerned with English as well as braille book summarization.

References

[1] "I'mnovation, "an e-reader for visually impaired," acciona." [Online]. Available: https://www.imnovation-hub.com/society/e-reader-visually-impaired-people/

[2] D. Freitas and G. Kouroupetroglou, "Speech technologies for blind and low vision persons," *Technology and Disability*, vol. 20, no. 2, pp. 135–156, 2008.

[3] N. Lalithamani, "Text summarization," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 3, pp. 1368–1372, 2018.

[4] P.-E. Genest and G. Lapalme, "Framework for abstractive summarization using text-to-text generation," in *Proceedings of the workshop on monolingual text-to-text generation*, 2011, pp. 64–73.

[5] G. Veena, D. Gupta, J. Jaganadh, and S. N. Sreekumar, "A graph based conceptual mining model for abstractive text summarization," *Indian Journal of Science and Technology*, vol. 9, no. S1, 2016.

[6] D. Miller, "Leveraging bert for extractive text summarization on lectures," *arXiv preprint arXiv:1906.04165*, 2019.

[7] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[8] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[9] S. S. Rani, K. Sreejith, and A. Sanker, "A hybrid approach for automatic document summarization," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 663–669.

[10] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," *Information Processing & Management*, vol. 43, no. 6, pp. 1606–1618, 2007.

[11] J. Jagadeesh and V. Varma, "Single document summarization using natural language processing," in *Proceedings of the 2nd Indian International Conference on Artificial Intelligence, Pune, India, December 20-22, 2005*, B. Prasad, Ed. IICAI, 2005, pp. 741–748.

[12] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (tf-idf)," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 4, pp. 285–294, 2016.

[13] E. D'Avanzo and B. Magnini, "A keyphrase-based approach to summarization: the lake system at duc-2005," in *Proceedings of DUC*, 2005.

[14] K. Shivakumar and R. Soumya, "Text summarization using clustering technique and svm technique," *International Journal of Applied Engineering Research*, vol. 10, no. 12, pp. 28 873–28 881, 2015.

[15] Y. Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.

[16] X. Wan and Y. Hu, "Braillesum: A news summarization system for the blind and visually impaired people," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 578–582.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] V. Shriamrut, K. Yamuna, D. Singh, and V. Gopalasamy, "English and braille interconverter," https://github.com/shriamrut/english-braille-interconverter, 2021.

[19] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 150–157.

Authorized licensed use limited to: UNIVERSITY OF ESSEX. Downloaded on July 10,2024 at 18:43:44 UTC from IEEE Xplore. Restrictions apply.