Headline Generation from News Articles using Recurrent Neural Network

Sumana Basu
McGill University
260727568
sumana.basu@mail.mcgill.ca

Abstract

Automatic headline generation is an abstractive text summarization problem that maps input sequence of words from news article to sequence of words constructing headline. In this paper we replicated global attention based multilayer LSTM stacked encoder-decoder model. Similar model proved to be successful in neural machine translation and data driven sentence summarization. We trained the model on a news dataset of 17,828 articles with 100 words of each article for 250 epochs on a GPU. Average BLEU score of predicted headlines is 0.6255. This proves that sequence to sequence mapping models are can be trained with much smaller corpus than Gigaword, though larger corpus undoubtedly generates better headline.

1 Introduction

Headline of a news article is a short statement that gives the reader a general idea about the main contents of the story it entitles (Colmenares et al., 2015). It should be brief, accurate and expressive. Automatic generation of headline that satisfies the above mentioned criteria is a demanding area of research due to its possible application in generating headlines as notifications for news applications on hand held devices.

Headline generation is a sub-task of text summarization. But since headlines are summaries shorter than a single sentence, extractive text summarization techniques cannot be used for headline generation. An abstractive summarization technique, which involves generating novel sentences for summary must be used to solve the headline generation problem.

1.1 Problem Formulation

Given an input sequence of words $x_1, x_2, ..., x_T$ defined over a fixed vocabulary V of size $|V|$, our aim is to build an automatic headline generator that takes $\boldsymbol{x}$ as input and generates an output sequence $y_1, y_2, ..., y_{T'}$ with $T' < T$.

1.2 Background

Recurrent neural networks are specialized for processing such sequence of values (Goodfellow et al., 2016). Unlike feed forward neural network, Recurrent Neural Network (RNN) can handle input and output sequence of variable dimensionality. It makes RNN a suitable candidate for processing and generating variable length sentences. RNN also maintains its internal state while reading a sequence of inputs. This enables RNN to keep track of the context. But regular RNN faces vanishing gradient problem (Hochreiter, 1998), which can be tackled by Long Short-Term Memory (LSTM).

Sutskever et al. proposed an encoder-decoder based architecture which uses a multilayered LSTM to map input sequence to an intermediate context vector and another deep LSTM to decode the output sequence from that vector. But this model treats each word equally. In the context of Neural Machine Translation, which is also a sequence to sequence mapping problem, (Luong et al., 2015) introduced the concept of global attention model which puts more importance on some words than the other.

Similar model has been used by (Rush et al., 2015) and (Lopyrev, 2015) for sentence summarization task.

1.3 Hypothesis

Inspired by Attention-based Neural Machine Translation model of (Luong et al., 2015), in this project we replicated their mutilayer LSTM based encoder-decoder model with global attention to test its applicability and performance in headline generation. (Rush et al., 2015) have created a similar model and trained it on Gigaword corpus (Graff and Cieri, 2003) with an objective to summarize single sentence. Gigaword (Graff and Cieri, 2003) contains around 9.5 million news articles sourced from various domestic and international news services over the last two decades. We conducted experiments described in this paper on a much smaller dataset of news articles. As input, our model takes first 100 words of each article, assuming main message is conveyed by first few words. We evaluated quality of headlines using BLEU. Motivation behind this project is to assess effect of corpus size and input sequence length on performance of attention based encoder decoder model.

2 Method

2.1 Data Description

We collected 17,828 articles and corresponding headlines from *The Hindu* which is an English newspaper circulated in India. Punctuation and non-ascii characters were removed from the data.

2.2 Theory

For sequence learning, input sequence is mapped to a context vector which is then mapped to output sequence. This is nothing but modeling the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ of generating target sequence $(y_1, y_2, ..., y_{T'})$ from source sequence $(x_1, x_2, ..., x_T)$. The basic Encoder-Decoder architecture consists of 2 steps :

1. **Encoder :** computes representation $\boldsymbol{s}$ for each input sequence $(x_1, x_2, ..., x_T)$.

