Deep Learning for Text Summarization

Ritwik Mishra, Dimpy Varshni, Aamir T. Ahmad
ritwikm@iiitd.ac.in,dimpy19022@iiitd.ac.in,aamir16001@iiitd.ac.in
IIIT-Delhi

1 INTRODUCTION

The process of reading an entire news article adds a cognitive load on a human mind. Tools for automated summarization of news article reduces the reading efforts. There are many state-of-art methods for text summarization in English language [1–3] but not much research has been done in summarizing the Hindi language, which is 4th most common natural language to be spoken in the world. Therefore, the intention is to develop a text summarization model for the Hindi language using deep learning techniques.

2 RELATED WORK

In literature, text summarization in Hindi has been done with methods like rule-based [4], topic modeling [5], and feature-based SVM classifier [6]. Deep learning techniques are not prevalent in the works related to Indic languages. We have found no regularity in the dataset and the evaluation metrics used in the previous works. To the best of our knowledge, no work has been done on the dataset that we have chosen for this project.

3 DATASET AND EVALUATION

Hindi Text Short and Large Summarization Corpus shall be used for training purposes. The dataset has been developed by Gaurav Arora [8]. The distribution of tokens in the dataset has been specified in Figures 1 and 2 in the appendix. Overall, the dataset has summaries having token length between 11 to 61, where the mode is near 27. And article text has token length 43-12k, where the mode is near 300 tokens. From a total of nearly 60k news articles, 60 percent are used for training, 20 percent for validation, and 20 percent for testing. In the extractive summarization part, the gold abstractive summaries are converted to extracts from the news article using lemmatization and word matching. We call this summary as silver extractive summaries. The silver summaries were majorly representative of the gold summaries as shown in figure 3. For evaluation of the developed models, the ROUGE score is used to evaluate different summaries.

4 METHODOLOGY

Here, we try to solve two main tasks - extractive and abstractive summarization for Hindi language articles. For abstractive, we have tried two different models - one is an attention-based seq2seq model and the other one is based on a pre-trained BERT encoder. For extractive, we have tried 4 different models explained in the subsection below.

4.1 Abstractive Summarization

Initially, an attention-based seq2seq model with teacher forcing (att-seq2seq-tf) is used to generate text for abstractive summaries. It was observed that the fully connected layer for vocabulary was learning nothing and kept on producing stopwords in the output. Gradient propagation is shown in figure 6. Further layers couldn't be added because of GPU memory constraints. The second model is an Encoder-Decoder architecture where Encoder has fine-tuned BERT [7]based embeddings for token indices and decoder is a linear layer outputting the probability distribution of every token in the vocab. The encoder part of the model is BERT fine-tuned on the original dataset only for articles having sequence length less than 512 (for faster computations) resulting in a total of 35470 train articles and 8784 test articles. The model consists of 12 layers comprising 110M parameters pre-trained on 104 languages including Hindi. The decoder part is just a linear layer that is responsible to learn the probability distributions of each token at the output end for the input tokens. Random sampling according to the probabilities obtained at the output is done for the resultant output word. The model is trained using 3e-05 learning rate for 15 epochs with batch size 2. To show a generalization of this model, a subset of 2500 articles (2000 train and 500 test) was randomly taken from the original dataset due to limited resources.

4.2 Extractive Summarization

Baseline : TextRank was implemented as the baseline. We achieved a Bleu score of 0.45 and rouge scores as indicated in the Results table. After that we have implemented the BERT model explained below. For extractive summarization, we have used the method proposed by [3]. BERT model is used for encoding the article sentences into sentence vector representations. For the summarization model, we tried two architectures namely simple classifier and RNN based classifier. For simple classifier, we have trained the model on the For RNN based classifier, we used an LSTM layer and linear layers with sigmoid in the end to predict the occurrence probability of each sentence of the article in the summary. The dropout probability of 0.2 is chosen between the linear layers and the learning rate is kept at 1e-4. At first, the model is trained with frozen BERT weights for 5

