Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

SUMMARY CONSTRUCTION STRATEGIES FOR HEADLINE GENERATION IN THE RUSSIAN

Malykh V. (valentin.malykh@phystech.edu)
Kazan Federal University, Kazan, Russia

Cherniavskii D., Valukov A.
Moscow Institute of Physics and Technology, Moscow, Russia

Key words: headline generation, summarization, Russian

DOI: 10.28995/2075-7182-2020-19-570-578

СТРАТЕГИИ СОСТАВЛЕНИЯ РЕФЕРАТОВ В ЗАДАЧЕ ГЕНЕРАЦИИ ЗАГОЛОВКА ДЛЯ РУССКОГО ЯЗЫКА

Малых В. (valentin.malykh@phystech.edu)
Казанский (Приволжский федеральный университет, Казань, Россия

Чернявский Д., **Валюков А.**
Московский физико-технический институт (научно-исследовательский университет), Москва, Россия

1. Introduction

In the modern world, texts are plenty in the everyday life of a person—the news articles, blogs, social networks. These texts could be long, for example, the typical length of a New York Times news article is more than 700 words [13]. The reading process could take significant time for even one article, so this raises a question of shortening this time. To handle the mentioned issue there were proposed techniques of extractive and later abstractive text summarization, i.e. the generation of a short text summary using longer original text.

There is an issue with most of abstractive and some of extractive summary generation strategies, they all need a training set, which could take time and labour to create, like CNN/DailyMail dataset initially presented in [5] and compiled for text summarization task in [10]. To overcome this issue there was presented a separate task of headline generation for news documents. Since the news documents are plenty, and they could be used with ease.

The headline generation task could be considered as a two-stage task. On the first stage, a summary of the article body is constructed and on the second stage, the headline is generated using the constructed summary. In this work, we concentrated on a headline generation task for the Russian language in an aspect of comparison summary construction techniques.

This work is composed as follows: related work, dataset and metrics description, base models description, summary strategies, experiments, and conclusion.

2. Related Work

There were already successive attempts in headline generation for different languages. For the English language, there are several works. The authors of [12] were to the best of our knowledge the first to apply neural networks to headline generation. In more recent work of Hayashi et al. [4], an encoder-decoder approach was presented, where the first sentence was reformulated to a headline. The related approach was presented in [11], where the approach of the first sentence was expanded with a so-called topic sentence. The topic sentence is chosen to be the first sentence containing the most important information from a news article (so-called 5W1H information, where 5W1H stands for who, what, where, when, why, how). This approach has a limitation that these sentences should be marked up beforehand. Tan et al. [18] present an encoder-decoder approach based on a pre-generated summary of the article. The summary is generated using a statistical summarization approach.

For the Russian language, there are a few works on this topic. In the work of [2] there were presented universal Transformer model, which used whole article body as input to generate a headline. The other works [3], [14], [16], which were resulted the shared task on headline generation, described in [8]. Sokolov and Stepanov in [14], [16] have used copy mechanism in encoder-decoder models to improve quality of the generation, while Gusev in [3] invoked phrase-based attention mechanism to improve the Transformer model itself. It need to be mentioned that all the previous works were using whole article body to generate headline hypothesises.

3. Dataset & Metrics

In this work, we decided to explore different summary construction strategies for Russian language dataset. There is only one dataset of significant size for Russian. It is "Rossiya Segodnya" News Dataset described in [2]. There are 1,003,869 news articles in the corpus with a mean title length 9.5 words, mean text length of 315.6 words, and mean 15.0 sentences. Following [2] we divided the dataset into three parts: 10,000 news documents were withheld as a validation set, 20,000 ones as a test set, and the rest was considered as a train set.

We are using ROUGE metric family, presented in [7]. Essentially, the ROUGE metric is counting common token sequences in ground truth and hypothesis sequences. There are three main variants: ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 and ROUGE-2 are using unigrams and bigrams respectively to compute a score. ROUGE-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 [7]:

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

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 in a set of reference summaries, and **`Count`** is a number of all n-grams in references' set.

In particular, the ROUGE-N formulae mentioned above are describing how much the hypothesis is capturing the reference summary and is often referred as the recall variant of ROUGE-N metrics, or simply ROUGE-N-Recall. As there are no control over the length of the hypothesis, so it can capture almost all of the reference summary while being excessively long. This issue is solved by the precision modification of ROUGE-N metrics that has the same formulae but the **`Count`** variable is now referred to the number of all n-grams in hypothesis' set. The ROUGE-N-F1 score is calculated as classical F_1 measure with ROUGE-N-Precision and ROUGE-N-Recall using harmonic mean.

