IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol-6, Special Issue-3, Apr 2018 E-ISSN: 2347-2693

A Review of Text Summarization using Gated Neural Networks

Touseef Iqbal[1*], Abhishek Singh Sambyal[2], Devanand[3]

[1*]Computer Science & IT, Central University of Jammu, Central University of Jammu, Jammu, India
[2]Computer Science & IT, Central University of Jammu, Central University of Jammu, Jammu, India
[3] Computer Science & IT, Central University of Jammu, Central University of Jammu, Jammu, India

[*]*Corresponding Author:* touseefiqba549@gmail.com, *Tel.: +91-9596506403*

Available online at: www.ijcseonline.org

***Abstract*—**There is an enormous amount of information available in the form of documents, articles, links, webpages, etc., which can't be read completely until effectively summarized. Different procedures are effectively used to separate the imperative information from data to produce summary. This paper gives a brief description of text summarization and deep learning approach called Recurrent Neural Networks (RNNs). Recent advances in Deep RNN methods like Sequence to Sequence, Generative Adversarial Networks, etc. show remarkable results for text summarization. Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are discussed to overcome the problem of Vanishing or Exploding Gradient.

***Keywords*—** RNN, Sequence to Sequence, Vanishing gradient, LSTM, GRU, Deep Recurrent Generative Decoder, GANs.

I. INTRODUCTION

People are repulsed with a lot of online data. Due to the increase in large amount of online data, fast and convenient summarization has become more important. Text summarization clarifies the most relevant information from the source document to provide a compressed version of specific task. There is a need to summarize in various domains like news articles, emails, webpages etc., as manually summarizing large documents is very difficult for the people. Text summarization aims to produce the short and condensed form of original documents [1]–[4]. The research on summarization of scientific documents occurred early and was proposed by Luhn in 1958, on text positions by Baxandale in 1958 and on key phrases by Edmundson in 1969. Since then different researches were presented that have worked on other areas as well, largely on newswire data. Literature based on document summarization is in huge amount that mainly focuses on extractive summarization. Old techniques for Extractive summarization are comprehensively arranged into Greedy methodologies for example, in 1998 Carbonell and Goldstein's diagram based methodologies (Radhev and Erkan 2004) and consistent enhancement approach (e.g McDonald 2007) [5]–[8]. The text summarization accomplishes great interest in Natural Language Processing (NLP). NLP is the way for computers to calculate, recognize and evolve the meaning of human language in a brilliant way. Automatic Text summarization frameworks can be sorted into a few unique types (Nenkova and McKeown, 2012; Saggion and Poibeau, 2013). The distinctive measurements of Text Summarization can be largely classified on input type (single or multi-record), a reason (Generic, Specified by Domain or based on the query) and output type (extractive or abstractive). Mainly text summarization is categorized into two distinct ways, Extractive and Abstractive. Extractive summarization is the determination of existing subset of words or numbers from a little information to make the summary while as Abstractive text summarization is the building of internal semantic representation about the data [4], [9]–[11]. Abstractive text summarization needs deeper analysis of data since it utilizes semantic techniques to investigate and delineate the content and finds the ideas and clarifications to portray it, by producing another shorter content which depicts the most vital data from the first content record [3].

II. BACKGROUND OF RECURRENT NEURAL NETWORKS (RNN)

A unique sort of Neural Network that came in 1980s called Recurrent Neural Networks (RNNs) which has revolutionized the field of NLP [12]. RNNs are the intense group of connectionist models that get time progression with cycles of Graph [13]. The generation of summary in a sequence needs some preceding information about the data. So, the Recurrent Neural Networks works well for the tasks showing data dependency [14]. In addition to data dependency this network works well for variable length input. The ability of RNN to recollect previous data has made it a dynamic sequence modeling tool. Fig. 1 shows unfolded RNN.

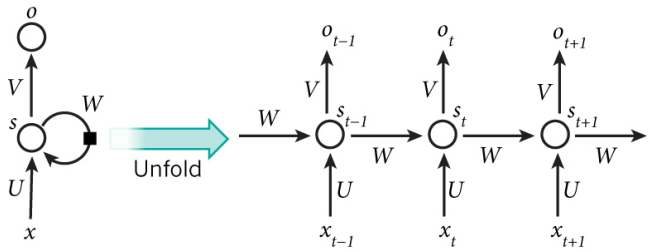

Figure 1. Figure of unfolded RNN.

