Using Generative Pretrained Transformer-3 Models for Russian News Clustering and Title Generation tasks

Maria Tikhonova[1,2] **Dina Pisarevskaya**[1]
Tatiana Shavrina[1,2,3] **Oleh Shliazhko**[1]
[1]SberDevices, Sberbank, Moscow, Russia
[2]National Research University Higher School of Economics, Moscow, Russia
[3]ANO «AI Research Institute», Moscow, Russia

Abstract

The paper presents a methodology for news clustering and news headline generation based on the zero-shot approach and minimal tuning of the RuGPT-3 architecture (Generative Pretrained Transformer 3 for Russian). The solution is presented in a competition for news clustering, headline selection and generation.
The following approaches are described: 1) zero-shot unsupervised classification based on pairwise news perplexity: the method requires no training or model fine-tuning and yields 0.7 F1-measure.
2) fine-tuning: news headlines generation with the best result 0.292 ROUGE and 0.596 BLEU.

Keywords: text clustering, text generation, evaluation track, ruGPT-3, generative pretrained transformer
DOI: 10.28995/2075-7182-2021-20-1214-1223

Модели Generative Pretrained Transformer-3 в Задачах Кластеризации Новостей и Генерации Заголовков

Мария Тихонова[1,2] Дина Писаревская[1]
Татьяна Шаврина[1,2,3] Олег Шляжко[1]

[1]SberDevices, Сбербанк, Москва, Россия
[2]НИУ «Высшая Школа Экономики», Москва, Россия
[3]АНО «Институт Искусственного Интеллекта», Москва, Россия

Аннотация

В работе представлена методика работы с кластеризацией новостных текстов и генерации заголовков на основе подхода zero-shot и минимального дообучения архитектуры RuGPT-3 (Generative Pretrained Transformer 3 for Russian). Решение представлено в рамках соревнования по кластеризации, выбору и генерации заголовков для новостей.
В работе рассмотрены следующие подходы: 1) zero-shot классификация без учителя основанная на перплексии пары новостей: алгоритм не требует обучения или дополнительного дообучения языковой модели и позволяет достичь f1-меры 0,7.
2) fine-tuning: генерация новостных заголовков с лучшим результатом 0.292 ROUGE и 0.596 BLEU.

Ключевые слова: кластеризация текстов, генерация текстов, ruGPT-3, generative pretrained transformer

1 Introduction

News articles are one of the most frequently used sources in the Natural Language Processing (NLP) evaluations tracks, as they are widely accessible, voluminous and have many practical applications originating from that kind of text data [1 - 3]. Among them are such areas as text summarization, text classification, thematic modelling, and text clustering. The 2021 Dialog Evaluation Shared Task [4]

brings all these tasks together under 3 subtracks on the Russian news data. News articles are commonly used for the quality assessment of NLP systems due to the fact that they tend to be characterized by specific complexity. Namely, news texts are abundant with factual information, mentions of named entities, basic facts and details, comments and opinions over the events, etc.

In the paper, we present the simplistic approach for news data clustering and headline generation based on the usage of pretrained language models with few-shot methods and minimal fine-tuning. For news clustering, we tested the zero-shot method based on pairwise news perplexity which requires no model fine-tuning at all and only a small amount of training examples for threshold selection. At the same time, such a simplistic method yields a reasonable F1 score of 0.7. The headline generation method presented provides 0.596 BLEU metric and 0.292 ROUGE metric.

This paper is structured as follows: in section 2 we present already existing research works related to the discussed topics; section 3 provides the general overview of the competition; section 4 is devoted to the news clustering track solution; section 5 describes our solution of title generation task; and section 6 concludes the paper.

2 Previous Work

2.1 News Clustering and Headline Generation

News clustering can be topic-based or story-based. Topic-based clusters may combine multiple stories about close events and with similar keywords. In recent years, most studies were focused on topic-based clustering. For instance, within the topic-based approach, [5] fine-tune pretrained multilingual models (such as BERT and XLM-R) as encodings for the task of Lithuanian topic-based news clustering.

