huawei-university/nlp-course-projects

Generating headlines of news articles in Russian with sequence to sequence network and attention

Vadim Vlasov

May 2020

Abstract

This document contains the final project report of the Natural Language Processing course from Huawei University. The project code is available at https://github.com/vadimvvlasov/nlp-project.

1 Introduction

With the rapid spread of online news, users can be overwhelmed with huge amounts of information. Understandind all of this data is time consuming. So summarizing can help us to process and understand these data. On the other hand headlines are becoming increasingly important to attract readers to news articles. News headline generation is a subtask of summarization which has been extensively studied recently. Seq2Seq is the most common model for generating text for various languages. At the same time, the use of Seq2Seq for texts in Russian is currently poorly understood. In this project I'm going to train a neural network to generate headlines with the "Rossiya Segodnya" news dataset.

1.1 Team

This project was completed individually by Vadim Vlasov.

2 Related Work

Extractive-abstractive

There are two approaches to text summarization: extractive and abstractive. In this article, we are dealing with abstract.

Encoder-Decoder architecture

Unlike sequence prediction with one RNN, the Encoder-Decoder (other name seq2seq) model frees us from the length and order of the sequence, which is suitable for summarizing texts. Encoder-Decoder architecture works quite well for short sentences, so we might achieve a relatively high Bleu score, but for very long sentences, longer than 30 or 40 words, the performance comes down. Long sentences, it doesn't do well on because it's just difficult to get in network to memorize a super long sentence.

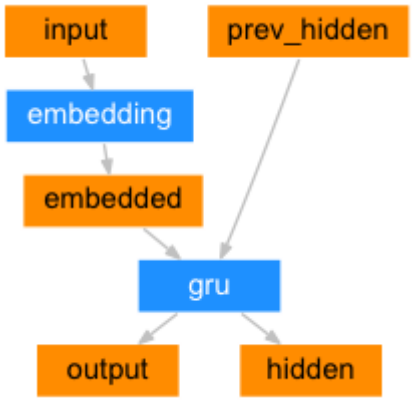
Attention Model

With the Attention Model we can improve the ability of neural network to memorize a long sentences. The model has attention weigths which tells us how much should we be paying attention to the different words from the input sentence.

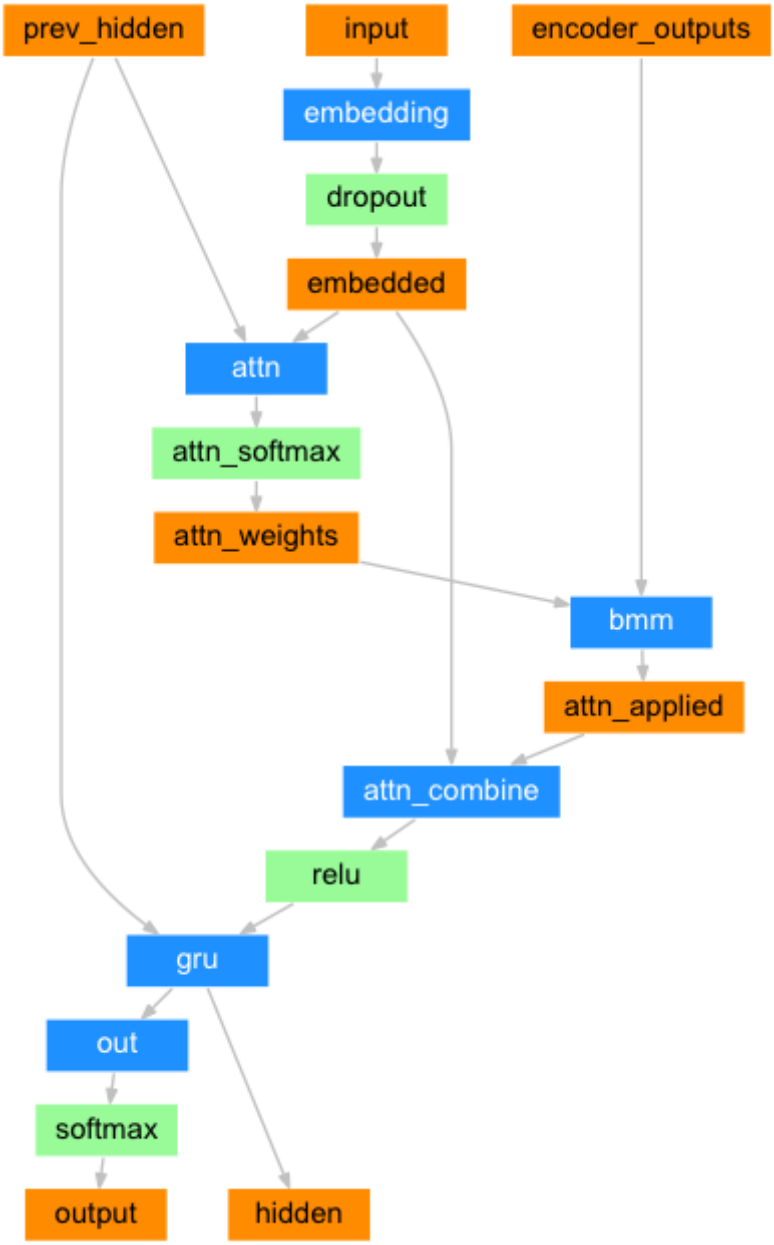
Transformer Model and transformer-based approaches