The unrolling of Recurrent Neural Network means that it is being written for the complete sequence. For the sequence of sentence containing 10 words the network will be unrolled or 10 layers, each for one word. Here are the formulas that carry out the computation: $S_t = F(U * X_t + W * S_{t-1})$, $O_t = \sigma (W_t * S_t)$ where input at time t is X_t and S_t represents hidden states for time step t. In addition, St can be thought as a memory of the network which has stored the information calculated in previous time steps. During processing, the RNN has strength of passing the information element wise in a selective manner over sequence steps. Sequence of elements which are dependent are represented as the input and/or output. In addition, RNN models sequential and time dependencies simultaneously on different scales. The applications of RNN can be found in many domains such as Sentiment analysis, Pattern recognition, Speech recognition and Synthesis Language Modeling [13], [15].

III. RELATED WORK

The recent techniques for text summarization are based on sequence to sequence learning like Attentional feed forward network given by Rush, Chopra and Weston (2015), RNN based encoder-decoder given by Nallapati, Zhou and Xiang (2016) [2] etc. Extractive procedures are more affordable and create syntactically and semantically redress sentences more often [2]. The recent techniques that made use of deep learning techniques for text summarization are:

A. Summarunner: RNN based Sequence model for text summarization

The extractive summarization is presented as a sequence characterization, where each sentence is visited sequentially in original arrangement and the decision is made for the sentence being as part of summary or not. The GRU is being used as the essential piece of this model. This model involves a two-layer bi-directional GRU-RNN in which main layer runs at the word level, and produces hidden state depictions for each word positions progressively, in perspective of the present word embedding's and the already hidden state, another RNN at the word level that keeps running backward from the last word to the initial is utilized. The combination of both forward and reverse RNNs is presented as a bidirectional RNN. The model in like manner contains another bi-directional RNN that continues running at the sentence-level and takes the average pooled, coordinated hidden states of bi-directional word-level RNNs as input, to encode the representation of sentences in the report, hidden states of another layer RNN are utilized. The portrayal of entire report is then shown as non-linear change in the average pooling of the connected hidden states of the bi-directional sentence-level RNN. For the purpose of classification every sentence is revisited sequentially in the sequence where a logistic makes the decision about the sentence to be chosen as a part of summary [2].

B. Attention-based Deep Recurrent Neural Networks

The collection of frameworks having various layers and individual units are examined, beginning with straight forward feed forward neural network (FFN) benchmark. At that point preparation of a 4-layer RNN in which GRU units are being presented as encoder cells with and without attention and 4-layer RNN with LSTM units with and without attention. The 3-layer RNN gets trained after removing one of the layer with or without attention in which GRUs or LSTMs are presented as cells respectively. Moreover, the trick with a 2-layer bidirectional RNN picked up a superior comprehension of productivity of models. It has been found that the parameters having hidden size of 512 and batch size of 128 perform better. RMSprop is utilized as the preparation updater, which separates the learning rates by an exponentially rotting average of squared gradients to determine Adagrads drastically lessening learning rates. Mean Square Error (MSE) is utilized as loss function [16].

C. Vanishing/Exploding Gradient, LSTM, GRU

1) *Vanishing/Exploding Gradient:* Recurrent Neural Network utilizes back propagation calculation for training as it is connected to each time stamp, we call it back propagation through Time (BPTT). RNN has an issues of adopting long term dependencies [17]. Through back propagation, we really endeavor to figure out the error and with that error, we compute the Gradient that is change in error as for the change in weight. There is an issue when the value of Gradient is less than one. Around then there is no weight updation in the network or there is no learning in the RNN. This is called Vanishing Gradient. Another issue which Recurrent Neural Network indicates is of exploding Gradient that is the point at which the value of Gradient is more prominent than one, so the new weight is entirely unexpected from previous weight. So again there is no learning in the network. Various techniques were proposed for mitigating the vanishing/exploding gradient issue, for example, utilization of second order enhancement strategies. (e.g., sans hessian streamlining (Martens and Sutskever, 2011)), particular training schedules (e.g., Greedy layer wise traning (Schmidhuber, 1992; Hinton et al., 2006; Vincent et al., 2008)) and extraordinary weight introduction strategies when training on both plain FFNs and RNNs (Glorot and Bengio, 2010; Mishkin and Matas, 2015; Le et al., 2015; Jing et al.,

2016; Xie et al., 2017; Jing et al., 2017). [18]. Gated Neural Networks additionally help to diminish the issue by acquainting Gates which control the stream of data through the network over the layers or successions. Examples are Long Short-Term Memories, Gated Recurrent Units and so on. [18].