Within the story-based approach, [6] propose a two-steps streaming system that first extracts keywords to create topics, and then combine local topics clusters into stories by comparing their keywords distribution; they use TF-IDF based method to compare articles. [7] apply similarity metrics and ranking for clustering an incoming stream of multilingual documents into monolingual and cross-lingual story clusters. [8] use fine-tuned BERT and ALBERT models to classify if sentences from two texts refer to the same event or not, getting initial scores. In the next step, the pair scores are recalculated by considering the relation of sentences in a pair concerning other sentences; final scores between these sentences are used to construct the clusters. [9] propose fine-tuned multilingual embedding using SBERT to create cross-lingual stories. [10] explore for classification of events, presented in short texts, the performance of TF-IDF-weighted character n-gram SVM-based model with SVMs trained on different pretrained word embeddings (GLOVE, BERT, FASTTEXT) as features.

For Russian, different story-based clustering techniques were investigated in [11]: results for text embeddings based on TF-IDF, pretrained language model (BERT), and multilingual embeddings pretrained on parallel corpora (LASER) are compared. Agglomerative clustering and DBSCAN are used. Additional approaches, such as clustering based on named entities as keywords, are also suggested. To the best of our knowledge, news clustering methods that apply GPT-3 perplexity were not used in the research for Russian before.

It is worth mentioning the problem of title composition/generation - for the Russian language, this problem already has a well-developed background based on the HeadlineGenerationEval-2019 competition[1] [3]. Such methods as machine translation, attentional transformer and Pointer-Generated networks were applied to the task, while the baseline of the first sentence choice as the headline of the news text was extremely powerful.

2.2 Few-Shot Methods for Russian

Along with the growing number of parameters of large language models and transformers, it is becoming more and more computationally complex to train the new tasks. Various techniques to reduce training costs include freezing models' layers, compressing and distilling models, and methods for manipulating model inputs during the inference: few-shot learning (several examples are combined into one input "prompt" and the target example is added to it), one-shot learning (one example in a prompt is combined

[1] http://www.dialog-21.ru/evaluation/2019/disambiguation/generation/

with a target example), and zero-shot learning (no examples passed at all). See an illustration of the prompt manipulation in Figure 1 taken from the GPT-3 work [11] that heavily popularized these methods.

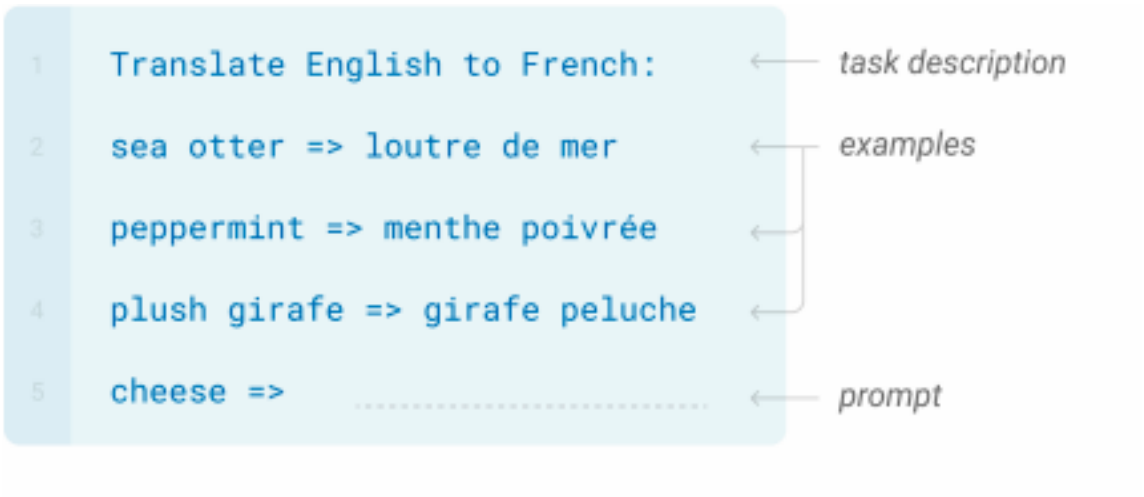

Figure 1: Prompting the pretrained model to perform machine translation

In the work [12] the pretrained language models were achieving outstanding results, including 71.8% overall score on SuperGLUE (GPT-3 175B parameters) with few-shot learning, 68.9% overall score with one-shot learning and 58.2% overall score with zero-shot learning (SOTA being 90.4%).

