Headline Generation Shared Task on Dialogue'2019

Malykh V.A. (valentin.malykh@vk.com), Kalaidin P.S. (pavel.kalaidin@vk.com)

VK, Saint-Petersburg, Russia

We present results for the headline generation shared task for Russian language which has been held as a part of the Dialogue 2019 conference. There were six active participants. Each of six participants' systems was evaluated according to the mean value of ROUGE-{1,2,L} on the private part of the test set. The best system got a score of 23.142.

Key words: shared task, headline generation, text summarization, Russian language

Дорожка по генерации новостных заголовков в рамках конференции Диалог-2019

Малых В.А. (valentin.malykh@vk.com), Калайдин П.С. (pavel.kalaidin@vk.com)

ВКонтакте, Санкт-Петербург, Россия

В статье изложены результаты дорожки по генерации новостных заголовков на русском языке в рамках конференциии Диалог 2019 года. В дорожке активно участвовало 6 команд. Лучшее решение по метрике среднее ROUGE-{1,2,L} набрало 23.142.

Ключевые слова: дорожка, генерация новостных заголовков, суммаризация текстов, русский язык

1 Introduction

The summarization shared task is a favoured a natural language processing community. It promises to reduce the amount of time spent on time-consuming reading of long texts. However the Russian NLP community still lacks of an available dataset for common knowledge domain. To overcome this issue and improve the current state of Russian summarization technology, we propose a headline generation shared task. The headline for a news article often contains the main piece of information from the article. Since there in an almost endless amount of articles published every day, it is possible to state headline generation task with a little effort for data collection and annotation.

2 Format

We have chosen a non-traditional format for this shared task, namely the test set is hidden from the participants. To be able to test submitted solutions in such environment we needed to run the solutions without the intervention of their authors. We chose Docker engine[1] to provide such isolated environments. The Docker environments have been isolated from the Internet, i.e. had no Internet access from inside and no Internet addresses to access from outside. The only data which a solution had access to was the hidden (private) part of a test set. In addition, a solution was able and supposed to write its output to a specified file in the file system. Each container had been provided with the same amount of computational resources, including 1 NVIDIA GTX 1080 GPU, to run. The participants uploaded their Docker packed solutions to private Docker registry and could only schedule their solution for validation after the upload.

[1] https://docker.com

3 Datasets

Up to this year, there were no published datasets for headline generation task in Russian. There is an unofficial Lenta.Ru dataset (Lenta) which is available on GitHub[2]. Unfortunately, this dataset has no licensing attached, so one cannot use it for research purposes. Fortunately, the other Russian dataset has been published recently and described in a work [3]. It is "Rossiya Segodnya" news dataset (RIA), available also on GitHub[3]. This dataset contains 1 million news articles from Rossiya Segodnya from January of 2010 up to December of 2014.

This dataset has been made public entirely, so there was no hold-out part available to the shared task organizers to make a private test set. Thus, we had to use some other dataset as the test set. There is ROMIP news dataset (ROMIP), which is described in a work [9]. It is much smaller than "Rossiya Segodnya" dataset, it contains only 32 thousand of news articles. These articles were split into two parts. First 16 thousand made public test set, on which each participant could test their solution. And the rest (15789 articles) had been used as the private test set. The score for the test set was published only after the official evaluation end.

There is a significant quantitative difference between two datasets: in "Rossiya Segodnya" dataset news articles contain the first sentence with a place and date for an original news article. An example of such a phrase:

> МОСКВА, 21 августа 2015.

To overcome this discrepancy there was a special note for the participants on that. In addition, that fact was explicitly used in the baselines.

4 Baselines

Two baselines were presented for the shared task.

4.1 First Sentence

First sentence baseline has used the first sentence from an article as a hypothesis for its headline. As the first sentence, we have used the first meaningful sentence, skipping technical (and repeating) information. This baseline is common for headline generation task and has been used in many articles devoted to the task [6, 11, 12]. The code for this baseline has been made publicly available[4].

4.2 Neural Machine Translation

Another approach which has been used as a baseline solution is a neural machine translation. In this approach, we use the first meaningful sentence as a source language input for a machine translation model and a headline as a target language output. This approach was presented in work [6] for the English language, and had shown reasonable results for the Russian language in work [3]. We have used OpenNMT framework, described in work [7], for this baseline. The code for this baseline using PyTorch OpenNMT implementation also made publicly available[5].

