Generation of headlines for news articles in Russian

Maria Ivanova

January 2022

Abstract

This document is the report for the project "Generation of headlines for news articles in Russian". A link to the project code here: `https://github.com/marii98/Generation-of-headlines-for-news-articles-in-Russian`.

1 Introduction

People nowadays read various information channels, online newspapers, etc. The amount of information that is published every day is huge, just to read it all, you can spend the whole day. That is why many people read only the headline of the news in order to understand whether this article is interesting to them. The title should highlight the key idea of the main text and make it clear to the reader what it is about.

Headline generation can significantly speed up and facilitate the work of journalists and all those who work with texts and make money on it. The author spends several minutes writing the title, while AI can write it in a second. The purpose of this work was to train the headline generator neural network on the Lenta.Ru news set.

1.1 Team

Maria Ivanova prepared this document.

2 Related Work

In fact, a heading is a text that is produced from one or more texts, which contains important information from them and whose length is significantly less than the original text.

One of the classifications applied to summarization methods divides them into extractive and abstract methods [Radeev, 2002]. The first methods are based on the selection of the most important sentences from the texts, while the second ones include the text generation method, since they generate new sentences for the abstract.

In this work[Здоровец, 2020], an algorithm was created for automatically creating titles for Russian-language materials based on the Seq2Seq model, which converts one sequence to another using the Encoder-Decoder architecture. An LSTM layer has been added to increase efficiency.

The data was split into training and test sets in the ratio 80/20, and the training was divided into training and validation in the same ratio. The Seq2Seq model had 2 layers in the encoder and in the decoder. Accordingly, there were 2 times more states of the last layer. The states of the last encoder layers were fed to the corresponding decoder layer (the state of the first was the initial state of the first, etc.).

As a result, by the end of the final epoch, the loss on the training set was 3.70, on the validation set - 4.07. The accuracy on the training was 0.45, on the validation - 0.43.

This paper [Шевчук, 2020] also presented the implementation of the Encoder-Decoder model and analyzed the results. To evaluate the results, the BLEU score metric was used, which reached a value of 0.1083239839959936 on the test sample. The results showed that the neural network copes with the generation of titles with the correct word order, they are related in meaning to the text. However, the generated results are inferior in quality to human-composed headlines.

3 Model Description

T5 transformer is inherently a simple encoder-decoder model. It is a "unified framework that converts every language problem into a text-to-text format". It is the model in the transformers series introduced by Google and Facebook.

The most notable feature of this model is its "text-to-text" nature. Unlike other transformers which take in natural language data only after converting it to corresponding numerical embeddings, T5 takes in data in the form of text only. And the outputs are also strings of characters, i.e. text again. This text-to-text nature enables the model to learn any NLP task without changing the hyperparameters and loss functions. Also it is "unified" in the sense that it can perform several NLG tasks simultaneously. It does not require separate output layers for different tasks like other transformers such as BERT and GPT2.

As already said the T5 model has an encoder-decoder based transformer architecture which is best suited for the text-to-text approach. The number of parameters is kept same as BERT (which is an encoder only model) by sharing them across decoder and encoder without a significant drop in performance.

The architecture of the T5 model is shown in Fig. 1.

The encoder consists of a stack of identical layers. Every layer is composed of two sub-layers. The first sub-layer of each encoder layer is a multi-head self-attention mechanism. The second sub-layer on the other hands is a fully connected position-wise feed-forward network. Residual connections are employed around these sub-layers, each followed by the normalization layer. Similar to the encoder, decoder also consists of a stack of identical layers.

In the decoder, a third sub-layer is also inserted, in addition to the two sub-layers already present in the encoder layer. This third sub-layer performs multi-head attention on the output received by the encoder stack. Here also residual connections are employed around these sub-layers, like that of encoder, each followed by the normalization layer.

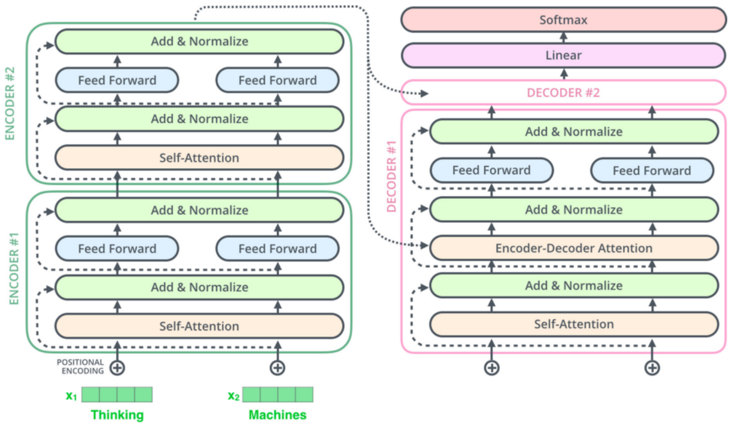


Figure 1: Architecture of T5 [Grover, 2020].

