Proceedings of 10th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: October Volume: 10

Extractive Method for Nepali Text Summarization Using Text Ranking and LSTM

Rishi Saran Khanal [a], Smita Adhikari [b], Sharan Thapa [b]

[a] *Department of Electronics and Computer Engineering,Institute of Engineering , Paschimanchal Campus Lamachaur*
Corresponding Email: [a] rishikhanal892@gmail.com, [b] smita.adhikari@pasc.tu.edu.np [b] sharant@wrc.edu.np

Abstract
Extracting the gist and main theme of any document is called summary. So text summarization process helps for understanding sense of any text file without reading all sentences. In machine learning there are two type of summarization method i.e abstractive and extractive method. In this research extractive method of summarization is followed for Nepali text document with help of word embedding. For word embedding purpose GloVe algorithm is used, which create embedding of Nepali word corpus. Before feeding to GloVe algorithm preprocessing of data is done to the given raw data. For text summarization of Nepali text, sentence embedding is created with help of GloVe embedding, after that similarity matrix is created between sentences using cosine similarity. Similarity matrix is used to create graph G(V,E), where sentences is treated as vertex and similarity value between two sentences is assigned as weight of edges. After that Text rank algorithm is applied to select top k sentences. These selected sentences become summary as result of system generated summary. Also attention based LSTM encoder decoder model is trained for text summarization. In this model, GloVE embedding of Neapli words is feed to encoder with three stacked LSTM encoder. Output of sequence embedding of encoder is taken as input to decoder. While using attention layer in encoder decoder model, it allows to create a context-vector at each time step, given the decoder's current hidden state and a subset of the encoder's hidden states. The context vector c_i depends on a sequence of hidden states to which the encoder maps the input sentence. Each hidden state h_i contains information about the whole input sequence with strong focus on the parts surrounding the i^{th} word of the input sentence. The context vector is computed as a weighted sum of these hidden states h_ii. After that sequence decoder produces summary as predicted summary. For the evaluation of system generated summary ROUGE-1 and ROUGE -2 is calculated with the reference of human generated summary. For this purpose, 14 different categories of Nepali news articles are collected as corpus for training to GloVe model.

Keywords
Text Summarization, GloVe, Text Rank Algorithm, LSTM Encoder Decoder, Attention, ROUGE

1. Introduction

Now a days many textual information is produced from anywhere through digital media due to which there is time consuming process regarding collecting important information about any topic by reading all the related document. So summarization process helps to extract key information from document. There are mainly two methods to summarize as abstractive summarization and extractive text summarization. Abstractive method is closer to human intelligence method because in this method summarization is performed based on semantic representation of text. In this method, natural language generation technique is applied. These summaries are grammatically correct and more coherent. Since this method generates new phrases and sentences which represent a most important sense of a given text document, it can help for generating a more accurate summary than that of extractive method. Also this method reduces chances of grammatical inaccuracies than of extractive method. In extractive method most important sections of text are extracted. This method calculates score for weight of sentences in text file by calculating average embedding value of sentences. Which sentences having a higher ranking score, these are selected as summary. In this method total 40% of

text is selected as summary.

In this summarization process, we have used GloVe(Global Vector) for vectorization of corpus ,Text Rank algorithm is used for sentence ranking to generate extractive summary given text .Text rank algorithm is unsupervised method. Also, LSTM with attention based method is used for supervised method for text summarization. In many research pre-trained word embedding provided by different researcher is used for text summarization. In this thesis Nepali News Corpus is trained on GloVe embedding. So main contribution is collecting Nepali news corpus and has to train to GloVe for word embedding. After that these trained embedding is used for generating extractive summary of Nepali text using Text Rank algorithm and attention based Encoder-Decoder LSTM Model. After that output of these models are evaluated with comparison each other and also compared with other researcher output.

In many research pre-trained word embedding provided by different researchers is used for text summarization. In this thesis, Nepali News Corpus is trained on GloVe embedding. So main contribution is collecting Nepali news corpus and has to train GloVe for word embedding. After that, this trained embedding is used for generating an extractive summary of Nepali text using the Text Rank algorithm and attention-based Encoder-Decoder LSTM Model. After that output of these models is evaluated with comparison to each other.

