Abstractive Headline Generation from Bangla News Articles Using Seq2Seq RNNs with Global Attention

Ruhul Amin*, Nabila Sabrin Sworna*, Md Nazmul Khan Liton*, Nahid Hossain*†
Dept. of Computer Science and Engineering
United International University
Dhaka, Bangladesh
ruhulamin6678@gmail.com, nsworna153069@bscse.uiu.ac.bd, liton079@gmail.com, nahid@cse.uiu.ac.bd

***Abstract*—Headline generation is the process of generating headlines automatically from text articles. We model a comprehensive abstractive headline generation technique using Seq2Seq Recurrent Neural Networks with Global Attention in this work. Despite being one of the most spoken languages globally, very few significant works have been done on this particular topic in the Bangla language. Thus, our model is solely based on the Bangla language, and we find that the performance of the model is highly satisfactory at the current stage. We also propose an extensive dataset consisting of 5,14,108 filtered Bangla news articles in full and other necessary information. The dataset has been created by scrapping several online reputed Bangla newspapers. Due to the unavailability of a proper and updated dataset, the proposed datasets are freely available at https://tinyurl.com/banglaHead**

***Index Terms*—Bangla, headline generation, seq2seq, gated recurrent unit, attention, news articles**

I. Introduction

In the 21st century, people have driven to know the unknown, and it is a century of information overabundance [1], [2]. Now, online newspapers, articles, blogs, and other e-contents have become very popular, and thus, we have a lot of text data available on the internet [3]. We mostly browse the web to get news about a particular topic. However, the search engine returns a large amount of news on that topic for the end-user. Therefore, an appropriate and specific headline of a news body plays an important role to reduce a user's scepticism and to stop wasting time reading irrelevant articles.

Automatic headline generation is an application of Automatic Text Summarization (ATS) technique [1]–[8]. ATS is the task of picking up the most informational elements from the reference article accurately. It converts the whole document into an encapsulated short script. In this version, all the core concepts of the primary text should be present. This system can be useful for blogs, newspapers, and any text-based platform. The headline generation system can be classified into an Extractive and an Abstractive approach. We focus on abstractive headline generation. This approach is more accurate to produce real-life headlines than the extractive approach.

The 7^{th} most spoken language in the world is Bangla (endonym Bengali) [4]. Apart from Bangladesh, there are some other regions all over the world where people speak in Bangla [5]. As the number of Bangla speakers and users of Bangla text data increased, it is now necessary to build a system that can generate the headline of Bangla articles automatically. Although many types of research have been done on headline generation in the English language, according to our study, no study has been found on this particular topic in Bangla language. This motivates us to work on automatic headline generator in Bangla language.

The rest of the paper is presented in the following way: In section 2, we have discussed the related works. We have demonstrated our proposed method, including dataset collection, preprocessing, numerical characterization, and methodology in section 3. In methodology, we have also explained the working procedures of seq2seq learning. In section 4, we have described experimental results and analysis of our system. Lastly, section 5 concludes the paper by mentioning limitations and future work on the headline generation system.

II. Related Works

Due to the unavailability of a sophisticated headline generation approach in Bangla, we have studied few papers for automatic text summarization in Bangla as well as in English. Abujar et al. proposed a new and efficient extractive Bangla text summarized approach. In this paper, they have derived a set of Bangla text analysis rules from the heuristics. They have presented the Bangla sentence clustering method, and sentence scoring processes for summarization [5]. Partha et al. proposed a rule-based approach to summarized Bangla news documents using the extractive approach. They have illustrated a graph-based sentence scoring feature for text summarization [1]. Abujar et al. illustrated the word2vector approach for text summarization in Bangla. They have mentioned that word2vec used for word embedding. They have applied two methods for word embedding, such as Skip-Gram and Continuous Bag of Word Model [6]. Efat et al. proposed a method for Bangla text summarization based on an extractive approach. They have firstly preprocessed Bangla dataset by tokeniza-