4 Dataset

The dataset was taken from Lenta.ru [web,a]. It includes about 800,000 records.

The data was collected from September 1999 to December 2019. The data includes the URL of the news article, the title of the article, the main text of the article, the subject and tags of the article, and the date of the article.

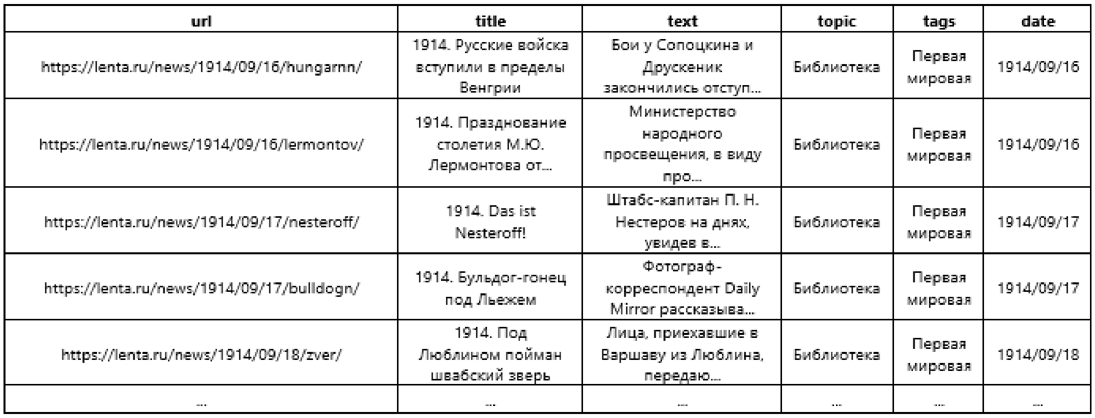
Due to specific formatting of some news, some texts were collected with errors. For the test sample, 20 percent of all data was given.

An example of data can be seen in Fig. 2 The distribution of words in the dataset texts and titles is shown in Fig. 3.

5 Experiments

5.1 Metrics

A metric called Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which is a standard for assessing the quality of summation results. As the



url	title	text	topic	tags	date
https://lenta.ru/news/1914/09/16/hungarnn/	1914. Русские войска вступили в пределы Венгрии	Бои у Сопоцкина и Друскеник закончились отступ...	Библиотека	Первая мировая	1914/09/16
https://lenta.ru/news/1914/09/16/lermontov/	1914. Празднование столетия М.Ю. Лермонтова от...	Министерство народного просвещения, в виду про...	Библиотека	Первая мировая	1914/09/16
https://lenta.ru/news/1914/09/17/nesteroff/	1914. Das ist Nesteroff!	Штабс-капитан П. Н. Нестеров на днях, увидев в...	Библиотека	Первая мировая	1914/09/17
https://lenta.ru/news/1914/09/17/bulldogn/	1914. Бульдог-гонец под Льежем	Фотограф-корреспондент Daily Mirror рассказыва...	Библиотека	Первая мировая	1914/09/17
https://lenta.ru/news/1914/09/18/zver/	1914. Под Люблином пойман швабский зверь	Лица, приехавшие в Варшаву из Люблина, передаю...	Библиотека	Первая мировая	1914/09/18
...	...	...	...	...	...