The remaining part of our article is organized as follows: Related works to this study are described in Section 2. In section 3, we have described our proposed methodology. Experiments and results are described in section 4. Finally, in section 5, we conclude our work.

2. Related Work

For word embedding model, GloVe model is developed by Stanford University[1]. In this model, word-word co-occurrence matrix is developed. This model was trained with a 2010 Wikipedia dump with 1 billion tokens, a 2014 Wikipedia with 1.6 billion tokens, a Gigaword 5 with 4.3 billion tokens and 42 billion tokens of web data. They trained corpus by setting Xmax=100, α=3/4 with initial learning rate=0.05 with iterations size as 50 for vector smaller than 300 dimensional, and iterations size as 100 otherwise and used window size as 10. Also this model is compared with other word embedding model, word2vec, CBOW. This research is based on English language. But in this thesis, our own Nepali News Corpus has to be trained with GloVe Model.

Aditya Jain teams perform text summarization using Word Vector Embedding[2].For vectorization of word,100 dimensional vector sized pre-trained GloVe embedding is used to vectorize every word present in sentence of document. Average sentence embedding score is calculated by using mean embedding dimension value of each word vectors. For test performance first 284 documents of DUC 2000 dataset are used for purposed model. ROUGE-1, ROUGE-2 and ROUGE-L score is calculated for performance matrix for this model. Sentence length is fixed as 6 of document for testing purpose.They used pre-trained GloVe vectors and evaluate with pre-defined summary. In case of this thesis, word-embedding has to be developed using Nepali News corpus and for evaluation purpose, reference summaries have to be generated with help of Nepali Subject faculty.

Myanmar News Headlines are generated by using RNN[3]. LSTM encoder-decoder model is applied in RNN model. This model forecasts a single word and it runs recursively for text summarization. This model generates news headlines as form of summary for news article of Myanmar. For training, total 5000 news article were collected of Myanmar news. For vectorization purpose GloVe embedding is used. ROUGE scores were calculated for performance matrix of RNN and seq2seq model. They found that result of RNN model was significantly better than that of seq2seq model. This research is focused on Maynmar News using RNN. This reference is taken for developing word embedding for Nepali News.

Text summarization is performed by using Word2Vector model in Bengali Language [4]. For this, word2vec model for Bengali language is developed. After that preprocessed text was trained with CBOW and Skip-Gram Model and used TSNE for visualizing the word. In other hand, document summarization is performed using word2vec Model [5]. In this research, 20 Newsgroups dataset with 18,846 labeled newsgroup documents were implemented for their model. For experiment, they applied Word2Vec model. In Word2Vec model they used Hierarchical Softmax framework and CBOW scheme as algorithm.They introduced the concept of documents partitions. It is found that some section of documents

which are not related to topics are eliminated. Experimental was found that this model decreases size of target documents. These researcher's contribution is on word2vec model. These references are taken for developing GloVe model in own Nepali news corpus. Xiangdong You (2019) generates summary of given text using Text Rank Algorithm [6]. In this research, each sentence in text is treated as node. The method to examine similarity of sentences following formula was applied.

$$Sim(Si,Sj) = (|[wk|wk \in Si \ \& \ wk \in Sj|)/(log(|Si|)+log(|Sj|)) \quad (1)$$

where Si,Sj represent two sentences, Wk represent words in the sentences. They cyclically calculated the similarity between any two nodes according to above formula. After that in obtained similarity matrix, page rank algorithm was applied to obtained highest TextRank value as summary. These references help for developing model for Nepali text summarization by using text ranking algorithm instead of page rank algorithm.

Researcher Shivam Patel team proposed text summarization model in 2020, April [7]. In this paper, they summarize text document by clustering its contents using topic modeling technique based on latent topics. After that extractive summary is generated from each cluster of text document.After that summary generated of each cluster were combined to form a final summary. Also they compared obtained results to other model like seq-to-seq ,Lead-3,TextRank by calculating ROUGE-1,ROUGE-2 and ROUGE-L. They have used clustering method for summarization but in this thesis, GloVe model is trained with Nepali text and for calculation of ROUGE value, summary of respective documents have to be developed by taking reference of Nepali Subject faculty. Abstractive Text summarization is implemented in 2019[8]. In this paper they used attention based Encoder-Decoder LSTM model to generate summary using "Amazon Fine Food Reviews" dataset. For word vectorization GloVe embedding is used. After that for encoding of input text, forward encoder and backward encoders are used. Output of encoder is feed to decoder process to produce English word. Local attention model is implemented to pay attention on context vector. For quantitative analysis ROUGE-1 and ROUGE-2 values are evaluated.

