Extractive Russian Text Summarization as Greedy Sentence Sequence Continuation Search with Probabilities from Pretrained Language Models

Irina Polyakova
Department of Algorithmic Languages
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
polyak@cs.msu.ru

Semyon Pogoreltsev
Department of Algorithmic Languages
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
pogorelcevsa@gmail.com

***Abstract*—Pretrained language models based on Transformer have defined new state-of-the-art result on a wide range of tasks being finetuned or used as embedders. Models with Multi-head Self-Attention mechanism have defined a new standart of quality on text summarization task in English, while Transformer based models potential for extractive Russian text summarization has been barely explored. We propose a new method for extractive Russian text summarization, reducing the task to the selection of the most probable sequence of sentences. The new method beats ROUGE-1 and ROUGE-L scores of other models such as SummaRuNNer, and mBART evaluated on Gazeta dataset and is more preferred in human evaluation poll.**

***Keywords*—Russian Text Summarization, Pretrained Language Models, BERT, Sentence-BERT, Next Sentence Prediction**

I. Introduction

Text summarization is the task of creating a shorter version of a document that captures essential information. Automatic summarization approaches can be extractive or abstractive.

Extractive methods form a summary as a combination of the original text's chunks. Extraction is usually reduced to classifying sentences of the initial document. The resulting summary is grammatically correct, especially in the case of sentence copying. But extractive methods can't produce generalized and paraphrased text, which is essential for high-quality compression. In addition, these algorithms usually don't structurize summaries.

Abstractive approaches generate a new text - a generalization of the original one's ideas. These models can generate new words that are not from the original text, which leads to better generalization abilities. It allows them to compress text in a better way via sentence fusion and paraphrasing, but overall complexity, errors in generated texts' grammar, unexpected results in some cases and weak interpretability still limit the use of these methods in various projects.

We introduce a new extractive summarization approach for Russian texts, which leverages pretrained BERT(Bidirectional Encoder Representations from Transformers) [1] and Sentence-BERT [2] models. The proposed method forms a summary that has human-natural storytelling order of sentences. The algorithm can easily be adapted for other languages or be multilingual provided that there are such BERT models.

II. Related works

A. Pretrained language models

Pretrained language models such as BERT [1] are now a key technology in NLP industry being extensively used with finetuning, few-shot learning, or as embedders. State-of-the-art approaches in most NLP tasks are based on neural networks with Transformer [3] architecture. The main building block of the Transformer model is the Multi-head Self-Attention mechanism

$$\mathrm{MHA}(Q,K,V) = \mathrm{Concat}(head_1, ..., head_n)W^o \quad (1)$$

$$\mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\,(\frac{Q*K^T}{\sqrt{d_k}})V \quad (3)$$

where V - values, K - keys, Q - queries, $d_k = dimK$ In the case of self-attention Q, K, V are from the same source, so relations between tokens of the sequence are estimated.

"Bidirectional Encoder Representations from Transformers(BERT)" [1] is a new language model which is trained with a masked language modeling and a "next sentence prediction” tasks on large unstructured single language or multilingual text corpora. BERT is a stack of Transformer encoders.

Sentence-BERT [2] is a modification of the pre-trained BERT model that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings. These semantic vectors are meant for comparison using cosine-similarity.

“Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language” [4] proved that Transformer-based models are useful for Russian. The authors of the paper also prepared pre-trained BERT models’ weights for Russian.

B. Extractive Summarization

SummaRunNNer [5] is one of the first approaches, using neural networks for extractive text summarization. It leverages an RNN-based encoder on sentence and document levels for semantic embeddings which are then used for binary sentence classification.

Refresh [6] model from “Ranking Sentences for Extractive Summarization with Reinforcement Learning” [6] uses CNN and RNN layers for sentence and document embedding. Refresh is trained with the new algorithm proposed in the paper, which globally optimizes ROUGE [7] evaluation score through a reinforcement learning objective.

“Single Document Summarization as Tree Induction” [8] leverages structured attention to induce a multi-root dependency tree representation of the document while predicting the output summary.

This idea was further studied in “HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization” [9].

Methods, proposed in the “Text Summarization with Pretrained Encoders” [10] made a big leap in ROUGE [7] metrics on the CNN/Daily Mail dataset. The extractive model used pre-trained BERT model, finetuned with inter sentence Transformers for sentence classification.