Figure 2: The first lines of the dataset

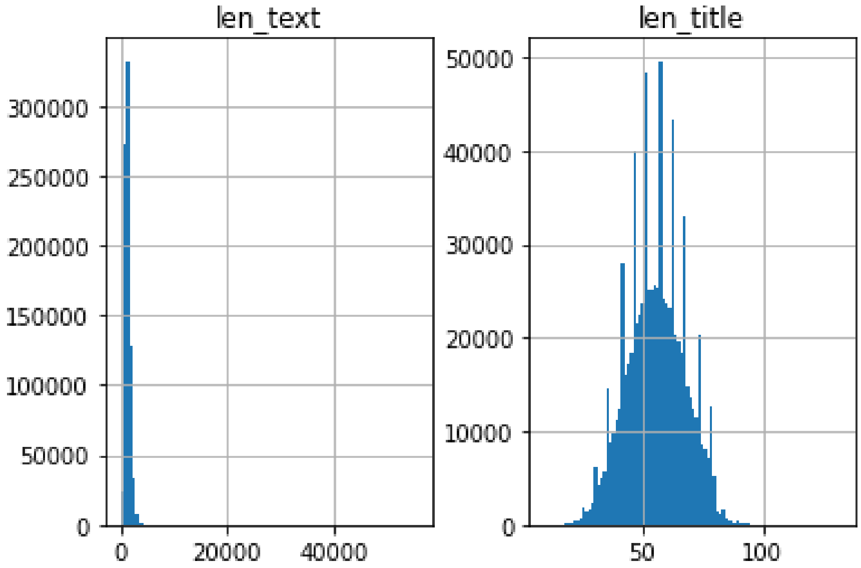
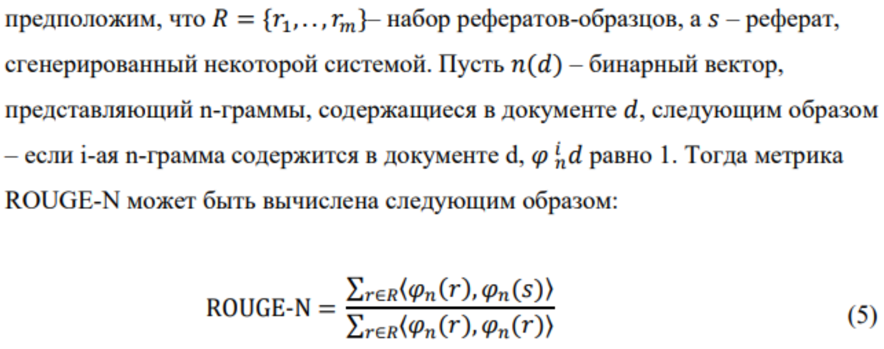


Figure 3: Word distribution in texts and titles.

name suggests, these metrics are recall-based. The ROUGE-N metric can be calculated as follows Fig. 4

Another metric from the set applies the concept of the longest substrings. The previous measure was based on the assumption that the number of matching n-grams reflects the similarity of automatic and manual abstracts. This one takes a general sequence as such a measure - the longer it is, the more similar the abstracts are to each other.



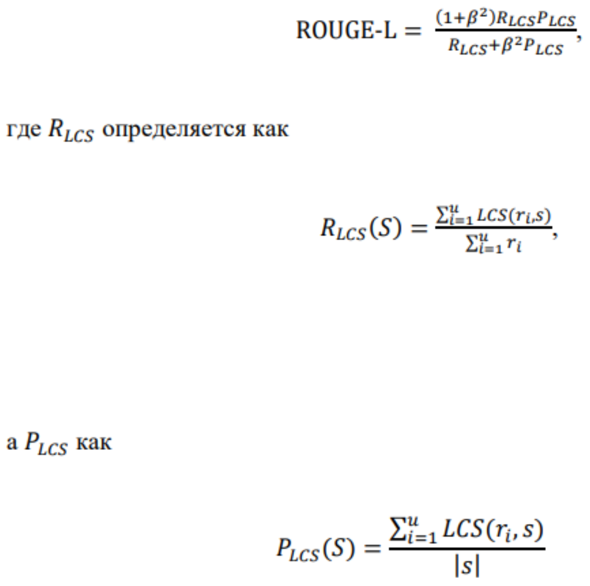
предположим, что $R = \{r_1, .., r_m\}$– набор рефератов-образцов, а s – реферат, сгенерированный некоторой системой. Пусть $n(d)$ – бинарный вектор, представляющий n-граммы, содержащиеся в документе d, следующим образом – если i-ая n-грамма содержится в документе d, $\varphi\,{}_n^i d$ равно 1. Тогда метрика ROUGE-N может быть вычислена следующим образом:

$$\text{ROUGE-N} = \frac{\sum_{r \in R} \langle \varphi_n(r), \varphi_n(s) \rangle}{\sum_{r \in R} \langle \varphi_n(r), \varphi_n(r) \rangle} \tag{5}$$

Figure 4: How to calculate Rouge-N.