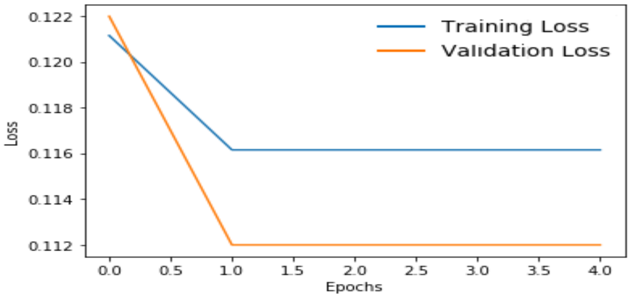


**Figure 1: =
Loss curved for BERTSUM-RNN when BERT weights were frozen.**

epochs. Then the model is retrained with BERT weights unfrozen for 5 more epochs. Loss curves for frozen and unfrozen BERT weights are shown in fig 2 and fig 3 respectively.

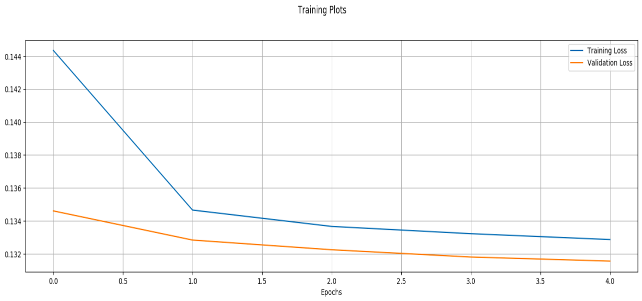


Figure 2: =
Loss curved for BERTSUM-RNN when BERT weights were frozen.

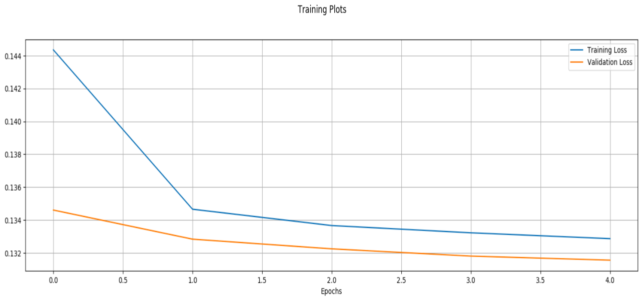


Figure 3: =
Loss curves for BERTSUM-RNN when BERT weights were unfrozen.

The number of epochs was kept low because of time constraints. An interesting observation is seen in figure 4a where training loss is greater than validation loss. One explanation for that could be the dropout layer deactivation during the validation set forward phase. As seen in figure 4b, the model was converging to optimum had it been given more time to converge.

All the models used for both the summarizations are trained using Adam optimizer.

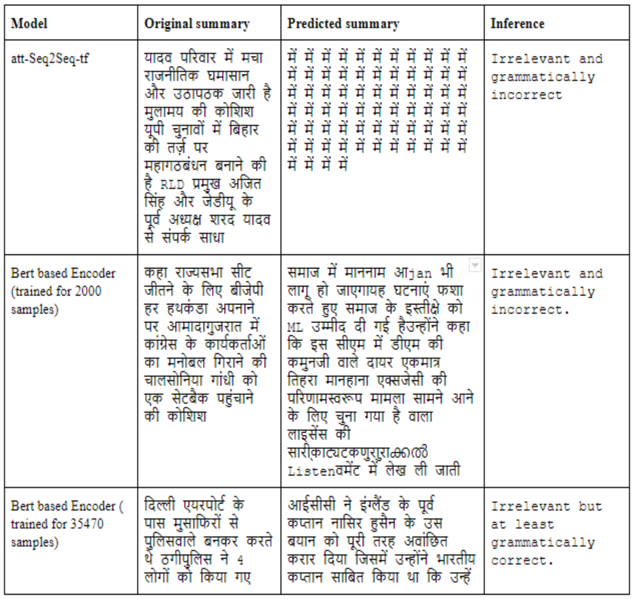
5 RESULTS AND ANALYSIS

This section summarizes the results of the two tasks -

5.1 Abstractive summarization

Methodology	Rouge-1	Rouge-2	Rouge-L
Bert based encoder	12.5	4.29	9.76
Bert based encoder for subset	12.85	8.51	9.75