*equal contribution
†corresponding author

tion, stop words removal, stemming. Afterwards, they have given the sentence's rank based on these significant features: Frequency, Position Value, Cue Words, and Skeleton of the document. They have also used a Linear Combination of these 4 features to obtain the final sentence score. Subsequently, ranked the sentences based on the final score to make the summary [7]. Sarkar et al. proposed the implementation of term frequency and semantic sentence similarity methods to summarize Bangla text. Firstly, they apply term frequency and then semantic sentence similarity to rank the sentences. After acquiring a bunch of top-ranked sentences, the system makes the summary [13]. Neural network-based methods have achieved great success in the Natural language generation tasks. Takase et al. proposed the attention-based AMR encoder-decoder method to improve headline generation from a longer text [16]. The encoded results which are obtained from an AMR (Abstract Meaning Representation) parser. They have used the reformed version of Tree-LSTM. In the next year, Shan et al. have given details about encoders, decoders, and neural model training strategies [17]. In their work, they have compared different methods to show the improvement of neural headline generation. In the same year, Tilk et al. illustrated a method in the improvement of headline generation [18]. They have focused on a small dataset. Usually, neural headline generation works better on an extensive dataset. They have improved headline quality on a small dataset. Hayashi et al. used a recurrent neural network in headline generation tasks [3]. They have presented a method that consists of LSTM (Long Short Term Memory) based encoder and decoder.

III. Proposed Method

In this section, we have described the development methodology of our proposed system.

A. Dataset Collection

The lack of a comprehensive Bangla headline generation approach and the lack of an extensive dataset of Bangla news articles motivates us to develop our own dataset by scraping online sources. We have used python's advanced Beautiful Soup library to make a spider. The spider automatically scrapes different renowned and reputed sources, i.e. 'Prothom Alo', 'Kalerkantho', 'Banglanews24' and 'Samakal'. The dataset contains Bangla news bodies and corresponding headlines. We have also constructed our own Bangla Stop Words dataset from these available sources [19]–[21].

TABLE I
Summary of the datasets.

Dataset	Number of Records
Training Dataset	433,356
Test Dataset	80,752
Stop Words Dataset	630

B. Data Preprocessing

Data preprocessing plays a vital role in any machine learning related task. This step is very crucial to enhance the quality of the model. Bangla is a language with lots of complex grammatical structures. Before feeding the textual data to our deep learning model, it is required some tweaking. The following measures have taken to preprocess our corpus.

- HTML Tags Removal: Since we have collected our data through web-scraping, it contains some HTML tags which need to be removed. These tags have no significance for our intended task, thus have been removed from our corpus.

- Punctuation Removal: Bangla text data contain many punctuation characters. These characters are less significant to acknowledge the context of the text. So, we have removed these punctuation (।,<,>,:,;,|,(,),{,},[,],&,!,etc.) signs from our corpus.

- Stop Words Removal: Stop words repeatedly emerge in the corpus but do not incorporate adequate information. Like other Languages, Bangla has its own stop words too. These type of words are ইহা, অথবা, আবার, আর, আরও, and so forth. These words have been removed only from news articles to emphasize significant words that define the context of the news.

- Numerical Characterization: Text cannot be passed directly to a deep learning model as it requires some numerical representations. For this, we maintain two lookup tables (vocabulary), which map each word from news articles and headlines to an integer number separately. We only consider the most frequent 10,000 words for news article and 4,000 words for headline's lookup tables respectively. We also remove any numbers and foreign words. Due to hardware limitations, we restrict the maximum number of words in each news body and headline to 60 and 10, respectively. We applied pre-padding to the news articles and post-padding to the headlines having a word count less than the defined limit. Moreover, we have added <START> and <STOP> tokens at the beginning and end of each headline.

TABLE II
Insights of the training dataset.

Encoder vocabulary size	10,000
News articles max. word limit	60
Decoder vocabulary size	4,000
News headlines max. word limit	10

C. Methodology

Deep Neural Networks (DNNs) have been applied in many domains of NLP, which evidently achieved excellent performances. However, these types of neural networks cannot be applied to portray one sequence to another. To deal with this sort of problem, Sutskever et al. [22] and Cho et al. [23] first proposed the sequence to sequence learning. The fundamental components of a seq2seq model are encoder and decoder network. The encoder steps through the inputs and encodes the inputs to a specified length vector, whereas the decoder extracts the target one by one conditioned by the encoder's output.