For generating Nepali news headlines, Encoder Decoder model[9] is used. In this model, GRUs based encoder-decoder is implemented on Nepali news article. For vectorization of input word tokens, FastText embedding model is used which is trained on Nepali news corpus. Result is evaluated by calculating BLEU score metric against human generated headlines. BLEU scores are generated on range between 0.18 to 1. It seems that this model is quite successful for generation Nepali news headlines.

Abstractive Text Summarization model is implemented using Sequence to sequence RNN [10]. The researcher's team uses a state-of-the-art Attentional Encoder-Decoder RNN architecture. For summarization, Novel models are proposed for showing an additional improvement in the performance instead of using a machine translated based model. In this model, a bidirectional GRI-RNN encoder and a unidirectional GRU-RNN decoder along with an attention mechanism and a soft-max layer are used. For identifying the key concept and key entities, named-entity tags, POS and TF-IDF statistics of words are collected. Gigaword corpus is used as a corpus and models are trained on this dataset. ABS and ABS+ models are used for comparison of model performance. Researcher Jianpeng Cheng teams researched on extractive text summarization using attention based neural network[11]. This research focuses on summarization of single document by extracting sentences or words. In this, attention based encoder is applied using hierarchical document for extracting summary. In this paper, continues sentence features and Neural Network are discussed for data driven approach. Attention helps for selecting the input words. This model summarizes from multiple sentences instead of single sentences of individual sentence. After that decoder selects expected output from text document.

3. Methodology

For Extractive based Nepali text summarization, we have implemented two methods as Text Ranking[12] and LSTM method. For this purpose, Nepali News articles are collected as a corpus. Preprocessing is applied in this collected corpus to generate single space-separated words. After that this preprocessed corpus is trained with GloVe embedding model. This embedding is used to generate the vector value of words present in Nepali text while using it for text summarization. so we have described methodology for text summarization in this section. This section

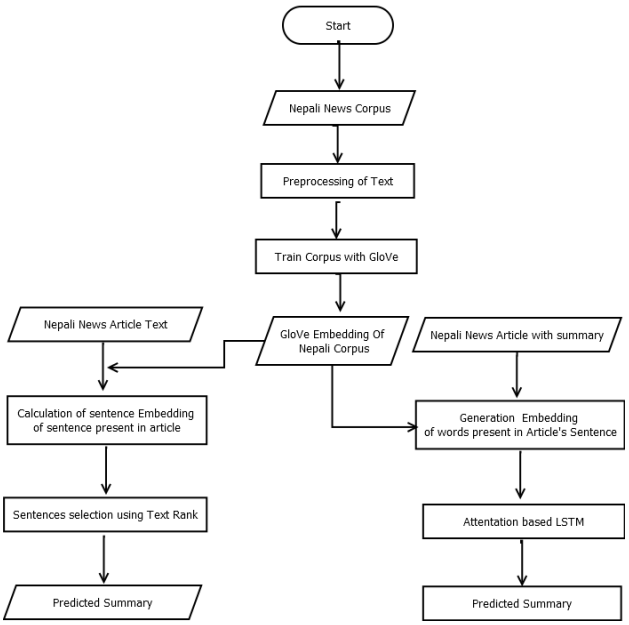

Figure 1: Overall methodology

mainly consists of following sub section:

3.1 Data Collection and Train corpus with GloVe

3.1.1 Data set Collection for word embedding

We have collected total 116K news article from different Nepali news article portal. There are total 14 categories from where articles are collected. Categories are automobiles, diaspora, economy, employment, entertainment, health, international, national, opinion, politics, society, technology and tourism.