2. **Decoder :** generates target word y_t at each timestamp. It decomposes the conditional probability as :

$$
\begin{aligned}
& p(y_1, y_2, ..., y_{T'}|x_1, x_2, ..., x_T) \\
& = \prod_{t=1}^{T'} p(y_t|y_1, y_2, ..., y_{t-1}, \boldsymbol{s})
\end{aligned} \tag{1}
$$

(Sutskever et al., 2014) stacked multiple layers of RNN with LSTM units in both encoder and decoder. In case of RNN probability of decoding each word y_t can be computed as:

$$p(y_t|y_1, y_2, ..., y_{t-1}, \boldsymbol{s}) = softmax(g(h_t)) \tag{2}$$

$$g(h_t) = \tanh(Wx_t + Uh_{t-1} + b) \tag{3}$$

where, W = input to hidden weight matrix
U = hidden to hidden weight matrix
b = bias vector

Training objective can be formulated as:

$$J_t = \sum_{(x,y) \in D} -logp(y|x) \tag{4}$$

3. **Attention :** While generating each word through decoder, at each time stamp t, decoder takes hidden state ($\boldsymbol{h_t}$) of top layer of LSTM as input and generates a context vector $\boldsymbol{c_t}$. (Luong et al., 2015) described a global attention mechanism that considers all hidden state of the encoder while deriving the context vector. By comparing each source hidden state ($\boldsymbol{h_s}$) with the target hidden state ($\boldsymbol{h_t}$), alignment vector $\boldsymbol{a_t}$ is calculated as follows:

$$\boldsymbol{a_t} = \frac{exp(score(\boldsymbol{h_t}, \boldsymbol{h_s})}{\sum_{s'} exp(score(\boldsymbol{h_t}, \boldsymbol{h_s})} \tag{5}$$

Score is calculated using dot product.

2.3 Experiment

We divided the dataset between training and test sets. Training set contains 14,828 articles and corresponding headlines. Based on the most frequent words appearing in articles and headlines of training set, we formed a vocabulary of 40,000 words. We embedded each of these words to continuous space with 100 dimensions using Glove (Pennington et al., 2014). For each word in the input sequence, we look up the corresponding word embedding. We approximate out of vocabulary words by its closest match inside vocabulary, if the cosine similarity between these two words is greater than 0.5. Training data has 1,10,687 different words of which 40,000 were inside vocabulary and rest were out of vocabulary. 33,723 out of vocabulary words were substituted using Glove similarity. Rest of the words were treated as $\langle unk \rangle$.

We used deep LSTMs with 3 layers and 512 nodes at each layer. We did not use regularisation. Initially a dropout of 20% has been used to avoid overfitting. But since it did not show any significant improvement, the final model does not use any dropout. Encoder generates a summarization vector which is fed into the decoder. Decoder generates headline, one word at a time, with a variation of beamsearch (Sutskever et al., 2014) sampling has been used to improve the quality of the generated text samples. Softmax activation layer and categorical cross-entropy loss has been used. Word generated at each time-stamp is fed into the decoder while generating the next word. Adam optimizer (Kingma and Ba, 2014) with initial learning rate of 0.0001 has been used. We trained the model for 250 epochs until loss stabilized.

3 Results

3.1 Relevance & Readability

During training, for the first 20 epochs predicted headlines were not related to the actual article. After the initial phase, model started to generate relevant headlines. However, length of headlines were very short. After 100 epochs grammatically correct headlines of considerable length started to generate. Few of the headlines generated during training and testing have been enumerated against human generated titles in Table1.

- Some of the predicted headlines exactly matched with human generated titles (example 1).

- Some predictions highlight facts which were not present in the original input article. This is because of the context the model has learnt from other articles of similar topic.
  (example 2 - Rohit Vemula has correctly been identified as 'Dalit'. But that fact was not present in the article.
  example 4 - article does not state that biker burnt to death)

- Some outputs are not relevant to the topic, but they share common words with the article.
  (example 2 is about Hyderabad, but not Mumbai
  example 3 - 'skill' word is common between headline and article. But prediction is unrelated to the topic)

3.2 Evaluation

Due to expensive and extensive nature of human evaluation, we used BLEU (Papineni et al., 2002) to evaluate predicted headlines. For each of the 3000 test items, we compared each candidate prediction with corresponding reference human generated headline and calculated BLEU score.

- Average BLEU score is 0.6255

- 4.57% of the test articles had BLEU score of 1

- For 0.80% of the test articles, BLEU score was zero

- Model could not generate any headline for 0.3% articles.