Variants of Recurrent Neural Network (RNN) such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) are preferred in seq2seq model because these are proficient in capturing long-term dependencies by overcoming the vanishing and exploding gradient problems that RNN often suffers from [24]. In this work, we have used Gated Recurrent Unit (GRU) in both the encoder and decoder network. Our choice of GRU over LSTM is to achieve competitive performance because LSTM is comparatively slower due to its complex inner layout [25]. The structure of a GRU unit is given below:

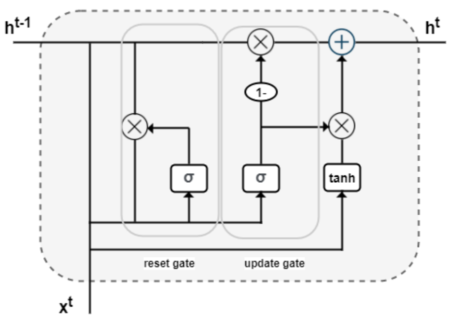

Fig. 1. Diagram of a GRU cell.

The encoder is required to compress every single essential information from the input sequence and transform it into a single concise vector. However, it often suffers from coping with long input sentences. Considering the properties of our corpus, we have used attention mechanism which enables the decoder to inspect the sources selectively. We have used the global attention technique in our work proposed by Bahdanau et al. [26]. We can categorize the whole process in the following 2 phases:

Training Phase: Initially, we set up the encoder-decoder networks along with the attention mechanism to build the model. Then, we train the model with our training and validation data and predict the headline offset by one timestamp.

Encoder: The encoder network in our model is built with 3 layers of stacked Bidirectional GRU on top of each other. These layers step through a single element of the input news article, assemble the information, and transmit it forward. Bidirectional GRU layers facilitate the learning by processing the input sequences in both forward and backward means simultaneously. This way, the encoder can learn the sophisticated sequential insights from the input article. Both the forward and backward output is concatenated before being conveyed to the decoder.

Attention Mechanism: Attention operates as a gateway between the encoder and decoder network. It offers the decoder insights into the source by observing the hidden states of the encoder. This enables the decoder to give selective preference on specific parts of the source. It facilitates learning an effective alignment between the source and the target. This assists the model in dealing with the issue of long input sequences. The attention layer uses alignment scores, weighting, and context vector to prioritize the specific hidden state from the input sequence. The alignment score maps how suitable each encoded input as regards the decoder's current output. This score is calculated from the encoder's hidden states and the decoder's output of the previous timestamp. For the first timestamp of the decoder, it will be zero. Then a softmax activation is applied to these alignment scores to get the attention weights. These normalized attention weights work as probabilities, which infer that each hidden state's likelihood is relevant to the decoder's current output. The context vector is then computed by utilizing the hidden states from the encoder and the weighted sum of the attention weights. This makes it possible to get a context vector that is refined explicitly for each output timestamp of the decoder.

Decoder: The decoder network in our model is a single unidirectional GRU layer and a temporal dense layer which steps through the target headline word by word and attempts to predict the same headline offset by one timestamp. This temporal dense layer performs the usual dense operation for every timestamp of the GRU layer's output separately. It has the same number of units as the output vocabulary, which produces a likelihood for each word being the next word in the headline. The decoder's initial input is the <START> token and the hidden states from the encoder's last layer. It is trained to anticipate the next word from the target headline given the previous output of the decoder and the customized context vector for the current timestamp.

Reasoning Phase: After completing the training process, the model is tested on new test news articles. For this, we need to set up a reasoning architecture to generate a headline for which the target is unknown. The procedures for generating a headline from a given article is described below. First, we encode the entire input article using the encoder. Then we initialize the decoder with <START> token as input along with the hidden states of the encoder. The decoder then steps one timestamp and produces a probability vector for each word in the vocabulary. We select the word with maximum probability. After that, the internal state gets updated with the current state. Then we pass the generated word to get the prediction for the next word. This process is repeated until the decoder generates the <STOP> token or reaches the maximum limit of words in a headline (10 in our case).

