Automatic Generation of Russian News Headlines

Ekaterina Tretiak[a]

[a] *Saint Petersburg State University, 7-9 Universitetskaya emb., St Petersburg, 199034, Russia*

Abstract

Text summarization is one of the key Natural Language Processing tasks. Automated text summarization has the potential to save time when creating reviews, abstracts etc. of the texts across multiple domains. Automatic headline generation is a challenging kind of text summarization. Basically, the distinction between extractive and abstractive summarization methods is drawn. Application of the extractive summarization techniques results in the extraction of relevant words or sentences from the original text. Abstractive summarization models synthesize a summary in which some of its material is not present in the input document. This paper deals with the fine-tuning the pretrained model based on Transformer architecture for the task of generation of Russian news headlines. Experiments discussed were carried out for the new dataset of Russian news which was automatically compiled from the "Bumaga" website. The paper contains the quantitative evaluation results using BLEU und ROUGE metrics as well as the human evaluation results. Finally, the paper presents error analysis and discussion of particular contexts.

Keywords

Headline generation, text summarization, abstractive summarization, Russian language, RuBERT

1. Introduction

In modern computational linguistics, text summarization holds a special place among the tasks of natural language processing (NLP). The aim of summarization is to produce a shorter version of the text that expresses the main idea of the source document. That is, given input text x, a model writes a summary y which is shorter than x and contains vital information from x. Text summarization makes it possible to access and process large amounts of textual data and extract the necessary information from a huge corpus of texts.

The automatic summarization problem can be addressed with two types of techniques, extractive and abstractive ones [1]. In extractive summarization, the most significant chunks of the source text are detected and extracted without any changes. That means that all words in the summary come from the input data. In contrast, abstractive summarization systems attempt to generate abstracts from new sentences, which may not even include words that occurred in the original. Although an abstractive model is much more complex than the extractive one, it produces detailed human-like summaries. It is this advantage that makes abstractive approaches increasingly popular today and, for this reason, we focus on them.

In this paper, we are concerned with the task of headline generation that tends to be considered as a special type of text summarization [2]. This is accounted for by the fact that the headline is a key component of the news text since it includes its main ideas. On the one hand, it should be quite informative, and on the other, encourage readers to spend their time on reading the full text. However, for digital media, it is especially essential to provide clear and informative headlines, since the user does not have time to guess what the hidden meaning was intended. In addition, the headline like any other text should have grammatical and lexical linking and meaningfulness.

IMS 2021 - International Conference "Internet and Modern Society", June 24-26, 2021, St. Petersburg, Russia
EMAIL: evtretyak1999@gmail.com
ORCID:

There are several sections in this paper. The review of current studies in the field of automatic summarization is presented in the Related work section. In the Methods section there is information about the corpus of news messages which was to be processed as well as the description of the used model. The Experiment section describes the present method of generation of Russian News Headlines. In the Results section there are examples of headlines predicted by our model, automatic and human evaluation results and error analysis. The Conclusion section provides the conclusions drawn up by the presented results.

2. Related work

Analysis of current research shows that the automatic summarization problem can be approached differently and a large number of papers covers it.

Many studies are devoted to extractive methods of text summarization [3][4]. [5] was one of the first to work on this issue in terms of detecting the most informative words relying on the word frequency. His idea was to count the frequency of words in order to find a list of the most meaningful words. However, the main disadvantage of using extractive methods to headline generation is that the abstracts they produce are hardly headlines. They cannot be shorter that the minimum text blocks used to compose them (a sentence or a paragraph). This is how neural models for abstractive summarization and text generation came into being.

Sequence-to-sequence (seq2seq) model is one of the most important recent concepts used in the current state-of-the-art applications in natural language processing. It is a type of encoder-decoder model using Recurrent neural network (RNN), that generates one sequence from the other after it has learned a great deal of sequence pairs. Developers from Google [6] demonstrated that translation models based on seq2seq outperform a standard statistical machine translation based (SMT-based) system. Not only machine translation benefits from seq2seq models; they do well on many other sequences learning problems, including text summarization and headline generation. In 2015, [7] proposed an approach called Attention-Based Summarization (ABS). It is a local attention-based model that generates the next word of the summary given the input sentence in terms of the combination of a neural language model and a contextual input encoder. Following them, [7] extended the ABS model using semantic and syntactic information about the source text in a standard neural attention-model.

Later, copying mechanism was presented [9] to improve RNN encoder-decoder model. It is designed to copy tokens from the source text. This model was taken as a basis in another study [10] and trained on the dataset of Russian news.