Despite the name and focus on recall, there is and is used the ROUGE-precision variant, which is calculated in a manner similar to ROUGE-n, however, unlike the latter, the denominator is not the number of n-grams in the abstract created by a person, but the number of n-grams in models.

In addition, ROUGE-L is defined as an F-measure based on the longest substring. Metric formulas are shown in the Fig. 5



$$\text{ROUGE-L} = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}},$$

где R_{LCS} определяется как

$$R_{LCS}(S) = \frac{\sum_{i=1}^{u} LCS(r_i, s)}{\sum_{i=1}^{u} r_i},$$

а P_{LCS} как

$$P_{LCS}(S) = \frac{\sum_{i=1}^{u} LCS(r_i, s)}{|s|}$$

Figure 5: How to calculate Rouge-L.

5.2 Experiment Setup

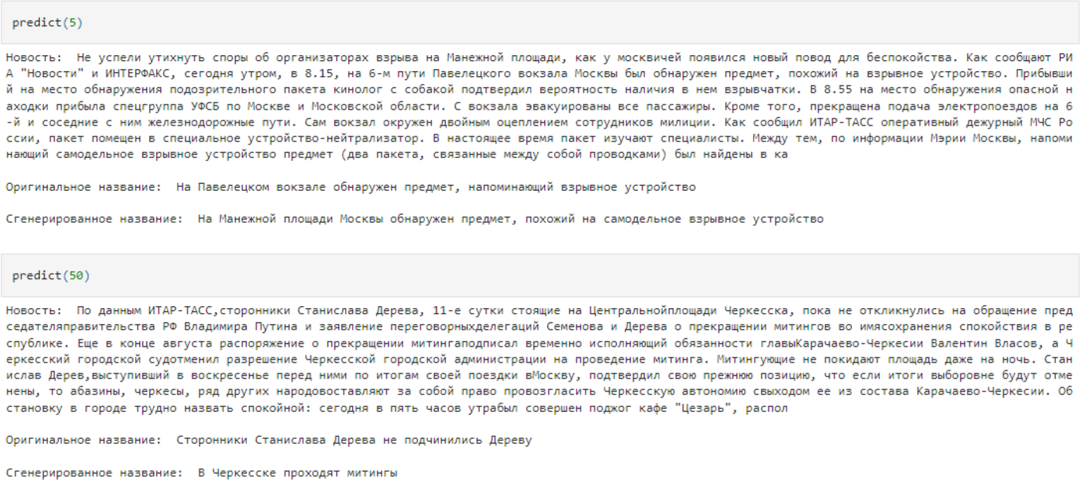
The data were divided into test and training sets. For the training sample, 80 percent of the data was allocated, the rest were used for the test sample, according to the results of which the metric was subsequently calculated.

5.3 Baselines

The Seq2Seq model was considered as a basis. The main difference between the Transformer architecture and the Seq2seq architecture is the rejection of recurrent blocks, such as RNN and LSTM, due to the fact that this network is entirely built on the multi-head self-attention mechanism. The transformer receives as input a set of key-value pairs, where the dimension of both components is equal to the length of the input sequence.