Table 1: Results for bert based models in two settings



Model	Original summary	Predicted summary	Inference
att-Seq2Seq-tf	यादव परिवार में मचा राजनीतिक घमासान और उठापठक जारी है मुलामय की कोशिश यूपी चुनावों में बिहार की तर्ज़ पर महागठबंधन बनाने की है RLD प्रमुख अजित सिंह और जेडीयू के पूर्व अध्यक्ष शरद यादव से संपर्क साधा	में में	Irrelevant and grammatically incorrect
Bert based Encoder (trained for 2000 samples)	कहा राज्यसभा सीट जीतने के लिए बीजेपी हर हथकंडा अपनाने पर आमादागुजरात में कांग्रेस के कार्यकर्ताओं का मनोबल गिराने की चालसोनिया गांधी को एक सेटबैक पहुंचाने की कोशिश	समाज में माननाम आjan भी लागू हो जाएगायह घटनाएं फशा करते हुए समाज के इस्तीक्षे को ML उम्मीद दी गई हैउन्होंने कहा कि इस सीएम में डीएम की कमुनजी वाले दायर एकमात्र तिहरा मानहाना एक्सजेसी की परिणामस्वरूप मामला सामने आने के लिए चुना गया है वाला लाइसेंस की सारीक़ाट्यटकणुराुराക്കൾ Listenवमेंट में लेख ली जाती	Irrelevant and grammatically incorrect.
Bert based Encoder (trained for 35470 samples)	दिल्ली एयरपोर्ट के पास मुसाफिरों से पुलिसवाले बनकर करते थे ठगीपुलिस ने 4 लोगों को किया गए	आईसीसी ने इंग्लैंड के पूर्व कप्तान नासिर हुसैन के उस बयान को पूरी तरह अवांछित करार दिया जिसमें उन्होंने भारतीय कप्तान साबित किया था कि उन्हें	Irrelevant but at least grammatically correct.

Figure 4: =
Prediction results with bert based model

The seq2seq model is not able to capture the complexity of Hindi words at all and fails to generate any sensible output sentences. The BERT based model first trained on a subset of articles and later on the large number of articles shown to obtain similar rouge scores but from Figure 4 it is clear that for a subset of articles model is not able to produce even grammatically correct sentences but when trained for more number of articles, it is shown to produce grammatically correct sentences but totally irrelevant with respect to input sentences at test time.

5.1.1 *Model Training.*

To further analyze the model while training, we can see that the model is able to generalize well for the train data by the evidence given by the train loss curve below and predicted samples at train time shown in the table below. Due to the limited resources, the results below are shown for model training on subset data.

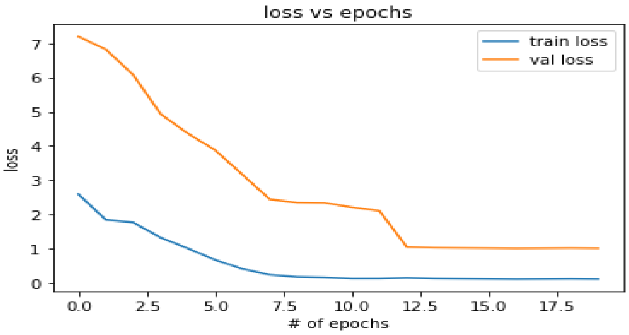
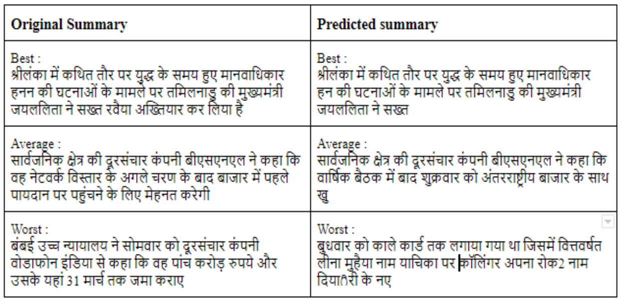