The Transformer architecture, originally developed for machine translation [11], is now applied to all the main tasks of natural language processing. There are many modified versions of Transformer. Thus, [2] adapted the Universal Transformer architecture [12], which is a modification of Transformer, to the task of headline generation.

Previous advances in abstractive text summarization have been made using pretrained language models based on Transformer architecture. In 2019, a new Bidirectional Encoder Representations from Transformers (BERT) [13] architecture was developed specifically for text summarization (BertSumExt and BertSumAbs for extractive and abstractive summarization, respectively) [14]. The BertSumAbs model is a standard encoder-decoder framework for abstractive summarization [15], where the encoder is the pretrained BertSum and the decoder is a 6-layered Transformer initialized randomly. In [16] RuBERT [17] was used as a pretrained BERT for fine-tuned on the Russian texts BertSumAbs model. Application of this approach to the task of Russian news headlines allowed to obtain state-of-the-art results on the RIA [2] and Lenta[2] datasets.

[2] https://github.com/yutkin/Lenta.Ru-News-Dataset

3. Methods

3.1. Data

We conduct our experiments on the new corpus of news messages in Russian. We have developed a programming algorithm for automatic corpus building from the website of the Russian online newspaper “Bumaga”[3]. It contains news messages from June, 2013 to April, 2021. In total, there are 38 499 news articles in the provided corpus which are supplied with additional meta information: title, date and link [4]. For the experiment, we split the Bumaga corpus into the train, validation, and test parts in a proportion of 90:5:5.

3.2. Model description

We examine the BertSumAbs model, which utilizes RuBERT as a pretrained BERT [16]. The original BertSumAbs model is a standard encoder-decoder framework that was fine-tuned for abstractive summarization task. The encoder is 6 stacked layers of BERT, while the decoder is a 6-layered Transformer that is initialized randomly. Thus, the encoder is pretrained and the decoder must be trained from the ground up. The model has more than 317M parameters.

We fine-tune a 40K checkpoint saved by the authors of [16], since its validation loss score was the best. That is, trained on the RIA dataset checkpoint is fine-tuned on the Bumaga dataset.

4. Experiment

4.1. Baseline model

First Sentence This model uses the first sentence of a news message as its hypothesis for a news message headline. It is the most naïve approach to headline generation. Its application is valid due to the fact that the structure of news articles is based on the principle of inverted pyramid. It means that the most valuable information can be found in the first sentence through the answers to key questions: Who? When? Where? Why? What? How?

4.2. Training

It has been mentioned that the encoder is pretrained while the decoder is trained from scratch in the BertSumAbs model. This mismatch between two parts of the Transformer can make fine-tuning unstable, as noted [14]. In order to overcome the difficulty, a new fine-tuning schedule was designed by [14] and then borrowed by [16][16]. This novel approach is characterized by using of different optimizers for the encoder and the decoder. Following [16], we separate the optimizers of the encoder and the decoder when training model on our dataset. We use two Adam optimizers [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and learning rates $lr_e = 0.002$ и $lr_d = 0.2$ for the encoder and decoder, respectively. When setting the parameters for training the model, we rely on the idea of [14] that the pretrained encoder must be fine-tuned with a smaller learning rate.

The model is fine-tuned with a batch size equals 128, gradient accumulation every 95 steps. The model was trained for 4,700 steps on a Tesla V100 GPU provided by Google's Colaboratory service [5]. The training of the model took about 24 hours.

[3] https://paperpaper.ru/
[4] The dataset is available at https://github.com/ekaterinatretyak/PreSumm.
[5] https://colab.research.google.com/

5. Results

In Table 1[6] we present results of the headline generation based on the Bumaga corpus for Russian. Despite the fact that our fine-tuned model makes mistakes, which are discussed in Section 5.3., relevant headlines still prevail.