4 Discussion & Conclusion

In this paper we trained an encoder-decoder LSTM network to generate headlines over a corpora of news articles taken from a leading Indian newspaper. Rush et al. trained similar model on Gigaword corpus, however the dataset we have chosen is much smaller in size. While in their experiment they have chosen first sentence of each article where average sentence length was approximately 20-30 words, we on the other hand have used first 100 words of each article to train our model, as the most relevant information regarding a news article can be found in the first paragraph. Hence, we were expecting sensible output even though we trained on a much smaller dataset. Since we trained on a much smaller dataset than Gigaword, our model did not had enough data to create proper context over the articles, but it showed significant improvements.

Article (First 100 words)	**Human Generated Headline**	**Auto Generated Headline**
The record breaking Star Wars opened on Saturday in China where it is far from certain to draw in enough movie goers to knock off Avatar as the world s all time biggest grossing movie Star Wars The Force Awakens is the highest ever grossing film in the North American market where it was released three weeks ago But internationally it still has a long way to go to beat Avatar James Cameron s science fiction movie with blue aliens The international box office of the latest Star Wars movie stands at 1 6 billion compared to the 2 8	Record breaking Star Wars movie opens in China	Record breaking Star Wars movie opens in China
It has been nearly 12 days since protests erupted after 26 year old research scholar Rohith Vemula committed suicide on the University of Hyderabad UoH campus The agitation has been sustained for so long because there is a well orchestrated team comprising several volunteers from the students Joint Action Committee JAC Anyone who has been to the Shopcom area at the university which is the epicentre of protests on the UoH campus will notice the information desk in a corner Managed by students from the communication department it is probably one of the most important organs of the JAC a	Volunteers sustain protest on Hyderabad varsity campus	Dalit scholar's suicide sparks protests in Mumbai
ICICI Bank which has fixed a target of training one lakh underprivileged students by March 2017 is set to achieve it with 60 000 students getting placement after the training The lender will spend Rs 200 crore on social sector projects this fiscal and meet the mandatory two per cent corporate social responsibility spend norms Chanda Kochhar managing director and chief executive officer of ICICI Bank said on Wednesday She opened the bank s 22nd skill development academy which was inaugurated by Maharashtra Chief Minister Devendra Fadnavis This is the first such centre in Mumbai and fourth in Maharashtra The	ICICI to train one lakh students	Communication skills must for techies
A mother and her 14 year old daughter were killed in a road accident near Moovattupuzha in the early hours of Wednesday Ibrahim 42 his wife Salma 38 and their daughter Mufsina were on a motorcycle when a lorry carrying gas cylinders hit them According to reports Salma was trapped under the lorry while her husband and daughter were thrown off on impact of the collision They were rushed to a nearby hospital The police said that Salma died on the spot Mufsina died later in the hospital police said The condition of Ibrahim is stated to be serious	Mother daughter die in road accident	Biker charred to death in collision

Table 1: Example Predictions

Our experiment generates grammatically correct and context-aware headlines for the provided articles, which proves the efficiency of stacked LSTM with attention mechanism trained over a corpora of text. Due to time and resource constraints, we could not explore deeper LSTM architectures, which coupled with bidirectional LSTM could have provided much coherent output. As evident for all neural network models, more data is the solution for better prediction, thus our model can be easily scaled up to a larger corpus to achieve better efficiency.

Acknowledgments

We would like to thank Ehud Ben-Reuven for providing the baseline LSTM code on which this work has been based on, which in turn was implemented following the paper by Konstantin Lopyrev. We would also like to thank the developers of Keras, Scikit learn and Numpy, popular multipurpose machine learning libraries which were used in this project. We are heavily indebted to Calcul Quebec and Compute Canada for providing GPU resources to run our networks. The operation of the supercomputers used is funded by the Canada Foundation for Innovation (CFI), ministre de lconomie, de la Science et de lInnovation du Qubec (MESI) and the Fonds de recherche du Qubec - Nature et technologies (FRQ-NT).

References

Carlos A. Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. Heads: Headline generation as sequence prediction using an abstract feature-rich space. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (NAACL'15)*, pages 133–142.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep learning. Book in preparation for MIT Press.

David Graff and Christopher Cieri. 2003. *English Gigaword LDC2003T05. Web Download. Philadelphia: Linguistic Data Consortium*.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 6(2):107–116, April.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *CoRR*, abs/1512.01712.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *CoRR*, abs/1509.00685.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.