Figure 5: =
Loss curve for bert based model



Original Summary	Predicted summary
Best : श्रीलंका में कथित तौर पर युद्ध के समय हुए मानवाधिकार हनन की घटनाओं के मामले पर तमिलनाडु की मुख्यमंत्री जयललिता ने सख्त रवैया अख्तियार कर लिया है	Best : श्रीलंका में कथित तौर पर युद्ध के समय हुए मानवाधिकार हन की घटनाओं के मामले पर तमिलनाडु की मुख्यमंत्री जयललिता ने सख्त
Average : सार्वजनिक क्षेत्र की दूरसंचार कंपनी बीएसएनएल ने कहा कि वह नेटवर्क विस्तार के अगले चरण के बाद बाजार में पहले पायदान पर पहुंचने के लिए मेहनत करेगी	Average : सार्वजनिक क्षेत्र की दूरसंचार कंपनी बीएसएनएल ने कहा कि वार्षिक बैठक में बाद शुक्रवार को अंतरराष्ट्रीय बाजार के साथ खु
Worst : बंबई उच्च न्यायालय ने सोमवार को दूरसंचार कंपनी वोडाफोन इंडिया से कहा कि वह पांच करोड़ रुपये और उसके यहां 31 मार्च तक जमा कराए	Worst : बुधवार को काले कार्ड तक लगाया गया था जिसमें वित्तवर्षत लीना मुहैया नाम याचिका पर कॉलिंगर अपना रोक2 नाम दियाितरी के नए

Figure 6: =
Train time predictions for bert based model

5.2 Extractive Summarization

As the summaries are produced as sentences extracted from the main news article, we kept our baseline as the leading three sentences from the news article as summary. This method is frequently used in the literature [9–11]. It can be seen that BERTSUM-RNN performs the best on the extractive summarization task. Figure 5 shows the most frequent sentence positions picked up by BERTSUM-RNN to generate extractive summaries. It shows that almost always the summary sentences are picked from the first three sentences. However, it still performs better than LEAD-3 baseline because BERTSUM-RNN predicts the same number of summary lines as that of silver summary lines whereas LEAD-3 predicts starting three sentences blindly which reduces its f1-score.

Methodology	Rouge-1	Rouge-2	Rouge-L
TextRank	49	37	46
LEAD-3	55	49	57
BERTSUM-Simple	66	32	49
BERTSUM-RNN	67	59	64
Bert based encoder for subset	12.85	8.51	9.75

Table 2: Results for bert based models in two settings

6 CONTRIBUTIONS

Data preprocessing and cleaning is done by all team members Ritwik: implemented seq2seq model and BERTSUM-RNN Dimpy - implemented bert based model. Aamir - implemented textrank and BERTSUM - Simple Classifier

REFERENCES

[1] Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

[2] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

[3] Liu, Y. (2019). Fine-tune BERT for extractive summarization. arXiv preprint arXiv:1903.10318.

[4] Gupta, M., & Garg, N. K. (2016, September). Text summarization of Hindi documents using rule based approach. In 2016 international conference on micro-electronics and telecommunication engineering (ICMETE) (pp. 366-370). IEEE.

[5] Kumar, K. V., & Yadav, D. (2015). An improvised extractive approach to hindi text summarization. In Information Systems Design and Intelligent Applications (pp. 291-300). Springer, New Delhi.