2) *Long short term memory (LSTM):* LSTM has shown an intense and extensible model for various learning issues related to sequential data that were not counteract by previous models. LSTMs are commonly versatile at handling long-term dependencies. They overcome the enhancement challenges by maintaining the constant error flow throughout the network that pushes the high advancement for some issues including language displaying and interpretation, acoustics demonstrating of discourse, discourse blend, protein auxiliary structure forecast, investigation of sound and video information among others. LSTM was first proposed by Hochreiter and Schmidhuber (1997), from that point onwards a number of minor changes to the first LSTM have been made. The execution of LSTM utilized nowadays is Graves (2013) [19]. It works better for the undertakings including long time slacks contrasted with conventional RNN [15]. LSTM is an uncommon sort of RNN that contains extraordinary units called memory blocks in recurrent hidden layer [20]. These memory blocks contain memory cells with self-associations which store fleeting condition of the network. In spite of memory cells, it contains unique multiplicative units called Gates that control stream of data [21]. Every memory block in the original architecture contains an input gate and output gate. The input gate contains the stream of information actuation into the memory cell. The output gate controls the output stream of cell activation into whatever remains of the Network. Later the forget gate was added to the memory block. The forget gate scales the internal size of the cell before adding it to the cell [22]. LSTM permits consistent error flow through the system that overcomes the issue of vanishing and exploding gradient [23].

3) *Gated Recurrent Unit (GRU):* Gated Recurrent Neural Network has demonstrated accomplishment for few applications including successive or fleeting information. GRU was first proposed by Cheng et al (2014) to make each repetitive unit adapt to the conditions of various time scales and for Neural Machine interpretation (Bahdanau et al. (2015)). At first, for Robot Reinforcement Learning (Bakker(2001)) [24] GRU was planned. The essential thought behind GRU is to make utilization of a few gates for controlling the stream of data from past strides to the present advances. By applying the gates, mapping starting with one point then onto the next can be learned by any repetitive unit. Like LSTM, the adjustment of stream of data is done by gated mechanisms inside GRU, Furthermore GRU contains separate memory cells [19], [25]. It is less difficult than LSTM and quicker to prepare. The GRU shares same component for calculation yet is less unpredictable than LSTM. It contains just two Gates, update gate and reset gate [22]. The Gated Recurrent Unit has a comparative idea to an LSTM, yet they accelerate training because of design rearrangements [15]. The Recent Techniques used for Text summarization which are utilizing Deep Recurrent Neural Network models, trained on different Data sets and working on different baseline approaches produces phenomenal results as shown in Table I.

Table I: Comparison of Recent Text Summarization Techniques on Recurrent Neural Networks

Ref	Method	Dataset	Evaluation measure	Baseline approach	Results
[27]	Deep Recurrent Generative Decoder	GIGA WORD,DUC-2004,LCSTS	ROUGE-1,ROUGE-2,ROUGE-L	Latent Structural Modelling	R-1,R-2,R-3 for Gigaword=36.2,17.57,33.6, R-1,R-2,R-3 for DUC-2004=31.7,10.75,27.48, R-1,R-2,R-3 for LCSTS=36.9,24.15.34.2
[30]	Generative Adversarial Networks (GANs)	CNN/DAILY MAIL CORPUS	ROUGE-1,ROUGE-2,ROUGE-L	Generator Discriminator	R-1=39.92,R-2=17.65,R-L=36.71
[2]	Attention Based Recurrent Neural Networks	GIGA-WORD CORPUS	ROUGE-1 and ROUGE-2	Feed Forward, A-GRU-4, GRU-3,ALSTM-4,LSTM-3	For Feed Forward R-1=9.31&R-2=1.77, ForA-GRU-4 R-1=13.64 & R-2=4.23, For A-LSTM-4 R-1=20.52 and R-2=5.48 & BRAT 2015 = (0.78, 0.65, 0.75)
[16]	Sequence classification approach	CNN/Daily Mail Corpus	ROUGE-1,ROUGE-2, ROUGE-L	Lead-3 Model	R-1= 39.6, R-2=16.2, R-L= 35.2