To create our model, we have used python's deep learning framework Keras with Tensorflow backend. 15% of training data (randomly selected) have been used as validation data. In our model, we first use the Embedding layer for both the encoder and decoder to learn the vector representations of

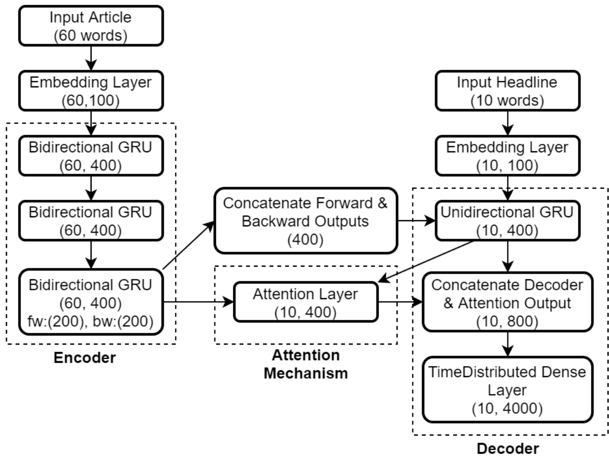

Fig. 2. A high level illustration of the headline generation model architecture.

words from the numerical characterization of source articles and target headlines. The number of GRU units in each layer of the encoder and the decoder is 200 and 400 respectively. We have used a dropout rate of 0.4 for both the input step and the recurrent state in each layer of the encoder and the decoder network. Dropout randomly leaves input units and recurrent state while calculating the linear transformation in every epoch to assist in subduing the overfitting issue. Hyperbolic Tangent (tanh) and Sigmoid activation function are used in the input step and recurrent step, respectively. For the final temporal dense layer, the softmax activation function is used. Adam has been chosen as the optimizer to update weights. The adaptive learning rate is used with an initial learning rate of 0.001 and a mini-batch size of 256 while training. Sparse Categorical Cross Entropy has been chosen as the loss function. This conserves memory by transforming the numerical input characterizations to one hot encoded vector on the fly. We have also used an early stopping mechanism to avoid overfitting.

IV. Experimental Results and Analysis

In this section, we discuss the experimental results. The experiment has been performed on Kaggle Notebook, which is a free cloud computational service with 5 GB disk space, 13 GB RAM and NVIDIA Tesla P100 GPU and 16 GB memory [27]. Table III shows some examples of headline prediction by our model. While dealing with language-based sequences, measuring the results or performance of the model becomes a lot more complex. There is no quantifiable or objective 'good' when it comes to the quality of a machine-generated headline from a news article. Human judgement is the benchmark for these types of tasks. So, the automatic evaluation metrics should be correlated with human judgement. We have used Bilingual Evaluation Understudy (BLEU), and Recall Oriented Understudy for Gisting Evaluation (ROUGE) score to quantify the effectiveness of the generated headlines. Initially, BLEU is proposed to evaluate the performance of machine translation [28], and ROUGE is proposed to evaluate summaries [29]. These metrics are well suited to our work. The average results obtained from our test dataset is given in Table IV.

TABLE III
Prediction Examples

Text	**Actual Headline**	**Predicted Headline**
সিলেটের কোম্পানীগঞ্জে অবৈধভাবে টিলা কেটে পাথর উত্তোলন ও গর্ত ধসে শ্রমিক হতাহতের ঘটনা পুলিশ ও প্রশাসন নয়. বিচার বিভাগীয় তদন্ত দাবি করেছে বাংলাদেশ পরিবেশ আন্দোলন (বাপা)। গতকাল বৃহস্পতিবার বিকেলে সিলেটের কেন্দ্রীয় শহীদ মিনারের সামনে মানববন্ধন চলাকালে সমাবেশে বক্তারা যাদের গাফিলতিতে ঘটনা ঘটল. তাদের দিয়ে তদন্ত কেন প্রশ্ন তুলে বিচার বিভাগীয় তদন্তের দাবি করেন। শাহ আরেফিন টিলায় গত সোমবার সকালে...	শ্রমিক হতাহতের ঘটনায় বিচার বিভাগীয় তদন্ত দাবি	শ্রমিক হত্যার তদন্ত দাবি
করোনার উপসর্গ জ্বর, ঠান্ডা, কাশি ও শ্বাসকষ্ট নিয়ে এক ব্যক্তির মৃত্যুর হয়েছে। আজ মঙ্গলবার সকাল সাড়ে সাতটার দিকে শরীয়তপুরের নড়িয়া উপজেলা স্বাস্থ্য কমপ্লেক্সের আইসোলেশন ওয়ার্ডে তিনি মারা যান। ওই ব্যক্তি (৫০) নড়িয়া পৌরসভার ৪ নম্বর ওয়ার্ডের বাসিন্দা। অসুস্থ অবস্থায় গতকাল সোমবার সন্ধ্যায় স্বজনেরা তাঁকে উপজেলা স্বাস্থ্য কমপ্লেক্সে ভর্তি করেন। করোনার উপসর্গ থাকায় উপজেলা স্বাস্থ্য বিভাগ তাঁর লাশ থেকে নমুনা সংগ্রহ...	করোনার উপসর্গ নিয়ে এক ব্যক্তির মৃত্যু	করোনা আক্রান্ত ব্যক্তির মৃত্যু