[6] Desai, N., & Shah, P. (2016). Automatic Text Summarization Using Supervised Machine Learning Technique for Hindi Langauge. IJRET: International Journal of Research in Engineering and Technology, 2319-1163.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[8] Gaurav. (2020, January 5). Hindi Text Short and Large Summarization Corpus. Retrieved from https://www.kaggle.com/disisbig/hindi-text-short-and-large-summarization-corpus

[9] Dohare, S., Karnick, H., & Gupta, V. (2017). Text summarization using abstract meaning representation. arXiv preprint arXiv:1706.01678.

[10] Kryscinski, W., Keskar, N. S., McCann, B., Xiong, C., Socher, R. (2019, November). Neural text summarization: A critical evaluation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 540-551).

[11] Zhang, X., Lapata, M., Wei, F., & Zhou, M. (2018). Neural latent extractive document summarization. arXiv preprint arXiv:1808.07187.

7 APPENDIX

7.1 Text Rank

We have implemented Text-Rank for clean 59k sentences. The Average BLEU Score= 0.454.

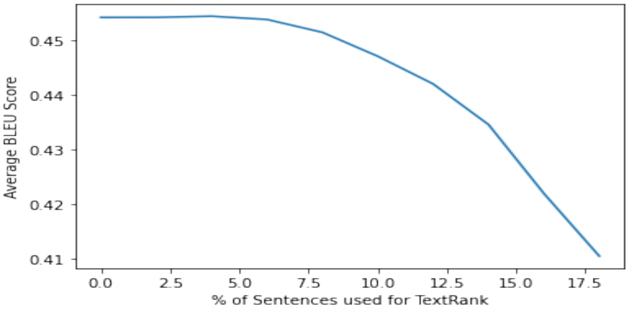


Figure 7: =
Plot for number of Top-K sentences vs Bleu Score

Here we can see that the most important information for finding the summary is captured only in the first few sentences of all the ranked sentences. So proceeding with this, we calculated the BleuScore for max(1,Top 10% of sentences for an article). Using this metric the following graph was obtained.

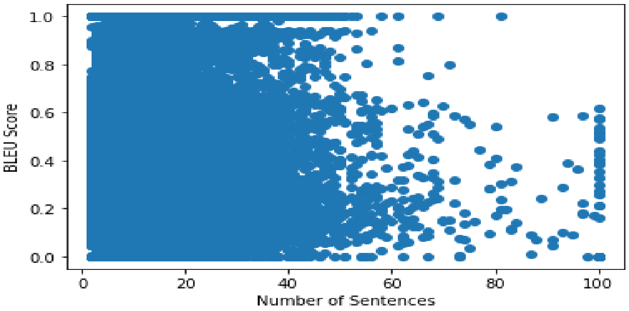


Figure 8: =
Plot for number of sentences in an article vs Bleu Score

Using the above plot we wanted to know if number of sentences had any correlation with the Bleu Score, as this one of few parameters which can be changed for finding text-rank. But the plot shows that there is a uniform spread for Bleu Score vs number of sentences in an article.

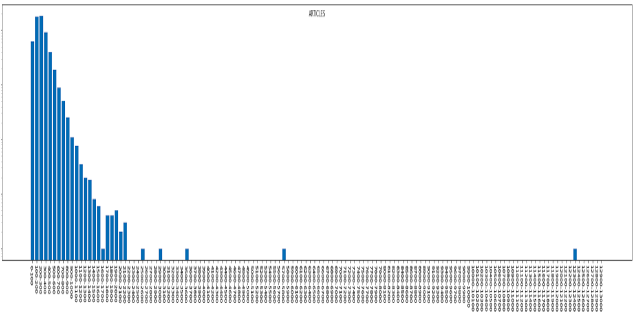
We observed that sentences with high Bleu Score had perfect matching between summary and Textrank sentences extracted. This is because the top-k sentences is essentially the summary. And on the other hand, low Bleu Score summaries had incorrect Sentence selected in top-k. This is because the way TextRank algorithm works. A sentence which is long or has lots of key words, which are used throughout the article, is automatically given a higher rating even though that sentence may not be a good summary of the overall article. Eg, In this article talking about filing court cases, the sentence about Tata's court case has lots of key words which are used throughout the article, and so has a very high rank, but isn't necessarily a good summary for the whole article.