The current state-of-the-art extractive model on the CNN/Daily Mail was introduced in ‘Extractive Summarization as Text Matching” [11] work. Sentence extraction task is reduced to semantic text matching with Sentence-BERT embeddings.

III. Sentence-BERT for Text Centrality Ranking

The algorithm uses cased Sentence-BERT, which was fine-tuned on SNLI [12] Google-translated to Russian and on the Russian part of XNLI [13] dev set.

Sentence embeddings from Sentence RuBERT are being ranked by cosine-similarity

$$\mathrm{sim} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}} \tag{4}$$

where A, B are the semantic representation vectors that are compared. Cosine-similarity is considered a default method f or semantic similarity comparison used for similar tasks, such as NLI(Natural Language Inference).

This metric is human-interpretable.

$$\mathrm{sim} \in [0, 1] \tag{5}$$

Higher cosine-similarity means that sentences are closer.

Central text embedding text-central is a selected sentence semantic vector or mean pooled representation from all sentences of the document. The experiments have show that for news texts it’s better to use the first sentence or mean pooling of all sentences and the last option is better in terms of ROUGE [7] scores.

$$\text{text-central} = \frac{1}{n} \sum_{i=0}^{n} \text{sent-emb}_i \tag{6}$$

where n is the number of sentences in text, sent-emb_i is the Sentence-BERT embedding of i-th sentence in the text

Then for each i-th sentence r-score_i is computed

$$\text{r-score}_i = \mathrm{sim}(\text{s-emb}_i, \text{text-center}) \tag{7}$$

where $i = 1, 2, ..., n$ These scores are min-max normalized

$$\overline{\text{r-score}_i} = \frac{\text{r-score}_i - \min * \text{r-score}_i}{\max * \text{r-score}_i - \min * \text{r-score}_i} \tag{8}$$

$\text{r-score}_i'$ is a normalized centrality measure, which is used in a latter formula.

IV. Sequential Next Sentence Prediction

A. Next Sententence Prediction

BERT models are pre-trained on the task of Next Sentence Prediction(NSP) as well as Masked Language Modelling. In the BERT training process, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document. During training, 50% of the inputs are a pair in which the second sentence is the subsequent sentence in the original document, while the other part consists of random sentences from the corpus. The assumption is that the random sentence will be disconnected from the first sentence.

The following steps are needed to estimate a probability that the sentence B is a continuation of sentence A:

1) A [CLS] token is inserted at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence.
2) A sentence embedding indicating Sentence A or Sentence B is added to each token. Sentence embeddings are similar in concept to token embeddings with a vocabulary of 2.

3) A positional embedding is added to each token to indicate its position in the sequence
4) The entire input sequence goes through the Transformer model.
5) The output of the [CLS] token is transformed into a 2×1 shaped vector, using a simple classification layer.
6) Calculating the probability with softmax.

Extractive summarization can be reduced to greedy sequential sentence selection maximizing the probability of each chosen sentence starting from the one that is the best generalization of the text. With this approach sentences are also ordered in a natural for human reading way, forming a story.

This heatmap(1) is an example of matrix on 5 sentence sample from the Gazeta [14] dataset. Each element is computed with formula (9)

$$\overline{P}(s_i|s_j) = \frac{P(s_i|s_j) - \min *P(s_i|s_j)}{\max *P(s_i|s_j) - \min *P(s_i|s_j)} \tag{9}$$

$$P(s_i|s_j) : i, j \in [0, n), i, j \in \mathbb{N}, i \neq j \tag{10}$$

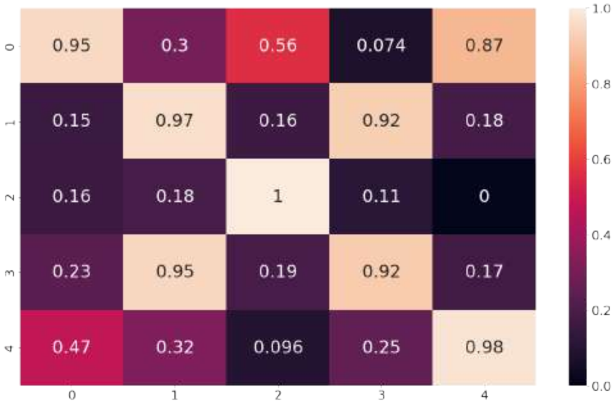

Figure 1. Example of a min-max normalized NSP matrix heatmap

B. Contextualized Next Sentence Prediction

In original NSP it's supposed that comparing a pair of sentences is made without context. For extractive summarization sentences which were already selected for the summary can be used. Self-Attention mechanisms benefit from additional context, so probabilities estimation is more precise in terms of summary, not just the last selected sentence.