D. Text summarization with deep Recurrent Generative Decoder

For the sequence to sequence learning, the neural network is structured as Encoder-Decoder in which the input is taken as a sequence like $X = X_1, X_2, X_3 \ldots X_n$ representing to the source content. The embedded word X_t is introduced arbitrarily and learns from optimization process. $Y = Y_1, Y_2, Y_3 \ldots Y_n$ represents output sequence for created abstractive summaries. GRU (Cho et al., 2014) is utilized as the fundamental succession demonstrating part for the encoder and the decoder. To display the inactive Network, authentic conditions on the latent variables of Variational Auto-Encoders (VAEs) [26] are included. To distill the latent structures inferred for objective summaries of training data a deep recurrent generative decoder (DRGD) is proposed. Now the decoding of abstractive summaries depends on the discriminative deterministic factors H and the generative latent structural data Z. The latent structure modeling system can be seen as a sequence generative model which gets isolated into two sections: interpretation (VEs) and formation (VDs). Input for the primary VAEs just contain the detected variable Y_t, the Variational-Encoder can outline to an inactive Variable $Z \in R * K_2$ utilized for recreating input. In the Sequence Decoder Component, the previous latent structure data is considered for building a powerful portrayal of next state to text summarization task [27].

E. Text summarization with Generative Adversarial Networks (GANs):

The preparation of two models is done all the while in an adversarial way: a generative model G and Discriminative model D [28]. The pre-preparing of the generative model is done by creating summaries given the source content. At that point, the discriminator is pre-prepared by giving positive cases from the human created summaries and the negative illustrations delivered from the pre-trained generator. Now after pre-training, alternatively the discriminator and generator are trained. For the generator model, it takes the input as $X = W_1, W_2, W_3 \ldots W_n$ and predicts the summary $Y = Y_1, Y_2, Y_3 \ldots Y_m$ in which length of the predicted summary is N. The Bi-Directional LSTM encoder is utilized to change over this info content X into a sequence of hidden states $H = H_1, H_2, H_3 \ldots H_n$ for time step t, LSTM decoder is then used for the calculation of hidden states, S_t of the decoder and a setting vector C_t. Generator parameters are collectively represented by Θ. The C_t is joined with S_t, and sustained through the completely associated layer and softmax layer to create the predicting word from the objective vocabulary at each time step t. Now the discriminator is a binary classifier and plans differentiating the input sequence as originally created by humans or incorporated by machines. The input sequence is encoded with Convolution Neural Network (CNN) [29] and for text classification it has shown great success. We utilize various channels with shifting window sizes to get diverse features and after that apply a max-over-time pooling activity over these features. These pooled features are passed to a completely associated softmax layer whose output is likely to be unique [30].

IV. CONCLUSION AND FUTURE SCOPE

Summarization of text in the understandable meaningful information is an intensive task. And due to the infinite growth of Internet, digital data is enormously increasing in the form of articles, documents, mails, news feed, blogs etc. To read such amount of data and to manually summarize is a very difficult and time-consuming approach. Various automated text summarizing tools are designed which can compress the data and produce meaningful information. These are called text summarization tools and are highly used and recommended. The techniques used for summarization of text are RNN based sequence model, deep recurrent generative decoder (DRGD) and GANs. RNN based sequence model is the extractive text summarization tool and has achieved better performance than the state-of-the-art. DRGD urges the summarization tools to include the latent structure information of summary to improve the quality of generated summaries. GANs also perform abstractive text summarization with the use of reinforcement learning to achieve the highest rouge score. RNN faces the problem of vanishing/exploding gradient where the weight updations show abnormal behavior and degrade the performance of model. This is a serious issue and is overcome by LSTM and GRU. LSTM architectures prevent any changes to the cell by simply turning off the gates and hence learn the long term dependencies. And when the gate is on, it updates the content of cell by an average value of present and previous stored value. The text summarization techniques so far have produced impressive results and is expected to improve with the further advancements and innovations.

REFERENCES

[1] D.K. Gaikwad, C.N. Mahender, *"A Review Paper on Text Summarization"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue. 3, 2016.

[2] R. Nallapati, F. Zhai and B.Zhou, *"SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents"*, In AAAI pp. 3075-3081, February 2017.

[3] V. Gupta and G.S. Lehal, *"A survey of text summarization extractive techniques"*, Journal of emerging technologies in web intelligence, 2(3), pp.258-268, 2010.

[4] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D.Trippe, J.B. Gutierrez and K. Kochut, "*Text summarization techniques: A brief survey*", arXiv preprint arXiv:1707.02268, 2017.

[5] D. Das and A.F. Martins, *"A survey on automatic text summarization"*, Literature Survey for the Language and Statistics II course at CMU, 4, pp.192-195, 2007.

[6] A.S. Asa, S. Akter, M.P. Uddin, M.D. Hossain, S.K. Roy and M.I. Afjal, *"A Comprehensive Survey on Extractive Text Summarization Techniques"*, American Journal of Engineering Research, Vol. 6, Issue-1, pp.226-239, 2017.

