Report: Abstractive Text Summarization using Seq-2-Seq Modelling

Virender Singh, Durga Madhuri Kottapalli, Laya Karimi chahrogh

March 28, 2021

1 Introduction

With the growth of the internet, we now have access to a wide range of data. But now, we are flooded with lots of information from various sources, such as the press, social media, office emails, etc. But it could be more helpful if there are summaries present for these data. We hope to provide a short and accurate description of the necessary document via this project. There are two main categories for Text summarization: extractive and abstractive summarization methods. We define relevant sentences or phrases from the text and extract only those in extractive summary. In contrast, an abstract description includes sentences or phrases that were not present in the original text.

Our project aims to use sequence-to-sequence modeling to produce an abstractive description of the text. This method, which has two major components: encoder and decoder, is widely used in sentiment classification, Neural machine translation (language translation), and named entity recognition. We'll use Long Short-Term Memory (LSTM) network model for both modules, and we'll work with the design to see what fits better.

2 Related Works

In a proceeding research paper(cited in our references section) about Information Extraction by an Abstractive Text Summarization for an Indian Regional Language, the authors try to combine the information retrieval tools with automatic text summarization and Natural language processing (NLP) is used as their main algorithm. Moreover, In a review(cited in our references section) about Neural network-based abstractive text Summarization models, the authors make comparisons between various neural network models and discuss based on different categories and approaches.

3 Materials and Method

3.1 Dataset and Features

We are using XSum(Extreme Summarization) dataset for implementing abstractive summarization of the text. Under this, we use ag-news-subset. Ag includes more than 1 million news articles. ComeToMyHead is an academic news search engine that has collected news stories from over 2000 news outlets over the course of a year of operation. The news topic classification dataset of AG includes the four largest groups of the original corpus. And, in each class, there are 30,000 training samples as well as 1,900 research samples. In total, there are 120,000 training samples and 7,600 research samples. Performing basic preprocessing is a very important step as it removes duplicates of the data. Using unclean and incomplete text data can be catastrophic. As a result, in this stage, we will delete all unnecessary symbols, characters, and other elements from the text that do not affect the problem's target.

3.2 Methods

The two main classifications of methods for text summarization are the extractive summarization method and abstractive summarization method. In the extractive summarization method, we recognize the important sentences or expressions from the original text and extract only those from the text. Then, those extracted sentences would be our summary. However, in the abstractive summarization method, we generate new sentences from the original text. And, the sentences generated through abstractive summarization might not be present in the original text.

Our goal is to implement a text summarizer where the input can be a long sequence of words, while the output would be a short summary of the input. Therefore, we can model this as a Many-to-Many Seq2Seq problem. The following image illustrates a typical Seq2Seq model architecture:

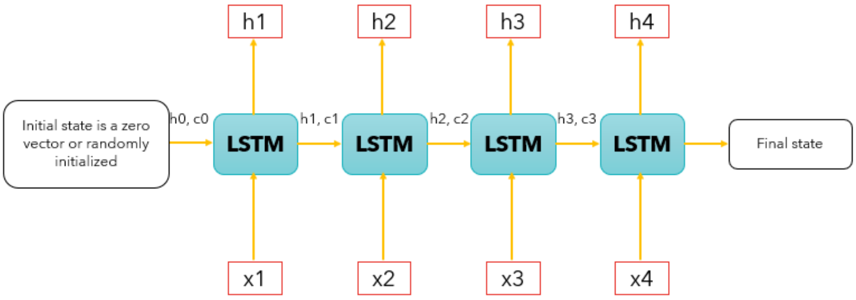

Figure 1: Encoder architecture

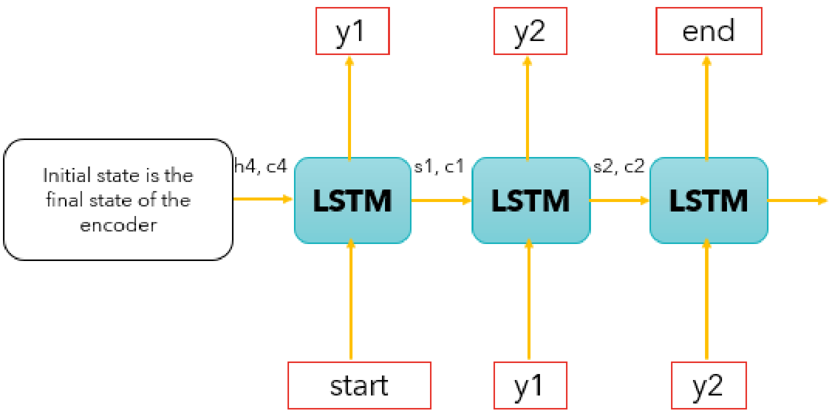

Figure 2: Decoder architecture

The two main components of a Seq2Seq model are Encoder and Decoder. The Encoder-Decoder architecture is mainly applied for solving the sequence-to-sequence (Seq2Seq) problems in which the input and output sequences have different lengths from each other.

The encoder turns the whole input sequence into a fixed-length vector. And next, the decoder predicts the output sequence. However, this would not work for long sequences because the decoder searches through the entire input sequence for the prediction. Furthermore, memorizing long sequences into a fixed-length vector is difficult for the encoder.