$$P(s_i|s_1, s_l, ..., s_k) = \text{NSP}(C + s_1 + \sum_{j=l}^{k}(s_k) + S + s_i + S) \tag{11}$$

s_i is the i-th sentence, C is a [CLS] token, S is a [SEP] token, $l \in [2, k-1], l \in \mathbb{N}$ 1 is a skipping numerator, $k \in [0, m-1], k \in \mathbb{N}$ is a number of sentences in the summary that were already chosen, except the first one, m is a number of sentences that is expected to be in a summary.

l sentences between the start and the last selected are skipped if the sequence of tokens exceeds the input size of BERT. The first and the last ones are always in the context for probability estimation, because the first one is the most generalizing and the last is the one that should be naturally continued.

The probabilites are also min max normalized

$$\overline{P}(s_i|s_1, s_l, ..., s_k) = \frac{P(s_i|..., s_k) - \min *P(s_i|..., s_k)}{\max *P(s_i|..., s_k) - \min *P(s_i|..., s_k)} \tag{12}$$

where min, max are selected from the probabilites of one step. It's generally better to normalize on the whole matrix, but to reduce computational complexity and memory usage only the list of next possible sentences is used.

C. Extractive Summarization of News Texts

The summary is formed iteratively with the sentences that locally maximise the $\text{SC}(s_1, ..., s_k \rightarrow s_i)$ score

$$\text{SC}(s_1, ..., s_k \rightarrow s_i) = \overline{P}(s_i|s_1, s_l, ..., s_k) + \alpha * \overline{\text{r-score}_i} \tag{13}$$

where α is a coefficient of central sorting importance, by default $\alpha = 0.05$,

The starting element is the first sentence of the text, because in news texts, like the ones in the Gazeta [14] dataset, it's the most important, generalizing the whole text.

Using $\text{r-score}'_i$ in the $\text{SC}(s_1, ..., s_k \rightarrow s_i)$ formula with a quite small α helps to control the semantic deviation from the central topics of the text. It's experimentally proven that choises based on pure $P(s_i|s_1, s_l, ..., s_k)'$, especially the variant without left context $P(s_i|s_k)'$, are less central to the main text topics in later iterations. It happens because heads of self-attention mechanisms are less useful with smaller context given as an input.

V. Results

A. Automatic Evaluation

In "Dataset for Automatic Summarization of Russian News" [14] some popular and recent summarization methods were evaluated on the test part of the dataset:

- TextRank [15]
- LexRank [16]
- LSA(Latent Semantic Analysis) [17]
- SumaRuNNer [5]
- Pointer Generator [18]
- CopyNet [19]

There were also published results of mBART [20] fine-tuned on the Gazeta dataset in the paper. mBART based method is a current state-of-the-art in abstractive English text summarization on the CNN/Daily Mail dataset.

The mean sentence number of human summaries in the dataset is 3, so our method compressed the texts into 3 sentences.

Here are the results on test part of the dataset compared with other methods that were evaluation in Gazeta's work [14]:

Approach	ROUGE-1	ROUGE-2	ROUGE-L
Lead-1	27.6	12.9	20.2
Lead-2	30.6	13.7	25.6
Lead-3	31.0	13.4	26.3
Greedy Oracle	44.3	22.7	39.4
TextRank	21.4	6.3	16.4
LexRank	23.7	7.8	19.9
LSA	19.3	5.0	15.0
SumaRuNNer	31.6	13.7	27.1
Proposed method	35.6	14.2	32.4
CopyNet	21.4	6.3	16.4
PG small	23.7	7.8	19.9
PG words	19.3	5.0	15.0
PG big	29.6	12.8	24.6
PG small+coverage	30.2	12.9	26.0
Finetuned mBART	32.1	14.2	27.9

All these values are F-scores of ROUGE [7] metric.

Lead-N as in the "Dataset for Automatic Summarization of Russian News" [14] paper is just N first sentences taken.

Greedy Oracle is a method that uses human-written summaries, choosing sentences which are maximizing ROUGE-2 [7] score. It's useful for training classifiers sentence classfiers and as a target value for extractive summarization algorithms, defining a max reachable score.