[7] N. Ramanujam and M. Kaliappan, *"An Automatic Multidocument Text Summarization Approach Based on Nave Bayesian Classifier Using Timestamp Strategy"*, The Scientific World Journal, 2016.

[8] A. Agrawal and U. Gupta, *"Extraction based approach for text summarization using k-means clustering"*, Int. J. Sci. Res. Publ.(IJSRP), 4(11), 2014.

[9] A. Nenkova and K. McKeown, *"A survey of text summarization techniques"*, In Mining text data pp.43-76. Springer, Boston, MA, 2012.

[10] S. Verma and V. Nidhi, *"Extractive Summarization using Deep Learning"*, arXiv preprint arXiv:1708.04439, 2017.

[11] M. Gambhir and V. Gupta, *"Recent automatic text summarization techniques: a survey. Artificial Intelligence Review"*, 47(1), pp.1-66, 2017.

[12] K.M. Tarwani and S. Edem, *"Survey on Recurrent Neural Network in Natural Language Processing,"* International Journal on Emerging Trends in Technology, Vol. 48, 2017

[13] Z.C. Lipton, J. Berkowitz, and C. Elkan, *"A critical review of recurrent neural networks for sequence learning"*, arXiv preprint arXiv:1506.00019, 2015.

[14] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. *"Speech recognition with deep recurrent neural networks"*, In Acoustics, speech and signal processing (icassp), ieee international conference on, pp.6645-6649. IEEE, 2013.

[15] V. Khomenko, O. Shyshkov, O. Radyvonenko and K. Bokhan, *"Accelerating recurrent neural network training using sequence bucketing and multi-GPU data parallelization. In Data Stream Mining & Processing (DSMP)"*, IEEE First International Conference on pp.100-103. IEEE, August, 2016.

[16] Yu, Hujia, Chang Yue and Chao Wang, *"News Article Summarization with Attention-based Deep Recurrent Neural Networks"* Stanford Natural Language Processing Group, Stanford University, pp.2746634, 2016.

[17 Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černocký and Sanjeev Khudanpur, *"Recurrent neural network based language model"*, In Eleventh Annual Conference of the International Speech Communication Association, 2010.

[18] Y. Hu, A. Huber, J. Anumula and S.C. Liu, *"Overcoming the vanishing gradient problem in plain recurrent networks"*, arXiv preprint arXiv:1801.06105, 2018.

[19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, *"Empirical evaluation of gated recurrent neural networks on sequence modeling"*, arXiv preprint arXiv:1412.3555, 2014.

[20] H. Sak, A. Senior and F. Beaufays, *"Long short-term memory recurrent neural network architectures for large scale acoustic modeling"*, In Fifteenth annual conference of the international speech communication association, 2014.

[21] Sak, Haşim, Andrew Senior, and Françoise Beaufays, *"Long short-term memory recurrent neural network architectures for large scale acoustic modeling"*, In Fifteenth annual conference of the international speech communication association, 2014.

[22] R. Dey and F.M. Salem, *"Gate-variants of gated recurrent unit (GRU) neural networks"*, arXiv preprint arXiv, 2017.

[23] S. Hochreiter and J. Schmidhuber, *"Long short-term memory"*, Neural computation, 9(8), pp.1735-1780, 1997.

[24] C.S. Saranyamol and L. Sindhu, *"A survey on automatic text summarization"*, Int. J. Comput. Sci. Inf. Technol, 5(6), pp.7889-7893.s, 2014.

[25] A. Graves, M. Liwicki, S. Fernndez, R. Bertolami, H. Bunke and J. Schmidhuber, *"A novel connectionist system for unconstrained handwriting recognition"*, IEEE transactions on pattern analysis and machine intelligence, 31(5), pp.855-868, 2009.

[26] D.P. Kingma and M. Welling, *"Auto-encoding variational bayes"*, arXiv preprint arXiv:1312.6114, 2013.

[27] P. Li, W. Lam, L. Bing and Z. Wang, *"Deep Recurrent Generative Decoder for Abstractive Text Summarization"*, arXiv preprint arXiv:1708.00625, 2017.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *"Generative adversarial nets. In Advances in neural information processing systems"*, pp.2672-2680, 2014.

[29] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, *"Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems"*, pp.2172-2180, 2016.

[30] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu and H. Li, *"Generative Adversarial Network for Abstractive Text Summarization"*, arXiv preprint arXiv:1711.09357, 2017.