3.1.2 Preprocessing of text document

Data preprocessing is a technique to generate useful, efficient and understandable format of data from collected raw data. In this process different noisy information like stop word, extra white space, special characters are removed. So it helps for cleaning unwanted information. Due to which efficiency of model performance is increased. In data preprocessing, white space, numbers and special symbol are removed. After that tokenization is performed and stop words are removed.

3.1.3 Constructing word –word co-occurrence matrix

After preprocessing of given corpus is completed, co-occurrence matrix of preprocessed data is constructed with window size 10, 12 and 15

3.1.4 Trains Corpus with GloVe

The GloVe model provided by Stanford University is trained on the non-zero entries of global word-word co-occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus. In this model, word-word co-occurrence matrix is developed. This model develops matrix by letting X. GloVe explains that, let

$$X_i = \sum_k X_{ik} \tag{2}$$

be the number of times any word appears in the context of word i. Finally, let

$$P_{ij} = P(j/i) = X_{ij}/X_i \tag{3}$$

be the probability that word j appears in the context of word i. Populating this matrix requires a single pass through the entire corpus to collect the statistics. For this, Preprocessed Nepali News corpus is used for training. Vector length is chosen to be 50,100, 200 and 300. After that training, word embedding is obtained.

3.2 Text summarization using Text Ranking

Text Ranking method is unsupervised method for extractive text summarization. It is inspired from Page Rank method. For text summarization average word embedding of sentences of given text is calculated using GloVe word embedding after preprocessing. After that similarity matrix is generated using cosine similarity. Formula for cosine similarity is

$$Similarity(x,y) = x.y/||x||||y|| \tag{4}$$

Where x and y are two sentences with their corresponding average word embedding. Obtained similarity matrix is feed to graph for generating graph. Text Ranking method uses this graph to calculate ranking score. For this purpose, we have proposed damping factor as 0.85 and convergence threshold value as 0.0001. The application of graph-based ranking[12] algorithms to natural language texts consists of following main steps: i. Identify the text units that best define the work at hand and add them to the graph as vertices. ii. Identify the relationships that connect these text units and use them to build edges between the graph's vertices. Edges can be weighted or unweighted, and they can be directed or undirected. iii. Iterate the graph-based ranking algorithm until it reaches a satisfactory conclusion. iv. Vertices are sorted by their final score. For

ranking/selection decisions, use the values associated with each vertex. After that total 40% of ranked value is determined as summary. For performance matrix, calculation of ROUG-1 and ROUGE-2 scores are calulated between system generated summary and reference summary. Reference summary is generated with help of Nepali literature teacher.

3.3 Text summarization using attention based LSTM model

We have used LSTM model in supervised method. So we have collected Nepali news article with their reference summary. For word embedding, already trained GloVe embedding is used. For this method,Total 5975 Nepali news article are collected. Data are collected from entertainment, business and sport categories.Preprocessing is applied in this both text and summary. Tokenization of articles are performed after text preprocessing. For this process maximum length of article is taken as 300 and summary is taken as 30. After that embedding matrix is generated by using pre-trained GloVe embedding of Nepali news corpus with window length 15 and vectorization size 300. After that attention based LSTM model is trained with these data. In LSTM model, 3 stacked LSTM is applied.Finally, this model is trained on 90 data pairs that means text and summary. Model is validated on 10 of data. Total 20 epochs are used with batch size 1024.For evaluation purpose ROUGE1 and ROUGE2 value is calculated.

4. Experiments and Result

Our model is trained using the steps discussed in methodology section. While training Nepali news corpus, it generated total 179954 word vectors.While training with GloVe embedding total 100 iteration is peroformed,and value of initial learning rate is set as 0.005 While using Text Ranking method for Extractive text summarization we have used total 13 sentences text as sample for text summarization. For this purpose, we have used GloVe embedding with window size as 10,12,15 and vector size as 200 and 300. For performance matrix, we calculated ROUGE-1 and ROUGE-2 score as follows where W denotes window size and V denotes vector size

