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

Vadim Vlasov

May 2020

Abstract

This document will provide you with guidelines for your project final report. You will learn how to structure the report and present your results. Please provide a link to your project code right here: `https://github.com/vadimvvlasov/nlp-project`.

1 Introduction

With the rapid spread of online news, users can be overwhelmed with huge amounts of information. Understanding 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. Encoder-decoder is the most common model for abstractive text summarization, but recently, self-attention models show better results for various languages. At the same time, the abstractive summarization of the texts in Russian shows less accuracy than for English.

This paper aimed to train the neural network of the headlines generator using the "Rossiya Segodnya" news dataset.

1.1 Team

This project was completed individually by **Vadim Vlasov**.

2 Related Work

Summarization problem is formulated as producing a shorter version of large text that retains most of its meaning. There are two approaches to text summarization: extractive and abstractive. Extractive summarization picks up sentences directly from the original document to form a brief summary. Abstractive

summarization is a more sophisticated NLP technique and more suitable way to generate headers.

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 Models.

With the Attention Model we can improve the ability of neural network to memorize long sentences (Fig. 1). The model has attention weights which tells us how much should we be paying attention to the different words from the input sentence.

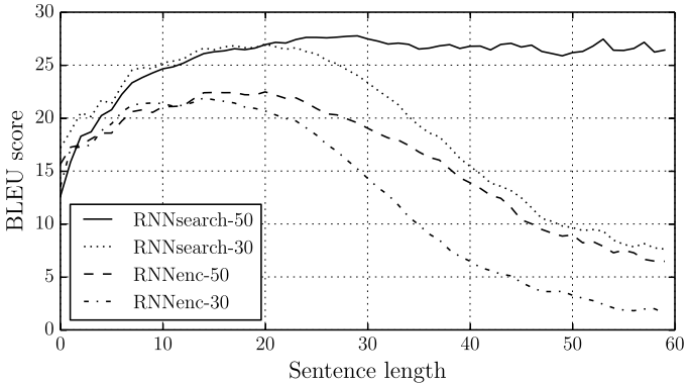

Figure 1: BLEU scores of model with fixed-length context vector and novel architecture [Dzmitry Bahdanau, 2016]

In [Vaswani et al., 2017] the new Transformer neural network architecture, based on a self- attention mechanism, was introduced. Transformer outperforms both recurrent and convolutional models and requires less computation to train.

To date seq2seq pre-training model ProphetNet [Yan et al., 2020] shows best performance on abstractive summarization. ROUGE-L 41.30 was achieved in the English language summarization tasks. ProphetNet is based on Transformer Seq2Seq architecture. It learns to predict future n-gram at each time step and uses n-stream self-attention mechanism.

Many papers are aimed at summarizing Russian texts on a RIA dataset. In [Gavrilov et al., 2019] Self-Attentive Model (Universal Transformer) with BPE is used. This approach has several advantages over RNN, it achieves ROUGE-L -score 36.81.

Phrase Based Attentional Transformer model in [M., 2010] on the RIA dataset shows ROUGE-l-f 40.02, which is closer to the state-of-the-art English language

ProphetNet ROUGE-L 41.30.

Thus, the Phrase Based Attentional Transformer shows the best results in abstractive summarization for the Russian language, while the ProphetNet model is promising.

3 Model Description

The model consists of two GRU networks shown on Fig. 2 , 3. Simple encoder network generates vector for each input word and output context vector. Context vector summarize the entire input text and is fed to the input of the decoder as the initial state. Input word vectors are used in the attention mechanism when generating output headers.

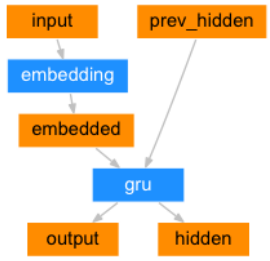

Figure 2: Encoder [web, c].

4 Dataset

Dataset under the project contains the first 1000 news documents from the full dataset. The dataset available for the research purposes here [web, b].

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 on Fig. 4 below

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

5 Experiments

This section should include several subsections.