TABLE IV
Summary of Experimental Results

Evaluation Metrics	**Precision**	**Recall**	**F1 Score**
ROUGE-1	33.65	31.02	31.79
ROUGE-2	23.11	20.37	21.63
ROUGE-L	28.74	25.97	26.48
BLEU	32.47		

$$Precision = \frac{\begin{array}{r} number\ of\ n-grams\ matches\ in \\ generated\ \&\ reference\ headline \end{array}}{\begin{array}{r} number\ of\ n-grams\ in \\ generated\ headline \end{array}} \tag{1}$$

$$Recall = \frac{\begin{array}{r} number\ of\ n-grams\ matches\ in \\ generated\ \&\ reference\ headline \end{array}}{\begin{array}{r} number\ of\ n-grams\ in \\ reference\ headline \end{array}} \tag{2}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

The reported result of the BLEU score is comprised of unigram, bigram, and trigram with weights 0.4, 0.3, and 0.3, respectively. BLEU score reflects how the generated headline is similar to the original one. It primarily focuses on strings. We have also used ROUGE-N (1-gram and 2-gram) and ROUGE-L to evaluate the generated headlines. It includes a set of metrics that are shown in equation 1, 2 and 3. ROUGE-N weighs the n-grams that match with the actual or reference headline and model generated headline. ROUGE-L measures the longest common subsequences shared between the actual and generated headline. Results of ROUGE-1 and ROUGE-2 in table IV indicate that two consecutive words in generated headlines appear less than a single word in reference to the original headlines. While assessing some of the generated headlines manually, we observed that several of them scored low on ROUGE and BLEU even though they look reasonable in terms of the original news article.

V. Conclusion

This paper has presented and demonstrated a comprehensive approach of headline generation on Bangla news articles based on Seq2Seq Recurrent Neural Networks with Global Attention mechanism. The model shows very encouraging results on our large novel dataset of Bangla news articles. Using only the first 60 words of a news article, the model generates very auspicious headlines even though BLEU and ROUGE scores are low similar to other state-of-the-art English works. Although we have reached our goal, we observed some limitations to our work. The only limitation of the system is the size of our dataset. Although we have collected a sufficiently large dataset of news articles using the crawling and scraping technique, we need a very extensive dataset for more efficient and accurate results. The authors are working to collect more data to propose a more substantial dataset soon. In addition to that, the authors are working on a model that combined both extractive and abstractive approach to further increase the accuracy rate for the upcoming version. The authors are also working on a web app of our proposed model. For the time being, our datasets are available for future researchers to use at https://tinyurl.com/banglaHead

References

[1] Ghosh, P.P., Shahariar, R., Khan, M.A.: A Rule Based Extractive Text Summarization Technique for Bangla News Documents. International Journal of Modern Education and Computer Science, Modern Education and Computer Science Press. vol. 10(12), pp. 44, Dec 2018.

[2] Patel, V., Tabrizi, N.: An Automatic Text Summarization: A Systematic Review , https://easychair.org/publications/preprint/z2ml, EasyChair Preprint no. 86, Mar 2019.

[3] Hayashi Y., Yanagimoto H. (2018) Headline Generation with Recurrent Neural Network. In: Matsuo T., Mine T., Hirokawa S. (eds) New Trends in E-service and Smart Computing. Studies in Computational Intelligence, vol 742. Springer, Cham

[4] Haque, M.M., Pervin, S., Begum, Z.: Enhancement of keyphrase-based approach of automatic Bangla text summarization. 2016 IEEE Region 10 Conference (TENCON), IEEE. pp. 42–46, Nov 2016.