3 Model Description

Encoder network



Attention decoder network



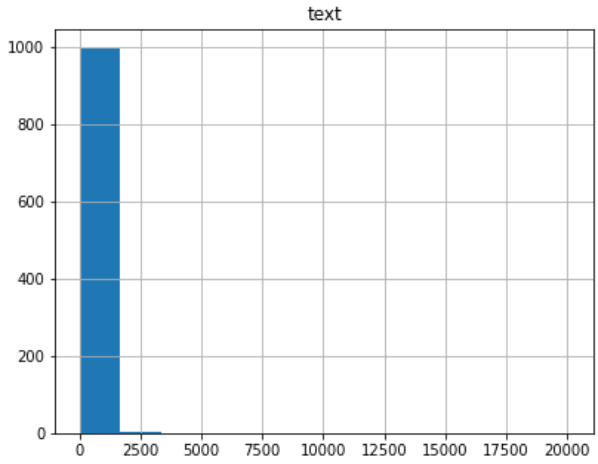
4 Dataset

Dataset under the project contains the first 1000 news documents from the full dataset. The dataset available for the research purposes here [https://github.com/RossiyaSegodnya/ria_news_dataset] Each row contains a JSON document that consists of two fields: text is a document body, while title is a news headline.

Each text in the data set contains HTML tags, so a pre-processing of texts is required before use it. Text pre-processing includes the following steps:

- normalize to lower case,
- remove all non-characters, including html markup
- split the texts, so have each word in it.

After preprocessing, the longest article has 20097 tokens. At the same time, 87.7% of texts contain less than 300 tokens. The distribution of words in the dataset texts is shown in the figure below



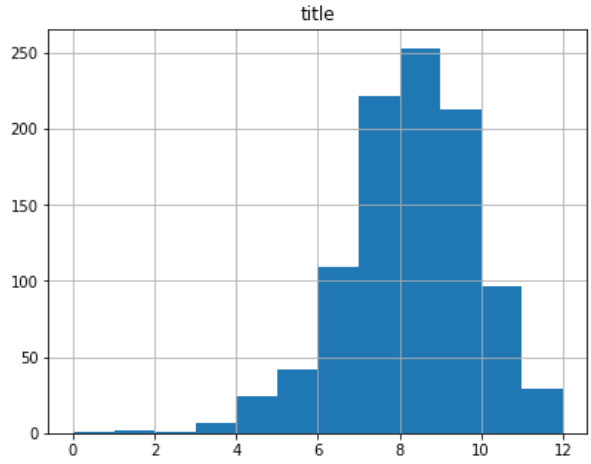


Table 1: Statistics of first 1000 news documents from the dataset ria_news_dataset

5 Experiments

This section should include several subsections.

5.1 Metrics

To assess the quality of the model, ROUGE metric is used. It measures n-gram overlap between predicted headlines and gold labels. R-1, R-2, R-L scores used here in terms of (p)recision $\frac{#overlap}{#predicted}$, (r)ecall $\frac{#overlap}{#gold}$ and F1 $\frac{2· p · r}{p+r}$.

Relative Length measures the ratio between the length of predicted headlines and the gold labels $\frac{LenPredict}{LenGold}$.

5.2 Experiment Setup

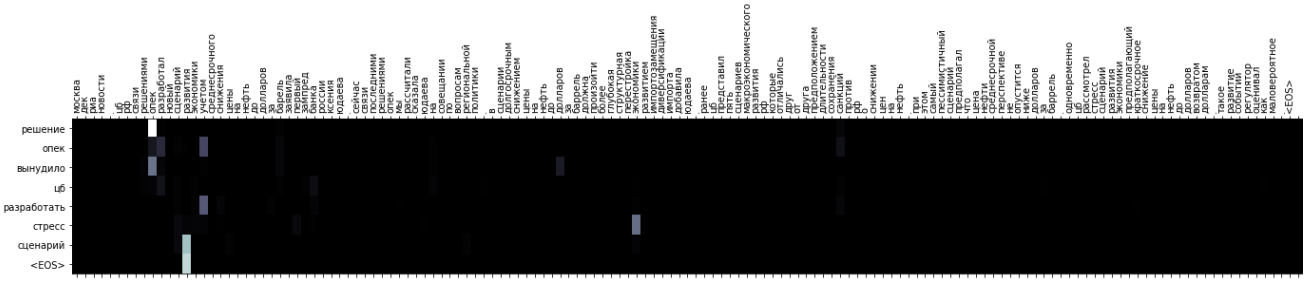
Trained model has 1 layer GRU with 256 hidden size in encoder and decoder.
Hyperparameter values are as follows:

- teacher forcing ratio = 0.5
- learning_rate = 0.01
- number of epoch iterations = 75 000
- Model trained on google “Colab” server with GPU. Training time is 424 minutes.

5.3 Baselines

According to [] state of the art solution shows the following score

6 Results



7 Conclusion

In this work RIA dataset was pre-processing and trimmed so that the texts do not contain non-symbols and its size limited by the maximum value. The sequence to sequence with attention approach applied to machine translation was adapted to the generation of article headlines.

Future work

- use pretrained word embedings
- use with more layers
- collect a larger dataset, for example, economic news from investing.com
- use newer transformer-based approaches

References

1. Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. Self-Attentive Model for Headline Generation. 41st European Conference on Information Retrieval, 2019. arXiv:1901.07786
2. "Rossiya Segodnya" news dataset. https://github.com/RossiyaSegodnya/ria_news_dataset
3. Nlp from scratch: Translation with a sequence to sequence network and attention. https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html