Table 1
Samples of headlines generated after fine-tuning BertSumAbs

№	Lang.	Original text	Original headline	Generated headline
1	ru	В Эрмитаже появились коты с именами Трамп и Хиллари…	В Эрмитаже появились коты Трамп и Хиллари	В Эрмитаже появились коты с именами Трампа и Клинтон
	en	Cats with the names Trump and Hillary appeared in the Hermitage…	Cats Trump and Hillary appeared in the Hermitage	Cats with the names of Trump and Clinton appeared in the Hermitage
2	ru	Совет Федерации назначил дату проведения президентских выборов в 2018 году — 18 марта…	Совет Федерации объявил дату проведения выборов президента в 2018 году	Совет Федерации назвал дату проведения президентских выборов в 2018 году
	en	The Federation Council set the date for the presidential elections in 2018 — March 18…	The Federation Council announced the date of the presidential election in 2018	The Federation Council named the date of the presidential elections in 2018
3	ru	Пожар на Васильевском острове затруднил дорожную обстановку в центре Петербурга, поскольку сотрудники ДПС перекрывали участок дороги… в коммунальной квартире в доме 31/22 по Кадетской линии горела одна из комнат…	На Васильевском острове скопились пробки из-за пожара на Кадетской линии	На Васильевском острове горела коммуналка, движение перекрыто
	en	The fire on Vasilyevsky Island complicated the traffic situation in the center of St. Petersburg, since traffic officers blocked traffic on a section of road... in a communal apartment in the house 31/22 on the Kadetskaya line, one of the rooms was burning…	Traffic jams have accumulated on Vasilyevsky Island due to a fire on the Kadetskaya line	On Vasilyevsky Island, a communal apartment burned, traffic was blocked
4	ru	В Петербурге 25 июля произошли прорывы труб на севере и юго-западе Петербурга… Водой залило перекресток улицы Симонова и проспекта Просвещения…	Улицы на севере и юге Петербурга затопило из-за прорывов труб	В Петербурге прорвало трубу. Машины оказались наполовину в воде
	en	In St. Petersburg, on July 25, there were bursts of pipes in the north and south-west of St. Petersburg… Water flooded the crossroad of Simonov Street and Prosveshcheniya Avenue…	Streets in the north and south of St. Petersburg were flooded due to bursts of pipes	A pipe burst in St. Petersburg. The cars were half in the water

[6] The texts of news articles are given in an abbreviated form

The generated headlines seem to have quite a high grammatical and semantic coherence. It should be noted that predicted headlines may consist of words that are not present in the text of the article. Moreover, the model effectively uses techniques from the theory of paraphrasing, e.g., use of converses, synonyms etc.

Among generated news headlines single-sentence headlines predominate over headlines with two and more clauses. It was found that when the model produces two simple sentences, this decreases the text quality due to a repetition of the already generated word or phrase. These problems seem to be related to the fact that the checkpoint used was trained on the RIA corpus which includes more than 1 million news headlines, consisting mainly of a single sentence. Thus, the increase of training examples, in which the headline consists of two sentences, is expected to contribute to better results for generation of headlines of more complex structure.

Nevertheless, the model is able to generate relevant headlines that consist of two sentences:

- Сайт об архитектуре Петербурга Citywalls снова не работает. <q> Петербуржцы встревожены
  (en) *The website about the architecture of St. Petersburg ‘Citywalls’ is not working again. < q> Petersburgers are alarmed*
- Финляндия заняла первое место в рейтинге самых счастливых стран. <q>Россия заняла 59-е место
  (en) *Finland placed first in the list of the happiest countries. <q> Russia took the 59th place*

The model generates complex sentences with subordinate clauses:

- На Гороховой улице открылся ресторан «Мука и вода», где можно попробовать пасту
  (en) *The restaurant "Flour and Water" has opened on Gorokhovaya Street, where you can taste pasta*
- Здание клуба «Камчатку», где работал Цой, расселят
  (en) *The residents of the building of the club "Kamchatka", where Viktor Tsoi worked, are going to be rehoused*

An analysis of the headlines produced indicates that the model performs best when generating information-rich headlines consisting of a single sentence that inform the readers about the main facts of a news article. In addition, the model is able to produce headlines that contain quotes. However, among the generated headlines, it is quite difficult to find the headlines that would contain an irony, wordplay or hidden author’s opinion. Some examples can be seen below:

- В РПЦ назвали позицию Эрмитажа по Исаакиевскому собору «провокацией»
  (en) *The Russian Orthodox Church called the Hermitage's position on St. Isaac's Cathedral a "provocation"*
- «Путин, спаси нас»: жильцы дома на Ремесленной
  (en) *"Putin, save us": residents of the house on Remeslennaya Street*
- На Гороховой улице восстановили под гостиницу дом Крутикова. <q> Посмотрите, как выглядит
  (en) *On Gorokhovaya Street, the Krutikov house was restored as a hotel. <q> See what it looks like*

5.1. Automatic Evaluation

For automatic quality evaluation we use BLEU score [19] and ROUGE score [20]. Since the Bumaga corpus has no previous art, in Table 2 we present results for the baseline and fine-tuned model. Moreover, we present evaluation results on the Bumaga dataset while model is trained on the RIA dataset in order to evaluate the success of the model in headlines generation given news articles with another structure.