Our algorithm is better than other ones at ROUGE-1 and ROUGE-L scores. [7] In ROUGE-2 [7] it's equal to SummaRuNNer and fine-tuned mBART methods. So, our method is the new state of the art on Gazeta, as there are no other results published on this dataset

B. *Human Evaluation*

Current summarization automatic evaluation methods do not score in any way the improvements in sentence order. We organized an anonymous opinion poll among students from other departments, most of whose are not related to natural language processing, machine learning and didn't have advanced algorithms courses.

We had involved 63 students to participate in the poll.

Each pole got an example from the test part of the used dataset. In both forms were given a headline, text, human-made summary, and the ones of our algorithm marked as "Summary 1" and "Summary 2".

Students answered the question "Which summary do you like more?" with the following possible options:

- Summary 1
- Summary 2
- Both great
- Both poor

Summary 1 and Summary 2 options are randomly mixed options Algorithm and Human.

The results are shown in the following table and are visualized(2).

Preferred	Count	Percentage
Algorithm	21	33.3%
Human	17	26.98%
Both great	16	25.4%
Both poor	9	14.29%

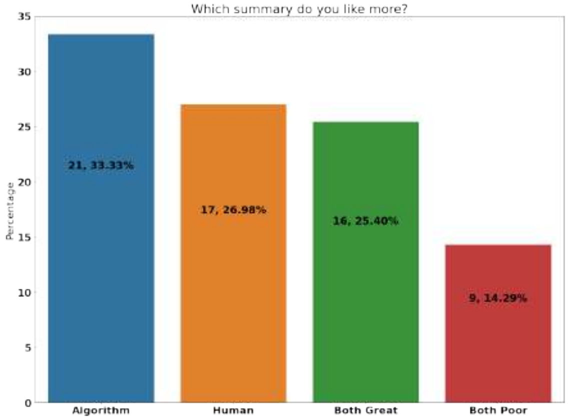

Figure 2. Student Poll Results

VI. Examples

A. *Gazeta Dataset*

Title: Почему мы остываем: температура человека снизилась за 200 лет

Original text: Температура человеческого тела снизилась за последние 200 лет, выяснили исследователи из Стэнфордского университета. О своем открытии они рассказали в статье в журнале eLife. Температура тела здорового человека колеблется в течение суток в пределах 35,5-37,2°C. Такая температура считается оптимальной для поддержания нормальной работы внутренних органов и протекания биохимических реакций, а также позволяет сдерживать грибковые инфекции. Хотя средней с XIX века считалась температура в 37°C, сегодня она считается повышенной и многим при таком показателе нездоровится. Раньше это списывалось на неточность измерений температуры в прошлом, но теперь оказалось, что люди с XIX века действительно «остыли». Чтобы выяснить, что на самом деле произошло, профессор Джули Парсоннет и ее команда объединили три набора данных. Первый из них охватывал почти 24 тыс. ветеранов Армии Союза времен Гражданской войны в США, температура которых измерялась в период между 1860 и 1940 годами. «Мне потребовалось много времени, чтобы найти данные XIX века, где была бы информация о температуре тела», — отмечает Парсоннет. Остальные наборы данных охватывали период с 1971 по 1975 год и с 2007 по 2017 год. В общей сложности команда проанализировала измерения температуры тела 677,5 тыс. человек. В среднем, температура тела людей снижалась на 0,03°C за десятилетие. У мужчин, родившихся в начале XIX века, температура тела была на 0,59°C выше, чем у мужчин сегодня. Данные по женщинам стали собираться несколько позже. Как выяснилось, температура их тела снизилась на 0,32°C с 1890-х годов. Средняя температура тела сегодня составляет около 36,6°C. Парсоннет предлагает два доказательства того, что дело именно в снижении температуры тела, а не в ненадежности старых термометров. Во-первых, тенденция к снижению температуры тела прослеживается и в поздних исследованиях, где использовались уже более точные термометры. «Изменения, которые происходили в 1860-1960-х годах, мы наблюдаем также в период с 1960-х годов по сегодняшний день, — говорит Парсоннет. — Я не думаю, что есть большая разница в термометрах между 1960-х годов и современными». Во-вторых, у пожилых людей температура тела была выше, чем измеренная в том же году у более молодых людей, причем разница была примерно одинаковой независимо от года. Также, сравнив несколько групп по возрастам, исследователи установили, что температура тела снижалась и у молодых, и у пожилых людей в одинаковой степени. Если бы проблема была в термометрах, то выявить такие точные различия было бы сложно, считает Парсоннет. В 1800-е годы люди страдали от малярии, туберкулеза, дизентерии, болезней полости рта и многих других продолжительных или хронических заболеваний, отмечает она. Сейчас с большинством болезней удалось справиться, что и могло повлиять на снижение температуры тела: организму не нужно больше постоянно бороться с инфекцией. «На мой взгляд, дело в том, что, с микробиологической точки зрения, мы очень отличаемся от людей прошлого, — говорит Парсоннет. — У современных людей меньше инфекций, благодаря вакцинам и антибиотикам, поэтому наша иммунная система не так активна, а ткани организма менее воспалены. Если это так, то температура тела должна была бы понизиться и в других странах, где здоровье людей улучшилось». Вряд ли в ближайшее время температура тела человека заметно понизится, считает Парсоннет. «До нуля мы не остынем, — шутит она. — Есть какой-то предел, просто я пока не знаю, где он». Как изменилась температура тела человека в большинстве других стран, еще предстоит узнать. Однако, например, результаты британского исследования 2017 года повторяют выводы Парсоннет. Измерив температуру тела 35,5 тыс. здоровых британцев, исследователи выяснили, что в среднем она составляет 36,6°C. Однако пожилые британцы, в отличие от американцев, были «прохладнее» молодых. Также температура зависела и от расовой принадлежности: темнокожие респонденты оказались «горячее».