[2] https://github.com/yutkin/Lenta.Ru-News-Dataset
[3] https://github.com/RossiyaSegodnya/ria_news_dataset
[4] https://github.com/deepvk/headline_gen_first_sent
[5] https://github.com/deepvk/headline_gen_onmt

5 Metric

Traditionally, for summarization tasks, for which a headline generation task is a special case, the ROUGE metric is used. ROUGE metric was described in [8]. Essentially, the BLEU metric is counting common token sequences in ground truth and hypothesis sequences. There are three main variants: BLEU-1, BLEU-2, and BLEU-L. BLEU-1 and BLEU-2 are using unigrams and bigrams respectively to compute a score. BLUE-L is using longest common subsequence for a reference and a hypothesis to compute the score. Here are the formulae for ROUGE metrics from original paper [8]:

$$\text{ROUGE-N} = \frac{\sum_{r \in \{\text{references}\}} \sum_{w \in r} \text{Match}(w)}{\sum_{r \in \{\text{references}\}} \sum_{w \in r} \text{Count}(w)}, \tag{1}$$

where N stands for the length of a n-gram w, `Match` is the maximum number of n-grams co-occurring in a candidate summary (hypothesis) and a set of reference summaries, and `Count` is a number of all n-grams in references' set. We use ROUGE-1 and ROUGE-2 for unigrams and bigrams comparison respectively.

$$\text{ROUGE-L} = \frac{\sum_{r \in \{\text{references}\}} \text{LCS}(r, h)}{m}, \tag{2}$$

where `LCS` is longest common subsequence in terms of words for r (a reference) and h (a hypothesis), and m is summed number of words for all references in the set.

As one can see ROUGE metrics are in fact classification metrics, and exactly Precision metrics. Analogously, there were proposed variants of ROUGE using Recall and their geometric mean F_1-measure. So, in the end, there are 9 variants of ROUGE metric, every single of them could be useful to score some aspects of hypothesis quality. Keeping this in mind, we decided to:

- use F_1 variants for ROUGE-{1,2,L}, as aggregated metrics for precision and recall variants, and
- make an aggregated metric of three mentioned F_1 metrics.

We ended with following formula:

$$score(r, h) = \frac{1}{3N} \sum_{i=1}^{N} (\text{ROUGE-1}(r_i, h_i) + \text{ROUGE-2}(r_i, h_i) + \text{ROUGE-L}(r_i, h_i)), \tag{3}$$

where r and h are references and hypotheses sets of length N each respectively.
All the scores in the section 7 are using this metric if other is not specified.

6 Participants

Four of our participants have provided the information of developed architectures. In this section, we describe the first four architectures.

6.1 DreamTeam

DreamTeam team's approach is called Phrase-Based Attentional Transformer. This approach is based on kernel convolutions for attention mechanism presented in [10]. These attention modification is formulated as following:

$$\text{QUERYK}(Q, K, V) = (\frac{\text{Conv}_n(KW_k, QW_q)}{\sqrt{d_k * n}})\ \text{Conv}_n(V, W_v), \tag{4}$$

where $W_q \in {}^{n \times d_q \times d_k}, W_k \in {}^{d_k \times d_k}, W_v \in {}^{n \times d_v \times d_v}$ are trainable weights. This modifies original attention mechanism (as described in Transformer paper [16]) using convolutions over attention hidden states.

Interestingly, the authors report a new state of the art results for the proposed model on RIA dataset shown in Tab. 1. More details on this model could be found in a work [14].

6.2 Black and Yellow

Black and Yellow team is using CopyNet mechanism in their approach. This mechanism was originally presented in the paper [4] and has shown itself as helpful for the text summarization task. The idea of this mechanism could be formulated as follows: there are a lot of rare words in the texts, which could be presented only once in the dataset. These words are really hard to predict and it could be better to just copy them from the input to the output. To become capable of copying a model should have an additional output which predicts the probability of copy operation for the specified word. A graphical representation of CopyNet model could be found in the Fig. 1.