The results obtained demonstrate that the BertSumAbs model fine-tuned on the Bumaga dataset **performs best** for all metrics. However, the Bumaga dataset evaluation results using model trained on the RIA dataset are the worst. This may indicate that the format and style of writing news texts and headlines differ from one news agency to another. Thus, one of the most noticeable differences is that

the headlines from the Bumaga corpus often consist of two sentences, while ones from the RIA corpus mostly consist of a single sentence.

Table 2
Bumaga dataset evaluation

Model	BLEU	R1	R2	RL	R-mean
			Bumaga		
First Sentence	41.06	38.9	22.8	36.8	32.8
BertSumAbs trained on the RIA dataset	25.84	21.8	9.0	20.5	17.1
BertSumAbs fine-tuned on the Bumaga dataset	**48.51**	**44.1**	**28.4**	**42.4**	**38.3**

5.2. Human evaluation

We have found out that the headline of digital media especially should be informative. We have established also that the headline should have grammatical and lexical cohesion (see Section 1). Since automatic quality evaluation methods evaluate the formal match of tokens, rather than the semantic one, it is hardly possible to use them to understand how well the headlines meet these requirements. Commonly, it is the degree to which native speakers perceive a text that is the main criterion when analyzing the results of text generation experiments. For this purpose, we perform a qualitative analysis by randomly sampling 190 examples of the news text, original headline and our fine-tuned model generated headline for human evaluation. We asked 5 annotators who are native speakers of Russian to choose the most preferred headline for a news article between the original headline (Reference) and the generated one (Hypothesis). If there was no preference, the annotators chose the third option (Tie). The annotators had no idea about the details of the experiment, including which of the headlines was the reference. The results can be seen in Table 3.

Table 3
Human evaluation of generated headlines

Reference	Tie	Hypothesis
36%	48.6%	15.4%

From the results obtained it might be inferred that in almost every second case (48.6%) our model reaches human parity. This means that the headlines generated by the fine-tuned BertSumAbs model are interpreted to the same extent as the ones written by the journalists. Based on the criteria for choosing the preferred headline, it might be concluded that such headlines are informative and relevant, they are perceived as a single grammatical text. In Figure 1 we present examples of headlines for which the annotators have chosen the Tie option.

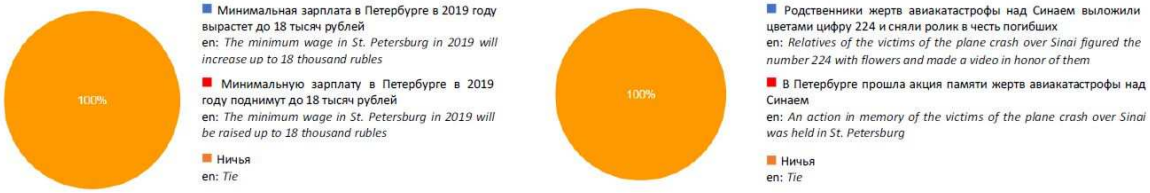

Figure 1: Examples of the headlines for which Tie option was selected

Analysing the aggregate statistics, we found that in 15.4% of cases, the annotators had a preference for generated headlines. This means that for some examples, such a headline was perceived more easily

and naturally than the reference one. Nevertheless, human generated headlines were chosen in 36%. Although we cannot yet claim that our model is completely equivalent to how a human produces headlines for news messages, this result is already pretty promising.

5.3. Error Analysis

The neural network makes several types of errors. In Table 4, we provide some examples of generated headlines. The most common mistakes are incomplete phrases, as in examples *a, b, c*. In example *d,* there is a factual error, which brings to the erroneous understanding. there is a factual error. One more type of errors is grammar mistakes. Thus, in example *e*, the model produces a sentence with incorrect verbal government. Example *f* shows the use of an erroneous noun case-form.