Human summary: Температура человеческого тела снизилась за последние 200 лет, выяснили американские исследователи. В 1800-е годы она составляла 37°C, причиной чему, вероятно, были продолжительные и хронические болезни. Сегодня люди более здоровы — и более «прохладны».

Machine summary: Температура человеческого тела снизилась за последние 200 лет, выяснили исследователи из Стэнфордского университета. У мужчин, родившихся в начале XIX века, температура тела была на 0,59°C выше, чем у мужчин сегодня. Раньше это списывалось на неточность измерений температуры в прошлом, но теперь оказалось, что люди с XIX века действительно «остыли».

B. Artice Introduction

This is a summary of this article's introduction section with multilingual BERT and Sentence-BERT models:

Text summarization is the task of creating a shorter version of a document that captures essential information. Extractive methods form a summary as a combination of the original text's chunks. These approaches are usually reduced to classifying sentences of the initial document. We introduce a new extractive summarization approach for Russian texts, which leverages pre-trained BERT(Bidirectional Encoder Representations from Transformers) [1] and Sentence-BERT [2] models. The proposed method forms a summary that has human-natural storytelling order of sentences.

VII. Conclusion

Most modern extractive text summarization techniques use pre-trained language models as a decent linguistic feature aware foundation, that can be fine-tuned with additional layers or used for embedding and next sentence prediction probability estimation.

There are many works about English and other popular languages that leverage modern methods which are significantly better than older approaches, but, as for the Russian there are barely any works that use attention-based models, which have been proved to work better with longer sequences than RNNs, whereas text summarization is all about sequences processing.

In this paper, we introduced a new approach to extractive text summarization, which uses pre-trained language models and can easily be used for other languages. The new algorithm orders sentences in a human-natural way, making summaries easier to read. The proposed summarization method achieved higher ROUGE-1 and ROUGE-L [7] scores on the Gazeta dataset and was preferred more in human evaluation.

As for the future, we plan to go on with exploring pre-trained language models' possibilities in semantically driven tasks, especially abstractive summarization.

Acknowledgment

The work was carried out within the framework of the research “Mathematical and software support of advanced systems for processing symbolic information with elements of artificial intelligence” which is led by the department of algorithmic languages on the faculty of computational mathematics and cybernetics of the Lomonosov Moscow State University.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019, arXiv preprint arXiv:1810.04805. Available at https://arxiv.org/pdf/1810.04805.pdf (accessed 2020, Dec)

[2] Nils Reimers, Iryna Gurevych Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 2019, arXiv preprint arXiv:1908.10084. Available at https://arxiv.org/pdf/1908.10084.pdf (accessed 2020, Dec)

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[4] Yuri Kuratov, Mikhail Arkhipov Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019, arXiv preprint arXiv:1905.07213. Available at https://arxiv.org/pdf/1905.07213.pdf (accessed 2021, Jan)