In addition to ROUGE, we decided to use an extraction score, presented in [1]. The extraction score is a metric of extractiveness of a summary. It searches for the long substrings from a source text in the summary. Extraction score is defined as follows:

$$ext\_score(S) = \sum_{s \in P(ACS_s)} s \times (e^{s-1} - \frac{1-s}{e}), \tag{2}$$

where S is a summary, a ACS_s is the set of all long non-overlapping common sequences between S and the document, $P(ACS_s)$ is a set, where each element is the length of a common sequence divided by the length of the summary.

4. Base Models

We have conducted experiments with two basic models which follow Encoder-Decoder approach presented in [17]. One is a recurrent neural network and another is Transformer network, described in [19], which are described below in more details.

4.1. Recurrent Model

The Encoder and the Decoder are both complied of two-layer bidirectional Long Short-Term Memory (LSTM) [6] recurrent networks. The encoder network recursively receives news body words (in both direction) as input and produces hidden states, one for each input word (since there are two reading directions, there are two hidden states for each word, these hidden states are concatenated to produce a whole hidden state for a word). Afterwards, it passes its final hidden state as the initial hidden state to the decoder network, and for each decoder step, an attention distribution is calculated upon the encoder hidden states. This distribution is used to predict the next word of the news headline. At the beginning of prediction, the decoder network receives a special <START> token, and later it uses the previously generated word as an input. This procedure is illustrated in **Fig. 1**.

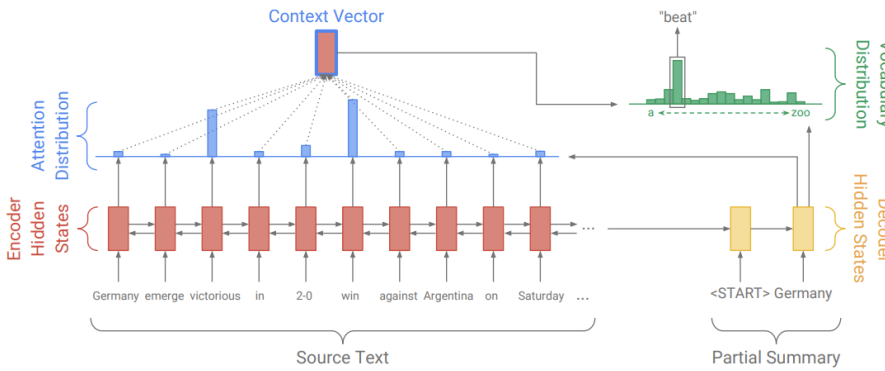

Figure 1: Seq2seq with attention from [14]

4.2. Transformer

In this case, the network receives the news body text not recursively but as a whole. The encoder and the decoder network both consist of 6 identical layers, described below. Firstly, the positional encoding is added to the input embeddings to ensure the difference of embeddings in a different part of the text. Then the multi-head self-attention is calculated upon the input. This attention is then added to the input of the multi-head module and a layer normalization is applied. After it is passed to the Feed Forward network. Final encoder layer output is then passed to each of layer the decoder network to the encoder-decoder attention block, which comes after the self-attention block. Complete architecture is shown in **Fig. 2**.

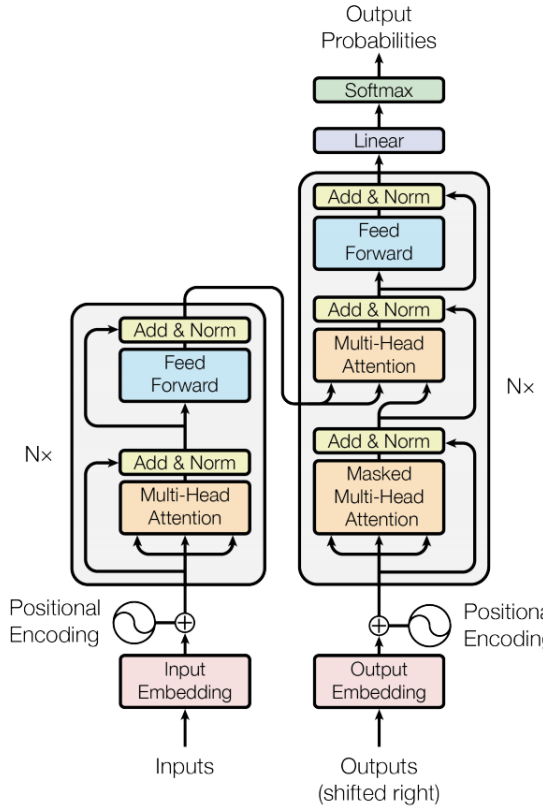

Figure 2: Transformer architecture from [19]

5. Summary Construction Strategies

In this section, we describe different strategies of summary construction for headline generation. The constructed summaries are used as input for encoder-decoder models.

Whole body. This is the simplest strategy, i.e. to use the whole article body text to generate a headline. This approach is the most common in the literature, e. g. [12]. This approach mentioned as "*full*" in **Tab. 1**.