Table 1: ROUGE score using Text Rank model

SN	GloVE used	ROUGE-1	ROUGE-2
1	W-15 and V-200	0.8381	0.8033

For text summarization using attention based LSTM, we have used total 3 stacked LSTM model. For training given text with summary, GloVe embedding with window size 15 and vector size 300 is used vectorization process. This model is trained on 90 percentage data pairs that means text and summary. Model is validated on 10 percentage of data. Total 20 epochs are used with batch size 1024. This model gives Figure 2 as a result. After that for performance

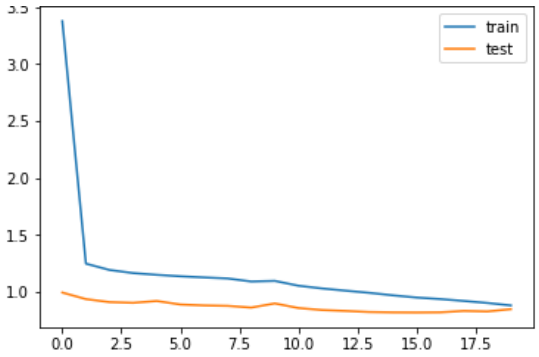

Figure 2: LSTM Training Loss vs Validation Loss

matrix, we calculated ROUGE-1 and ROUGE-2 score as follows:

S.N	System Generated Summary	Reference Summary	ROUGE-1 score	ROUGE-2 score
1	दूरसञ्चार बक्यौता उठाउन	दूरसञ्चारको बक्यौता उठाउन निर्देशन	0.857	0.799
2	खडेरी पीडित गाउँपालिका क्षतिपूर्ति	खडेरी पीडितलाई गाउँपालिकाले क्षतिपूर्ति दिने	0.88	0.857
3	क्रिटिक्स अवार्ड कालो पोथी	क्रिटिक्स अवार्ड "कालो पोथी" लाई	0.88	0.857
4	रियल म्याड्रिड लिग	रियल म्याड्रिड च्याम्पियन्स लिग फाइनलमा	0.749	0.333

Figure 3: ROUGE score for LSTM Result

5. Conclusion

In this paper, we have trained Nepali news corpus with GloVe embedding with different window size as 10,12,15 and vector size 100,200 and 300. For extractive text summarization Text Ranking and attention based LSTM model is used. Total 116K news article from different Nepali news portal are collected. There are total 14 categories from where articles are collected to train with GloVe embedding. ROUGE-1 and ROUGE-2 scores are calculated for evaluation matrix. These result shows effectiveness of our model. In future result of both Text Ranking and LSTM has to be compared with same data. Data size has to be increased to train with GloVe embedding.

Acknowledgments

We would like to acknowledge to the Electronics and Computer Engineering Department, Paschimanchal Campus for direct and indirect guidelines. Also we want to thank all Nepali News portal for facilitating to collect news dataset.

References

[1] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[2] Aditya Jain, Divij Bhatia, and Manish K Thakur. Extractive text summarization using word vector embedding. In *2017 International Conference on machine learning and data science (MLDS)*, pages 51–55. IEEE, 2017.

[3] Yamin Thu and Win Pa Pa. Generating myanmar news headlines using recursive neural network. In *2020 IEEE Conference on Computer Applications (ICCA)*, pages 1–6. IEEE, 2020.

[4] Sheikh Abujar, Abu Kaisar Mohammad Masum, Md Mohibullah, Syed Akhter Hossain, et al. An approach for bengali text summarization using word2vector. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2019.

[5] Zhibo Wang, Long Ma, and Yanqing Zhang. A novel method for document summarization using word2vec. In *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 523–529. IEEE, 2016.

[6] Xiangdong You. Automatic summarization and keyword extraction from web page or text file. In *2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET)*, pages 154–158. IEEE, 2019.

[7] Kastriot Kadriu and Milenko Obradovic. Extractive approach for text summarisation using graphs. *arXiv preprint arXiv:2106.10955*, 2021.

[8] Puruso Muhammad Hanunggul and Suyanto Suyanto. The impact of local attention in lstm for abstractive text summarization. In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 54–57. IEEE, 2019.

[9] Kaushal Raj Mishra, Jayshree Rathi, and Janardan Banjara. Encoder decoder based nepali news headline generation. *International Journal of Computer Applications*, 975:8887.

[10] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[11] Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*, 2016.

[12] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.