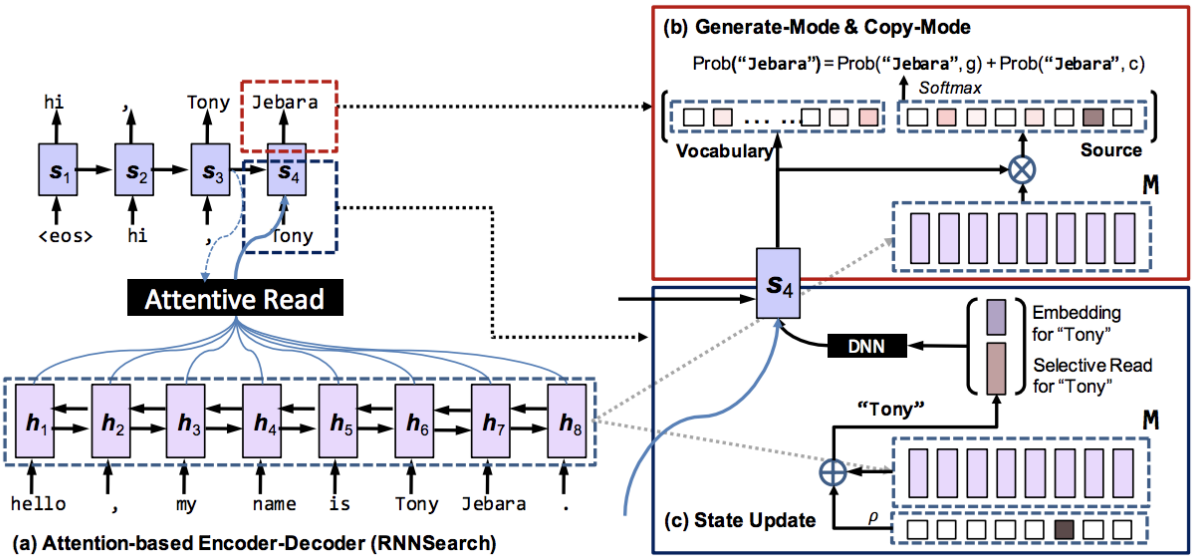

Figure 1: CopyNet model architecture from [4]

We should mention that Black and Yellow team has used Lenta dataset to tune their model. The team also provided testing results on RIA dataset which could be found in Tab. 1. The details and additional results for the Black and Yellow team solution could be found in [5].

6.3 Burning Headlines

Burning Headlines team has used Pointer-Generated networks initially presented in [13]. The approach of pointer generation is close to CopyNet approach described in section 6.2, so we address a reader to the team's publication [15] and original work for details. It is important to mention that there were proposed some extensions to the Pointer Generation model. These are a usage of grammemes and specific stem+flexion embeddings. The grammeme is a word

with an explicitly stated grammatical form. Authors use a word as an input and predict a grammeme for it. In the Fig. 2 one could find a scheme for this proposed extension.

According to authors the best score by shared task metric had shown the vanilla Pointer-Generation model, which results is shown in Tab. 2. Interestingly, authors also have used Lenta corpus and provide extension testing results on it. This and other details could be found in the authors report [15].

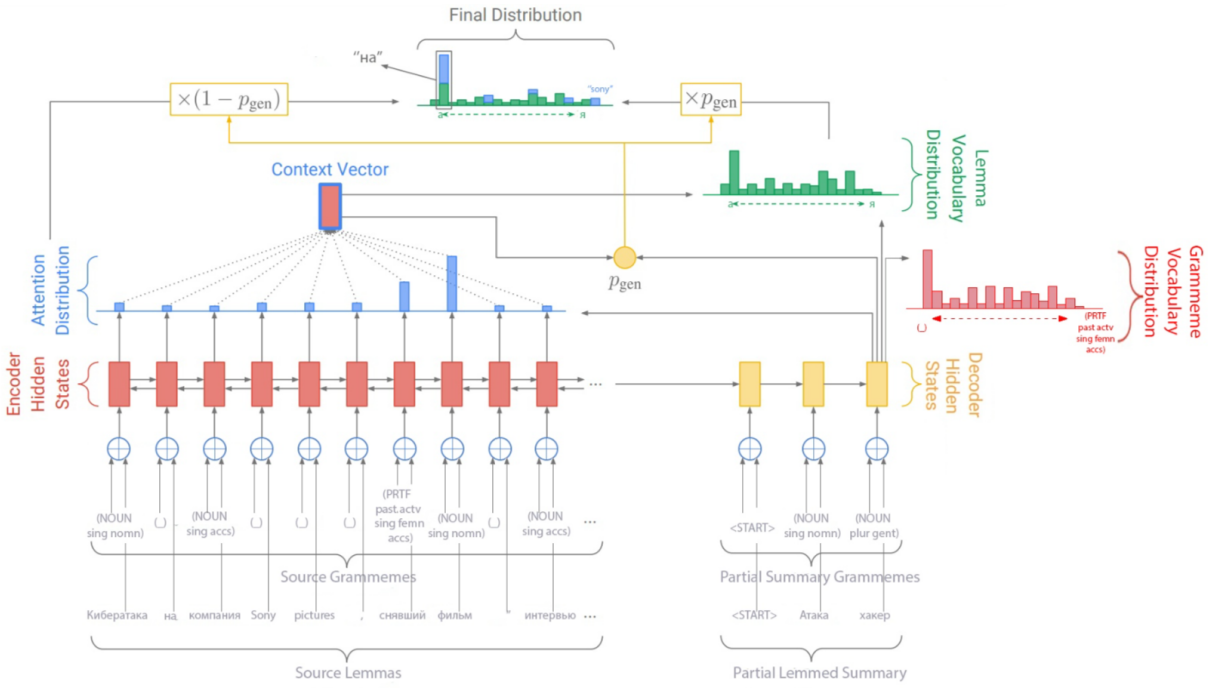

Figure 2: Grammeme using extension for Pointer Generation model from [15].