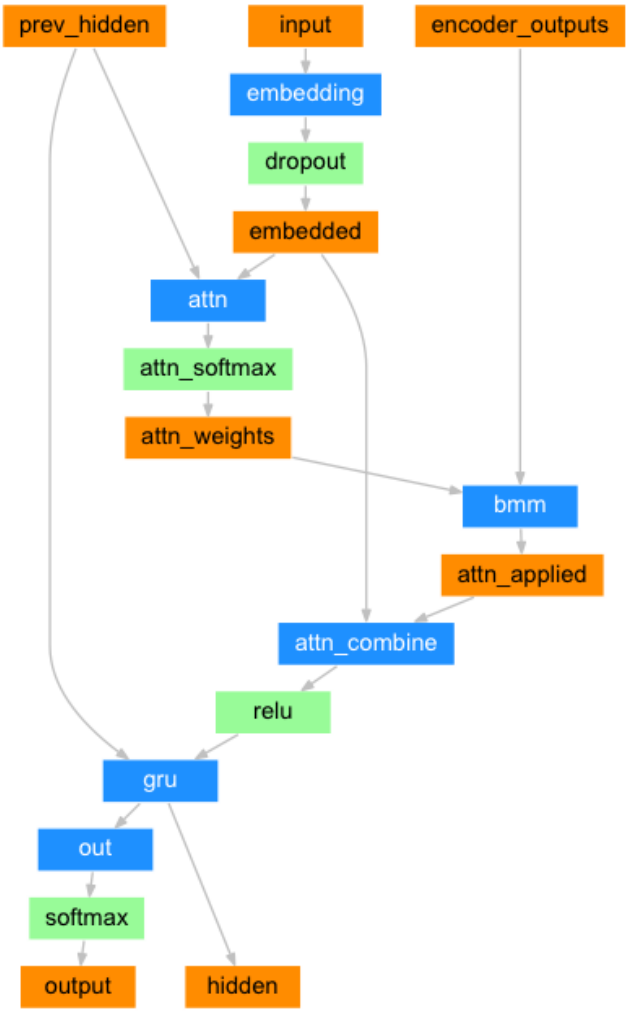

Figure 3: Decoder with attention [web, c].

	Train	Valid
Articles	789	88
Tokens	129,215	14,681
Vocabulary size	25,551	

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

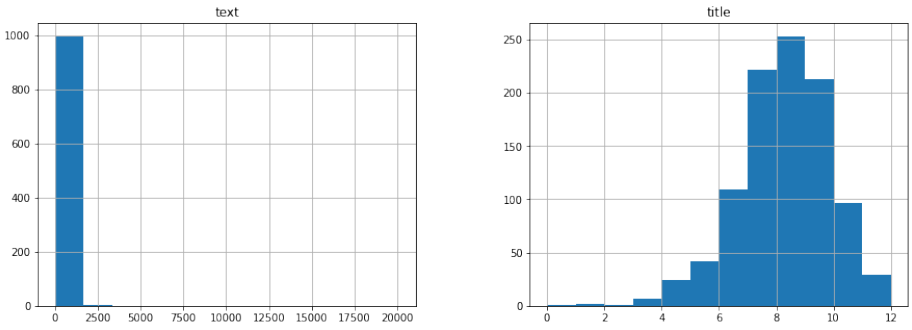

Figure 4: Word distribution in texts and headings.

5.1 Metrics

First of all, you should describe the metric(s) you were using to evaluate your approach. Most likely a metric description will include a formula.

Evaluation 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 $\frac{2\Delta p \Delta r}{p+r}$.

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 [web, a] state of the art solution is ProphetNet [Yan et al., 2020], a sequence-to sequence pretraining model. It learns to predict future n-gram at each time step and shows the best performance in abstractive summarization. Its ROUGE-L score is 41.30.

6 Results

Attention visualization shows which words of the input text are more important when generating the title

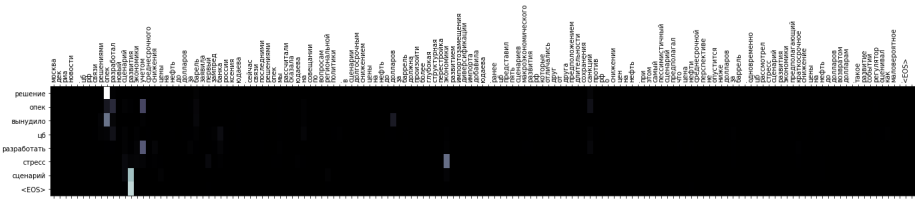

Figure 5: Attention visualization.

Some examples of results on a trained model can be found in Tab. 2.

= бородюк надеется что торпедо рассчитаются по долгам командой
< бородюк российский футбол копирует самое плохое из европейского <EOS>
= более тонн зараженных вредителями томатов из египта не пустили рф
< более тонн зараженных вредителями томатов из египта не пустили рф <EOS>
= ндв недвижимость подарит кв за каждые кв в новостройках
< ндв недвижимость подарит кв за каждые кв в новостройках <EOS>
= обсе украинские военные ополченцы лнр договорились прекратить огонь
< почти тыс абонентов донецке остаются без газа из за обстрелов <EOS>
= в бахрейне правозащитница осуждена за нападение на полицейских
< в бахрейне правозащитница осуждена за нападение на полицейских <EOS>

Table 2: Output samples.

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
- handle out of vocabulary words with different subword algorithms
- use with more layers
- collect a larger dataset, for example, economic news from investing.com
- use newer transformer based approaches

References

[web, a] Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

[web, b] Rossiya Segodnya news.

[web, c] NLP from scratch: Translation with a sequence to sequence network and attention.

[Dzmitry Bahdanau, 2016] Dzmitry Bahdanau, Kyunghyun Cho, Y. B. (2016). Neural machine translation by jointly learning to align and translate.

[Gavrilov et al., 2019] Gavrilov, D., Kalaidin, P., and Malykh, V. (2019). Self-attentive model for headline generation. In *Proceedings of the 41st European Conference on Information Retrieval*.

[M., 2010] M., S. A. (2010). Phrase-based attentional transformer for headline generation. computational linguistics and intellectual technologies. In *Proceedings of the International Conference "Dialogue" 2019*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need.

[Yan et al., 2020] Yan, Y., Qi, W., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. (2020). Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training.