7.1.1 Analysis. :

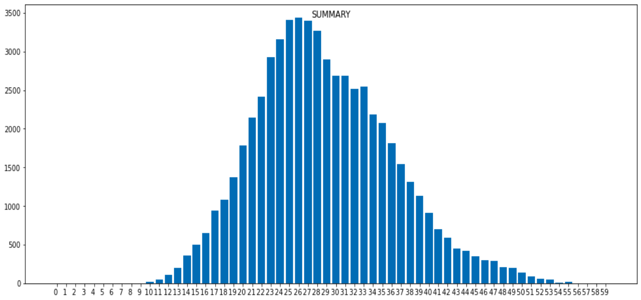
Textrank, which stems from Pagerank, suffers a lot of problems associated with Pagerank. Sentences which are long and contain a lot of keywords automatically gain a higher rank because they are important sources of most information, which isn't always required while writing summaries. So while they contain a lot of keywords contained in the actually summary, because the way Bleu score works (which rewards closeness to the final sentence rather than % of words matched), they get a poor score.

This fact precisely explains the reasoning for the varied Bleu scores we observe. The articles have a hit and miss chance of having a good TextRank summary. If the top-k sentences have a long sentence which also happens to be the summary, they get a good score. Otherwise the long sentences which sometimes contain explanations, examples, etc and have a lot of keywords but aren't the summary get a bad score.

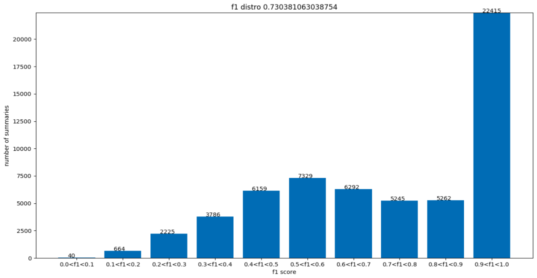
7.2 Images for Other Experiments



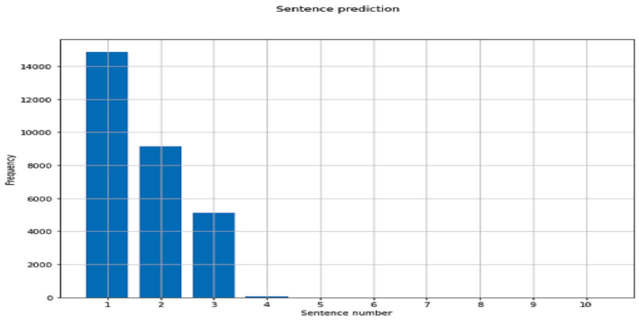
Distribution of token length in the article text. The X-axis represents token lengths in brackets of 100. The Y-axis represents the frequency.



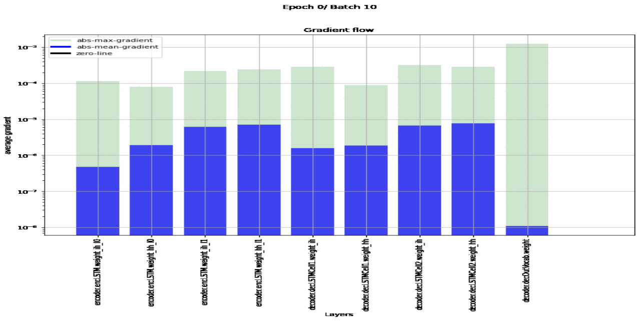
Distribution of token length in summary text. The X-axis represents token lengths. The Y-axis represents the frequency.



Distribution of F1-score in the silver extractive summaries from gold abstractive summaries.



Frequency of Sentences selected by model



Gradients of model being updated