Attention mechanism can be used for predicting a word by looking at a few specific parts of the sequence only instead of searching through the entire sequence. The two main classes of attention mechanism based on the way the attended context vector is derived are Global Attention and Local Attention. We will use global attention in this project.

In global attention, the attention is located on all the source positions. More specifically, all the hidden states of the encoder are considered to obtain the attended context vector:

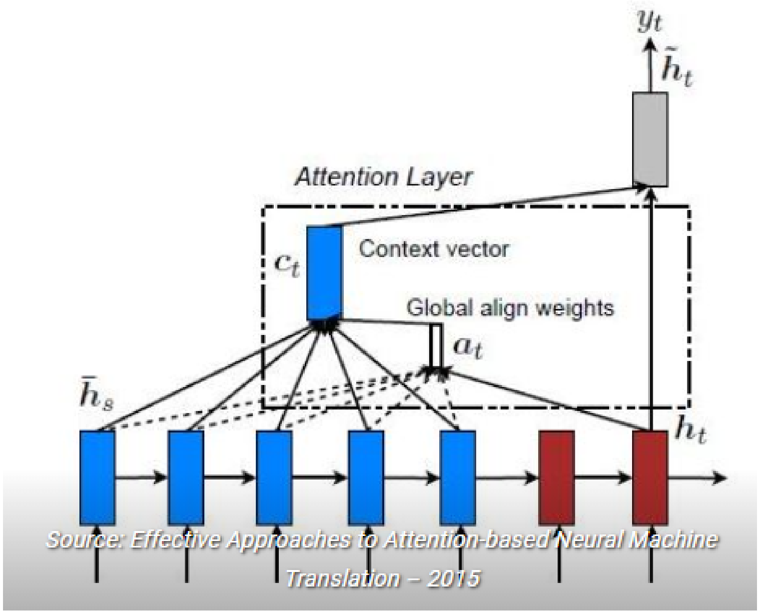

Figure 3: Global attention layer structure

In this project, in order to build our Text Summarizer, we will use a **4 layers stacked LSTM structured for encoder with a global attention layer**. In Stacked LSTM, multiple layers of LSTM are stacked on top of each other, which results in a better representation of the sequence.

4 Preliminary Results and Next Steps

We performed some preprocessing on our dataset, which included removing null and duplicate values. Then we replaced multiple grammatical forms of words with a singular form using a dictionary and removed any special characters or stop words. After all these steps, we used "matplotlib" to plot the word count frequencies for our dataset, which can be seen in figure 1.3 below. Lastly, we used a tokenizer to convert our word sequences to integer sequences.

Our next steps would be preparing our deep learning model with encoder and decoder and also implementing an attention layer and concatenating their outputs for better results.

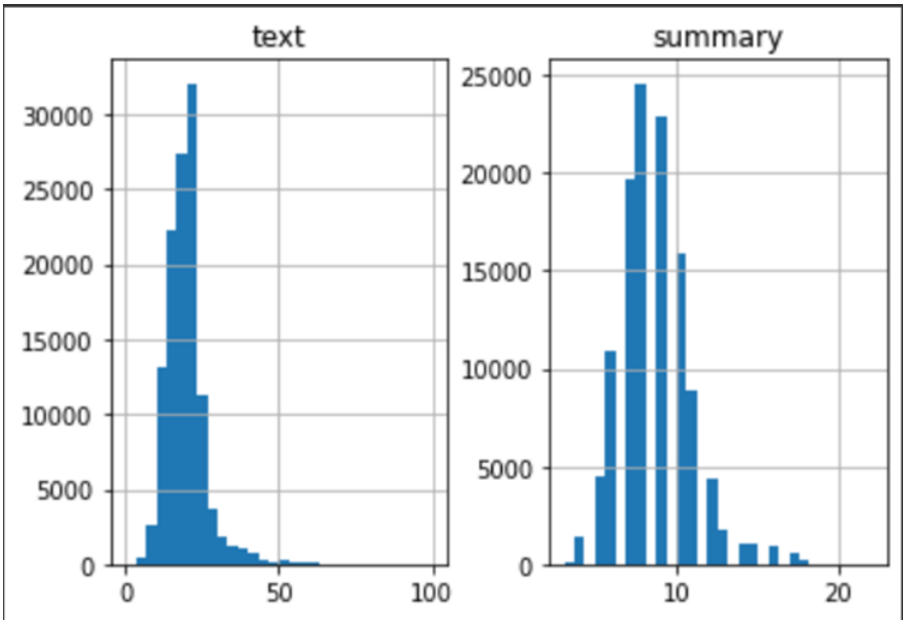

Figure 4: Word length frequencies of the ag-news dataset after preprocessing.

5 References

https://ieeexplore.ieee.org/abstract/document/9033912/

https://towardsdatascience.com/text-summarization-using-deep-learning-6e379ed2e89c

J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

J. Tandel, K. Mistree and P. Shah, "A Review on Neural network-based Abstractive Text Summarization models," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033912.

https://github.com/thushv89/attention_keras

J. S. Kallimani, K. G. Srinivasa and E. Reddy B, "Information extraction by an abstractive text summarization for an Indian regional language," 2011 7th International Conference on Natural Language Processing and Knowledge Engineering, Tokushima, Japan, 2011, pp. 319-322, doi:10.1109/NLPKE.2011.6138217.