[5] Ramesh Nallapati, Feifei Zhai, Bowen Zhou SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. 2016, arXiv preprint arXiv:1611.04230. Available at https://arxiv.org/pdf/1611.04230.pdf (accessed 2020, Nov)

[6] Shashi Narayan, Shay B. Cohen, Mirella Lapata Ranking Sentences for Extractive Summarization with Reinforcement Learning. 2018, arXiv preprint arXiv:1802.08636. Available at https://arxiv.org/pdf/1802.08636.pdf (accessed 2021, Mar)

[7] Chin-Yew Lin ROUGE: a Package for Automatic Evaluation of Summaries. Association for Computational Linguistics, 2004, vol. Text Summarization Branches Out, pp. 74–81

[8] Yang Liu, Ivan Titov, Mirella Lapata Single Document Summarization as Tree Induction. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, vol. 1 (Long and Short Papers), pp. 1745-1755

[9] Xingxing Zhang, Furu Wei, Ming Zhou HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. 2019, arXiv preprint arXiv:1905.06566. Available at https://arxiv.org/pdf/1905.06566.pdf (accessed 2021, Dec)

[10] Yang Liu, Mirella Lapata Text Summarization with Pretrained Encoders. 2019, arXiv preprint arXiv:1908.08345. Available at https://arxiv.org/pdf/1908.08345.pdf (accessed 2021, Apr)

[11] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, Xuanjing Huang Extractive Summarization as Text Matching. 2020, arXiv preprint arXiv:2004.08795. Available at https://arxiv.org/pdf/2004.08795.pdf (accessed 2021, Apr)

[12] Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D. Manning A large annotated corpus for learning natural language inference. 2015, arXiv preprint arXiv:1508.05326. Available at https://arxiv.org/pdf/1508.05326.pdf (accessed 2021, Mar)

[13] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, Veselin Stoyanov XNLI: Evaluating Cross-lingual Sentence Representations. 2018, arXiv preprint arXiv:1809.05053. Available at https://arxiv.org/pdf/1809.05053.pdf (accessed 2021, Mar)

[14] Ilya Gusev Dataset for Automatic Summarization of Russian News. 2020, arXiv preprint arXiv:2006.11063. Available at https://arxiv.org/pdf/2006.11063.pdf (accessed 2021, Feb)

[15] Rada Mihalcea, Paul Tarau TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp. 404–411

[16] Gunes Erkan, Dragomir R. Radev LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 2004, vol. 22, pp. 457-479

[17] P. W. Foltz Latent semantic analysis for text-based research. Behavior Research Methods, Instruments, & Computers, 1996, vol. 28, pp. 197–202

[18] Abigail See, Peter J. Liu, Christopher D. Manning Get To The Point: Summarization with Pointer-Generator Networks. 2017, arXiv preprint arXiv:1704.04368. Available at https://arxiv.org/pdf/1704.04368.pdf (accessed 2021, Feb)

[19] Jiatao Gu, Zhengdong Lu, Hang Li, Victor O.K. Li Incorporating Copying Mechanism in Sequence-to-Sequence Learning. 2016, arXiv preprint arXiv:1603.06393. Available at https://arxiv.org/pdf/1603.06393.pdf (accessed 2021, Feb)

[20] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer Multilingual Denoising Pre-training for Neural Machine Translation. 2020, arXiv preprint arXiv:2001.08210. Available at https://arxiv.org/pdf/2001.08210.pdf (accessed 2021, May)

Извлекающее автореферирование русскоязычных текстов с применением предобученных языковых моделей

Полякова И.Н., Погорельцев С.А.

Предобучееные языковые модели на основе архитектуры Transformer установили новые рекорды на многих задачах обработки естественных языков при их дообучении или использовании для получения контекстуализированных семантических векторов. Модели с механизмом многоголового внутреннего внимания стали лучшими на задаче автореферирования англоязычных текстов, однако возможности применения подобных подходов для русского языка слабо изучены. Мы представляем новое решение задачи автореферирования текстов на русском языке, которое достигает лучших результатов по нескольким метрикам относительно других моделей, таких как SummaRuNNer и дообученный mBART для генерирующего автореферирования на датасете Gazeta. Результаты работы предлагаемого алгоритма является более предпочтительным вариантом в опросе среди студентов.

Received 31.05.2021