6.4 Symmetrical potato

Symmetrical potato team has used a machine translation baseline code described in section 4.2. They achieved a boost in the performance of that model using other byte pair encoding trained on Wikipedia dump. More details about their experiments and results could be found in the paper [1].

7 Results

There are two types of results which we are presenting in this section: the reported results for participants' models (Tab. 1) and the official shared task testing scores (Tab. 2). Unfortunately, the results in the former table aren't directly comparable due to differences in the test environments, so they are shown here for reference and we address the reader to participants' reports for the environment details.

Model	ROUGE-1-f	ROUGE-2-f	ROUGE-L-f
DreamTeam	42.96	25.43	40.02
Black and Yellow	41.61	24.46	38.85
Symmetrical potato[6]	39	22	36

Table 1: Results of participants' models on RIA dataset.

The official testing results for the shared task has been presented in Tab. 2. These results have been achieved by participants solutions on the unseen (private) part of the test set in the same environment.

Team	Score
DreamTeam	20.267
Black and Yellow	**23.142**
Burning Headlines	20.293
Symmetrical potato	20.268
Зульфат Мифтахутдинов	20.267
L&M	20.267

Table 2: Official shared task results on the private test set.

As one could see from the Tab. 2 there are three participants with the same score on the private test set. There three participants sent the first sentence baseline as their solution. The other their submits hadn't shown improvement over this baseline.

8 Conclusion

The shared task has shown that there is an interest in the headline generation task in the Russian natural language processing community. In the participants' solutions we could see different approaches both presented in the literature and original ones. Some of the presented models show improvement over a universal transformer model, originally described in the work [2] and adopted for this dataset in the work [3]. And we should mention that the first sentence baseline again had been proved as a strong baseline for headline generation task.

It is interesting to mention that an approach which shown best results on the RIA dataset had shown mediocre results on the private test set, taken from ROMIP dataset.

The used format had shown itself as appropriate and convenient for the participants, so we hope to use it in the next challenges and that it will be adopted for other shared tasks in the future.

As future work authors see usage of other news sources, on the different languages, like Belorussian closely related to the Russian language. The task of cross-lingual summarization for related languages has not been solved yet.

Acknowledgements

The authors are thankful to the shared task co-organizers Ivan Smurov and Ekaterina Artemova for useful discussions during shated task run, and also to Ira Shubina and Ivan Karabakin who helped with platform for participants' submissions.

References

[1] Nikita Churikov and Elena Sannikova. Headline generation: first sentence vs neural machine translation. In *Computational Linguistics and Intellectual Technologies*, 2019.

[2] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[6] The authors provided the results without decimal places after the dot.

[3] Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. Self-attentive model for headline generation. In *Proceedings of the 41st European Conference on Information Retrieval*, 2019.

[4] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1631–1640, 2016.

[5] Ilya Gusev. Importance of copying mechanism for news headline generation. In *Computational Linguistics and Intellectual Technologies*, 2019.

[6] Yuko Hayashi and Hidekazu Yanagimoto. Headline generation with recurrent neural network. In *New Trends in E-service and Smart Computing*, pages 81–96. Springer, 2018.

[7] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*, 2017.

[8] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[9] Igor Nekrestyanov and Marina Nekrestyanova. Overview of the romip'2006. In *Proceedings of Fourth Russian Seminar of Information Retrieval Methods Evaluation (ROMIP)*, 2006.

[10] Phi Xuan Nguyen and Shafiq Joty. Phrase-based attentions, 2018.

[11] Jan Wira Gotama Putra, Hayato Kobayashi, and Nobuyuki Shimizu. Experiment on using topic sentence for neural news headline generation. 2018.

[12] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Empirical Methods in Natural Language Processing*, pages 379–389, 2015.

[13] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017.

[14] Andrej Sokolov. Phrase-based attentional transformer for headline generation. In *Computational Linguistics and Intellectual Technologies*, 2019.

[15] Matvey Stepanov. News headline generation using stems, lemmas and grammemes. In *Computational Linguistics and Intellectual Technologies*, 2019.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.