[5] Abujar, S., Hasan, M., Shahin, M., Hossain, S.A.: A heuristic approach of text summarization for Bengali documentation. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) IEEE. pp. 1–8, Jul 2017.

[6] Abujar, S., Masum, A.K.M., Mohibullah, M., Hossain, S.A, et al.: An Approach for Bengali Text Summarization using Word2Vector. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE. pp. 1–5, Jul 2019.

[7] Efat, M.I.A., Ibrahim, M., Kayesh, H.: Automated Bangla text summarization by sentence scoring and ranking. 2013 International Conference on Informatics, Electronics and Vision (ICIEV), IEEE. pp. 1–5, 2013.

[8] Nallapati, R., et al.: Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023, Aug 2016.

[9] Nallapati, R., et al.: Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. ArXiv:1602.06023 [Cs]. Aug. 2016. arXiv.org, http://arxiv.org/abs/1602.06023.

[10] Song, S., Huang, H., Ruan, T.: Abstractive text summarization using LSTM-CNN based deep learning. 2Multimedia Tools and Applications, Springer. vol 781, pp. 857–875, Feb 2019.

[11] Thakkar, K.S., Dharaskar, R.V., Chandak, M.B.: Graph-based algorithms for text summarization. 2010 3rd International Conference on Emerging Trends in Engineering and Technology, IEEE. pp. 516–519, 2010.

[12] Yousefi-Azar, M., Hamey, L.: Text summarization using unsupervised deep learning. Expert Systems with Applications, Elsevier. vol. 68, pp. 93–105, Aug 2017.

[13] Sarkar, A., Hossen, M.S.: Automatic Bangla Text Summarization Using Term Frequency and Semantic Similarity Approach. 2018 21st International Conference of Computer and Information Technology (ICCIT), IEEE. pp. 1–6, Dec 2018.

[14] Uddin, M.N., Khan, S.A.: A study on text summarization techniques and implement few of them for Bangla language. 2007 10th international conference on computer and information technology, IEEE. pp. 1–4, 2007.

[15] Ullah, S., Hossain, S., Hasan, K.A.: Opinion Summarization of Bangla Texts using Cosine Simillarity Based Graph Ranking and Relevance Based Approach. 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE. pp. 1–6, sept 2019.

[16] Takase, S., et al.: Neural headline generation on abstract meaning representation. Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 1054–1059, Nov 2016.

[17] Shen, S., et al. Recent advances on neural headline generation. Journal of Computer Science and Technology, Springer. vol. 32(4), pp. 768–784, jul 2017.

[18] Tilk, O., Alumäe, T.: Low-resource neural headline generation. arXiv preprint arXiv:1707.09769, Jul 2017.

[19] Bengali Stopwords, https://www.ranks.nl/stopwords/bengali

[20] Stopwords-bn, https://github.com/stopwords-iso/stopwords-bn

[21] Bengali Stopwords, https://sanjir.com/6202/

[22] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks. arXiv preprint arXiv:1409.3215v3 (2014).

[23] Cho, K., et al.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1409.3215v3 (2014).

[24] Shewalkar, A., Nyavanandi, D., Ludwig, S. A.: Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU. Journal of Artificial Intelligence and Soft Computing Research. 9(4), 235-245 (2019) doi:https://doi.org/10.2478/jaiscr-2019-0006

[25] Viswanathan S., Anand Kumar M., Soman K.P.: A Sequence-Based Machine Comprehension Modeling Using LSTM and GRU. In: Sridhar V., Padma M., Rao K. (eds) Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol 545. Springer, Singapore (2019). doi:https://doi.org/10.1007/978-981-13-5802-9_5

[26] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).

[27] "Notebooks Documentation." Kaggle, www.kaggle.com/docs/kernels?fbclid=IwAR0a4pxbvpQetZ6N O8rvsLXQ9SOzcqCJG1a4DP8N-Dy-ZzOfxrFDwdYkwrM#technical-specifications.

[28] Papineni, K., et al.: BLEU: a Method for Automatic Evaluation of Machine Translation.(2002). doi: 10.3115/1073083.1073135.

[29] Lin, C.: ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, Association for Computational Linguistics. 2004, pp. 74–81.