6 Results

After reviewing several generated headers (Fig. 6) we can conclude that the model works well. Headings are readable, convey the main meaning of the text and are grammatically adequate, and the main idea is clear.



```
predict(5)
```

```
Новость:  Не успели утихнуть споры об организаторах взрыва на Манежной площади, как у москвичей появился новый повод для беспокойства. Как сообщают РИ
А "Новости" и ИНТЕРФАКС, сегодня утром, в 8.15, на 6-м пути Павелецкого вокзала Москвы был обнаружен предмет, похожий на взрывное устройство. Прибывши
й на место обнаружения подозрительного пакета кинолог с собакой подтвердил вероятность наличия в нем взрывчатки. В 8.55 на место обнаружения опасной н
аходки прибыла спецгруппа УФСБ по Москве и Московской области. С вокзала эвакуированы все пассажиры. Кроме того, прекращена подача электропоездов на 6
-й и соседние с ним железнодорожные пути. Сам вокзал окружен двойным оцеплением сотрудников милиции. Как сообщил ИТАР-ТАСС оперативный дежурный МЧС Ро
ссии, пакет помещен в специальное устройство-нейтрализатор. В настоящее время пакет изучают специалисты. Между тем, по информации Мэрии Москвы, напоми
нающий самодельное взрывное устройство предмет (два пакета, связанные между собой проводками) был найдены в ка

Оригинальное название:  На Павелецком вокзале обнаружен предмет, напоминающий взрывное устройство

Сгенерированное название:  На Манежной площади Москвы обнаружен предмет, похожий на самодельное взрывное устройство
```

```
predict(50)
```

```
Новость:  По данным ИТАР-ТАСС,сторонники Станислава Дерева, 11-е сутки стоящие на Центральнойплощади Черкесска, пока не откликнулись на обращение пред
седателяправительства РФ Владимира Путина и заявление переговорныхделегаций Семенова и Дерева о прекращении митингов во имясохранения спокойствия в ре
спублике. Еще в конце августа распоряжение о прекращении митингаподписал временно исполняющий обязанности главыКарачаево-Черкесии Валентин Власов, а Ч
еркесский городской судотменил разрешение Черкесской городской администрации на проведение митинга. Митингующие не покидают площадь даже на ночь. Стан
ислав Дерев,выступивший в воскресенье перед ними по итогам своей поездки вМоскву, подтвердил свою прежнюю позицию, что если итоги выборовне будут отме
нены, то абазины, черкесы, ряд других народовоставляют за собой право провозгласить Черкесскую автономию свыходом ее из состава Карачаево-Черкесии. Об
становку в городе трудно назвать спокойной: сегодня в пять часов утрабыл совершен поджог кафе "Цезарь", распол

Оригинальное название:  Сторонники Станислава Дерева не подчинились Дереву

Сгенерированное название:  В Черкесске проходят митинги
```

Figure 6: Examples of generated titles.

The following results were obtained. Metric ROUGE1 = 0.07278906926406926, ROUGE2 = 0.019727777777777778, ROUGE-L = 0.07197240259740262, ROUGE-Lsum = 0.07247694805194807. That is a pretty good result.

Having calculated the average value ROUGE = 0.0592415494227994, we see that it slightly exceeds the initial value obtained in the article [Аишева, 2021] and is equal to 0.05800769685750703. However, this result cannot compete with the results obtained by larger laboratories and users, for example, in the article [Shevchuk, 2019] the result is 0.2314152364

7 Conclusion

In the course of this work, the Russian-language news corpus Lenta.ru was used. The data set included about 800,000 records. To evaluate the result, the ROUGE metric was applied.

The results of the experiment showed that the neural network copes with generating a header with the correct word order. In addition, the resulting headlines are related in meaning to the text of the news and often resemble the original headlines.

And although the initial result of the metric was obtained higher than that obtained from the authors, on the basis of which the studies were conducted. Unfortunately, later works were found in which the results significantly exceed the obtained result.

8 References

[Radeev, 2002] Radev D.R., McKeown K., Hovy E. Introduction to the Special Issue on Summarization // Computational Linguistics. 2002. № 2 (28). C. 399–408.

[Здоровец, 2020] Здоровец А. И. и др. Генерация заголовков с помощью многоуровневой Seq2Seq модели //ACTUAL PROBLEMS OF LINGUISTICS AND LITERARY STUDIES. – 2020. – С. 96.

[Шевчук, 2020] Шевчук А. А. и др. АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИИ НОВОСТНЫХ ЗАГОЛОВКОВ С ПРИМЕНЕНИЕМ НЕЙРОННОЙ СЕТИ ENCODER DECODER //ACTUAL PROBLEMS OF LINGUISTICS AND LITERARY STUDIES. – 2020. – С. 100.

[Grover, 2020] Grover K. et al. Deep learning based question generation using t5 transformer //Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10. – Springer Singapore, 2021. – С. 243-255.

[Аишева, 2021] Аишева Д. А. и др. Модификация нейронной сети Transformer для генерации новостных заголовков на русском языке: магистерская диссертация по направлению подготовки: 45.04. 03-Фундаментальная и прикладная лингвистика. – 2021. [Shevchuk, 2019] Shevchuk A., Zdorovets A. TEXT SUMMARIZATION WITH RECURRENT NEURAL NETWORK FOR HEADLINE. 2019

[web,a]https://github.com/marii98/Generation-of-headlines-for-news-articles-in-Russian

[web,b]https://www.kaggle.com/datasets/yutkin/corpus-of-russian-news-articles-from-lenta