First Sentence. The first sentence could be a strong hypothesis for the headline, as shown in [2]. We should mention that in "Rossiya Segodnya" news documents the first sentence is a formal statement of date and place, for example:

МОСКВА, 21 августа 2015.

We skip this formal intro and use the next informative sentence as the first sentence in our setup. This approach mentioned as "*1sent*" in **Tab. 1**.

Three First Sentences. The first sentence baseline although showed good performance could not contain enough information for headline generation. For example, in paper [18] authors stated that the first sentence is not informative enough. So we decided to use the first three sentences as a summary. Again, the formal intro is skipped in this setup. This approach mentioned as "*3sent*" in **Tab. 1**.

Unsupervised Summary. Work [18] showed that unsupervised summary could be a good hypothesis for a headline. In our work we have used classic TextRank algorithm, described in publication [9], to generate a summary from an article body. This approach mentioned as "*unsup*" in **Tab. 1**.

NER Summary. We propose a novel approach to construct a summary using named entity recognition (NER). We use PullEnti pre-trained model [15] to mark up the "Rossiya Segodnya" corpus. The mark up contains mentions of persons, organizations and locations. To construct the summary we extracted the sentences which contain at least one named entity. This approach mentioned as "*NER*" in **Tab. 1**.

Table 1: ROUGE-1,2,L scores for Recall (r) and F-measure (f) variations; also extraction score[1]

Model | Metric	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r	ext. score
1sent	23.395	44.055	10.302	20.716	16.291	40.390	0.427
3sent	15.235	**53.039**	5.836	24.089	8.698	**49.656**	0.477
unsup	14.095	48.003	5.110	20.286	8.507	44.772	0.367
NER	12.499	36.168	4.124	13.797	7.797	33.362	**0.241**
Seq2seq+1sent	39.866	38.671	23.111	22.480	37.058	36.758	0.551
Seq2seq+3sent	42.545	41.584	25.131	***24.668***	39.613	39.539	0.627
Seq2seq+full	41.927	40.641	24.639	23.944	39.002	38.663	0.582
Seq2seq+unsup	36.147	35.093	19.643	19.134	33.448	33.223	0.425
Seq2seq+NER	25.556	24.104	13.142	12.547	23.287	22.884	***0.269***
Transformer+1sent	41.075	40.557	24.593	24.372	38.319	38.488	0.719
Transformer+3sent	***42.922***	***41.863***	**25.476**	**24.908**	***39.996***	***39.784***	0.673
Transformer+full	39.627	37.945	21.153	20.328	36.525	35.852	0.423
Transformer+unsup	34.090	32.764	17.583	16.967	31.422	30.936	0.363
Transformer+NER	28.501	27.688	14.705	14.387	26.298	26.142	0.379
Gavrilov et al. [2]	39.75	37.62	22.15	21.04	36.81	35.91	—
Sokolov [14]	**42.96**	—	***25.43***	—	**40.02**	—	—
Stepanov [16]	25.23	25.79	10.33	10.60	22.82	24.08	—
Gusev [3]	41.61	40.33	24.46	23.76	38.85	38.51	—

[1] For the ROUGE scores higher is better, for extraction score lower is better. **Bold** marks up the best result, while ***Italic-Bold*** marks up the second best result. All the metrics are computed on test set.

6. Experiments

As a Transformer model we have trained a basic 6-layer Transformer architecture model with 8 heads. The dimension of fully connected layer was 2,048. We trained it with a batch size of 4,096 for 110k training steps. For Seq2seq model we took a default 2-layer LSTM with 500 hidden units on both of the encoder and decoder. To train these models we used single nVIDIA Titan X (Pascal) GPU with 12Gb of RAM.

6.1. Results

The evaluation scores for both architectures and all summary construction summaries are presented in Tab. 1. The best results by recall in ROUGE-1 and ROUGE-L are showed by the 3sent baseline. This fact could be considered trivial. But interestingly, the Transformer approach over 3sent summary achieves best results by ROUGE-2 metric, including recall score. In addition this model shows the second best results in ROUGE-1 and ROUGE-L both recall and F-measure. We interpret this as Transformer is being very extractive—its variants are consistently more extractive than other two approaches, and 3sent baseline has the highest scores for the recall, so Transformer has a better choice to copy from the input text. Interestingly, the Transformer model achieves the best performance in terms of ROUGE-2-Recall, even better than 3sent baseline. We also could draw the reader's attention to the fact that Transformer models lower extraction score with an extension of input text from one first sentence to the whole text. While Seq2seq models do not follow this regularity.

Regarding the extraction score, the lowest one is demonstrated by NER baseline, but this also accompanied with the lowest ROUGE metrics. The Seq2seq approach over NER basic summary drastically improves the ROUGE results, but also gain some extraction score, showing the second best one. Interestingly, a Transformer model has a much higher extraction score with this summary as input.