Table 4
Examples with errors

a	Увольняемые сотрудники Ford в Ленобласти провели пикет против **«суверенного** (en) *Dismissed Ford employees in Leningrad Oblast held a picket against the **“sovereign***
b	Активиста «Весны» арестовали на 10 суток за акцию с манекенами на **Марс** (en) *The activist of “Spring” was arrested for 10 days for the action with mannequins on the Mars*
c	Россия с 1 апреля возобновляет регулярное авиасообщение с Германией, Шри-Ланкой и еще четырьмя (en) *Russia resumes regular flights with Germany, Sri Lanka and four other from April 1*
d	**Новостное сообщение:** На улице Тамбасова, 5 в **Красносельском районе** Петербурга в ночь с 31 января на 1 февраля произошел сильный пожар в павильоне киностудии «Ленфильм» … (en) **News text:** *On Tambasova Street, 5 in the Krasnoselsky district of St. Petersburg, there was a strong fire in the pavilion of the “Lenfilm” film studio on the night of January 31 to February 1…* **Сгенерированный заголовок:** В **Приморском районе** Петербурга произошел сильный пожар в павильоне «Ленфильма» (en) **Generated:** *In the Primorsky district of St. Petersburg there was a strong fire in the pavilion of “Lenfilm”*
e	Минобороны официально подтвердило **об** уничтожении военного штаба в Сирии
f	На мосту Александра Невского с грузовика упал меш**ка** с песком и цементом

6. Conclusion

In this paper, we explored the effectiveness of application of the fine-tuned pretrained Transformer-based model, that as a pretrained BERT uses RuBERT, to the task of neural generation of Russian news headlines. We showed that predicted headlines are highly grammatically and semantically coherent and resemble original news headlines. We also present a newly gathered Bumaga corpus and provide results achieved by the BertSumAbs model applied to generation of headlines for news articles from this dataset.

7. Acknowledgements

I would like to thank PhD, Associate Professor O.A. Mitrofanova (Saint Petersburg State University) for useful discussions and for comments that greatly improved this paper.

8. References

[1] H. Saggion, T. Poibeau, Automatic text summarization: Past, present and future, 2013. URL: https://hal.archives-ouvertes.fr/hal-00782442/document.
[2] D. Gavrilov, P. Kalaidin, V. Malykh, Self-Attentive Model for Headline Generation, 2019. URL: https://arxiv.org/abs/1901.07786.
[3] E. Alsentzer, A. Kim, Extractive Summarization of EHR Discharge Notes, 2018. URL: https://arxiv.org/abs/1810.12085.
[4] S. Xu, S. Yang, and F. C. M. Lau, Keyword extraction and headline generation using novel word features, in: AAAI, 2010, pp. 1461–1466.
[5] H.P. Luhn, The automatic creation of literature abstracts, in: IBM Journal of Research and Development, 2 (2), 1958, pp. 159-165.
[6] I. Sutskever, O. Vinyals, V. Le Quoc, Sequence to sequence learning with neural networks, 2014. URL: https://arxiv.org/abs/1409.3215.
[7] A.M. Rush, S. Chopra, J. Weston, A Neural Attention Model for Abstractive Sentence Summarization, 2015. URL: https://arxiv.org/abs/1509.00685.
[8] S. Takase, J. Suzuki, N. Okazaki, T. Hirao, M. Nagata, Neural headline generation on abstract meaning representation, 2016. URL: https://arxiv.org/abs/1603.06393.
[9] J. Gu, Z. Lu, H. Li, V.O. Li, Incorporating copying mechanism in sequence-to-sequence learning, 2016. URL: https://arxiv.org/abs/1603.06393.
[10] I.O. Gusev, Importance of copying mechanism for news headline generation, 2019. URL: http://www.dialog-21.ru/media/4599/gusevio-152.pdf.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: https://arxiv.org/abs/1706.03762.
[12] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, L. Kaiser, Universal transformers, 2018. URL: https://arxiv.org/abs/1807.03819.
[13] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805.
[14] Y. Liu, M. Lapata, Text Summarization with Pretrained Encoders, 2019. URL: https://arxiv.org/abs/1908.08345.
[15] A. See, P.J. Liu, C.D. Manning, Get to the point: Summarization with Pointer-Generator networks, 2017. URL: https://www.aclweb.org/anthology/P17-1099/.
[16] A. Bukhtiyarov, I. Gusev, Advances of Transformer-Based Models for News Headline Generation, 2020. URL: https://arxiv.org/pdf/2007.05044.pdf.
[17] Y. Kuratov, M. Arkhipov, Adaptation of deep bidirectional multilingual transformers for russian language, 2019. URL: https://arxiv.org/abs/1905.07213.
[18] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2015. URL: https://arxiv.org/abs/1412.6980.
[19] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, 2002. URL: https://www.aclweb.org/anthology/P02-1040/.
[20] C.Y. Lin, Looking for a few good metrics: ROUGE and its evaluation, 2004. URL: https://research.nii.ac.jp/ntcir/ntcir-ws4/NTCIR4-WN/OPEN/OPENSUB_Chin-Yew_Lin.pdf.