Table 2: Samples of headlines generated by the studied models. We present only unique generated headlines

Original text, truncated:	пожар, произошедший в среду в ресторане в центре москвы, ликвидирован, пострадавших нет, сообщил риа новости источник в правоохранительных органах столицы. «пожар в ресторане „эль гаучо“ на садовой-триумфальной улице в двухэтажном здании ликвидирован. по предварительным данным, горели жировые отложения в вентиляции. возгорание произошло в вентиляционной системе», — сказал собеседник агентства. в настоящее время причины пожара устанавливаются. по данным представителя мчс, сообщение о пожаре поступило на пульт дежурного «01» в 21.25 мск. он отметил, что, благодаря своевременной эвакуации, никто из посетителей и сотрудников ресторана не пострадал.
Original headline:	пожар в ресторане в центре москвы ликвидирован, никто не пострадал
Transformer+1sent:	пожар в ресторане в центре москвы потушен
Transformer+3sent:	пожар в ресторане в центре москвы ликвидирован, пострадавших нет
Transformer+full:	пожар в ресторане в центре москвы ликвидирован
Transformer+ner:	пожар в центре москвы потушен

Some samples of headlines produced by different models are presented in Tab. 2. As one could see, the approaches are differ with details, and quality of a headline is seemingly correlated with BLEU score, for example, the named entity recognition approach for summary constructing (the worst one by BLEU score) suffers from lack of useful words, such as «потушен» or «ликвидирован».

7. Conclusion

We have presented a comparison of summary construction strategies, where the constructed summaries are used as input for headline generation. We have studied the classic first sentence strategy and extended it to the three first sentences one. The latter strategy shows the best performance by the means of recall itself and also gives a boost for Transformer architecture model which achieves new state of the art ROUGE-2 results. This model outperforms other approaches even those which are using whole text as input. The Seq2seq models are consistently gaining lower ROUGE scores in all variants in comparable setups, although they have lower extraction score also.

As the direction for future research authors see two main ones: an application of the proposed approach to other languages, and its modification for abstractive summarization task itself, which has significantly different text structure and so states open question of applicability of the proposed approach.

7.1. Acknowledgements

The work of the first author was funded by RFBR, project number 19-37-60027.

References

1. *Cibils, A. et al.:* Diverse beam search for increased novelty in abstractive summarization. CoRR. abs/1802.01457, (2018).
2. *Gavrilov, D. et al.:* Self-attentive model for headline generation. 41st European Conference on Information Retrieval. (2019).
3. *Gusev, I.:* Importance of copying mechanism for news headline generation. In: Computational linguistics and intellectual technologies. (2019).
4. *Hayashi, Y., Yanagimoto, H.:* Headline generation with recurrent neural network. In: New trends in e-service and smart computing. pp. 81–96 Springer (2018).
5. *Hermann, K. M. et al.:* Teaching machines to read and comprehend. In: Advances in neural information processing systems. pp. 1693–1701 (2015).
6. *Hochreiter, S., Schmidhuber, J.:* Long short-term memory. Neural computation. 9, 8, 1735–1780 (1997).
7. *Lin, C.-Y.:* Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out. (2004).
8. *Malykh, V., Kalaidin, P.:* Headline generation shared task on Dialogue'2019. In: Proceedings of the international conference "Dialogue 2019". (2019).

9. *Mihalcea, R., Tarau, P.:* Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. (2004).
10. *Nallapati, R. et al.:* Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of the 20th SIGNLL conference on computational natural language learning. pp. 280–290 Association for Computational Linguistics, Berlin, Germany (2016).
11. *Putra, J. W. G. et al.:* Experiment on using topic sentence for neural news headline generation. In: Proceedings of 24th annual conference of japanese association for natural language processing. (2018).
12. *Rush, A. M. et al.:* A neural attention model for abstractive sentence summarization. In: Empirical methods in natural language processing. pp. 379–389 (2015).
13. *Sandhaus, E.:* The new york times annotated corpus ldc2008t19. DVD. Linguistic Data Consortium, Philadelphia (2008).
14. *Sokolov, A.:* Phrase-based attentional transformer for headline generation. In: Computational linguistics and intellectual technologies. (2019).
15. *Starostin, A. et al.:* Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016". (2016).
16. *Stepanov, M.:* News headline generation using stems, lemmas and grammemes. In: Computational linguistics and intellectual technologies. (2019).
17. *Sutskever, I. et al.:* Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014).
18. *Tan, J. et al.:* From neural sentence summarization to headline generation: A coarse-to-fine approach. In: Proceedings of the 26th international joint conference on artificial intelligence. pp. 4109–4115 AAAI Press (2017).
19. *Vaswani, A. et al.:* Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017).