For the Russian language, such a method of prompt application became available with the advent of the Russian-language generative models ruGPT-2[2] and ruGPT-3[3]. For the first time, the SOTA result was obtained based on zero-shot learning on the Russian SuperGLUE benchmark [13]: model perplexity was used to weight the answer options for tasks, received on the text of the task joint with the answer option. The variant with the least model perplexity was chosen as the model's response, resulting in a 53.5% overall score, with the random choice being 38.5% score.[4]

This approach has the potential in its simplicity, although it has unavoidable drawbacks: for the best result, the developer needs to investigate the prompts and choose the optimal one. As shown in [14], the success of the few-shot learning approach depends significantly on the number of examples in the prompt, their class balance and order of the examples – these factors can lead to the quality deterioration or improvement up to 30%.

Fine-tuning for this purpose also remains appropriate, especially in tasks requiring work in a specialized domain, subject, and strict data formatting.

3 Dialog Evaluation 2021 Track

The main track criteria state that a pair of news texts refer to the same event when they match:

1. *time of the event*;
2. *numbers:* such as the stock price of a company or the number of victims;
3. *locations:* for example, the location of an event or the location of an accident.

A couple of news items are not related to the same event if they either have *inconsistent facts* (the time or place of the event, significantly distinguishing the number of victims, etc.) or there is *a description of an event* in one of the news and *a commentary on this event* by some person in another.

The training data, including text clusters, titles and news texts are originally published within the Telegram contest[5]. Below is a training example (news texts were shortened for demonstration purpose):

[2] https://github.com/mgrankin/ru_transformers

[3] https://github.com/sberbank-ai/ru-gpts

[4] At the time of this writing, a new result has been achieved with the help of tuning English-based and multilingual models, the new SOTA being 67.9%

[5] https://contest.com/docs/data_clustering2/ru

{“first_url”: “ https://news-r.ru/news/stavropol_krai/479592”,
“second_url”: “https://stavropolye.tv/news/134284”,
“first_title”: “Мужчина заработал почти 3 млн рублей на незаконном АЗС на Ставрополье”,
“first_text”: “Житель Курского района открыл автозаправку, хотя лицензии на это не имел Ставрополь, 25 мая. Житель Курского района Ставропольского края незаконно открыл автозаправочную станцию и заправлял автомобили сжиженным углеводородным газом. [...]”,
“first_timestamp”: “2020-05-25 14:33:00”,
“first_host”: ’news-r.ru’,
“second_title”: ’На Ставрополье задержан владелец незаконной автозаправки’,
“second_text”: “Уголовное дело по факту незаконного предпринимательства расследуется на Ставрополье. Об этом сообщили в понедельник в краевом управлении МВД РФ. [...]”,
“second_timestamp”: “2020-05-25 13:36:00”,
“second_host”: “stavropolye.tv”}

The training data of both of the tracks originate from 620 websites, including the major news sources like tass.ru, lenta.ru, mk.ru, rosbalt.ru, etc. and also smaller and more thematically homogeneous sources like champions.football.ua, dota2.ru, brodude.ru, and even galleryporn.d3.ru.

4 News Clustering

4.1 Data and metrics

News texts are taken from the Telegram Data Clustering Contest. Train set contains 14838 URL pairs for news texts. All texts were written on 25 May 2020. Pairs are annotated by organizers: each pair has a label if both texts refer to the same cluster or not. It was also allowed to use additional data - news from other 11 days provided by the organizers - for models pre-training or fine-tuning.

Public and private test sets contain 8493 and 8480 text pairs respectively. Each of them corresponds to one day: the public test set includes URL pairs for news from 27 May 2020, and the private test set contains URL pairs for news from 29 May 2020. F-score for positive examples is used as metrics.

Titles, texts and some metadata information (such as publication timestamp) were extracted for the URLs using the library provided by the task organizers [6].

4.2 Methods and results

For this task, we used approaches based on RuGPT-3 and intentionally aimed to check how RuGPT-3 can perform clustering in an unsupervised setting, that is without fine-tuning on the contest data. RuGPT-3 models were, in general, trained on a huge dataset of Russian, including news.

For every news story from the news clustering task, such details as title, text, URL, DateTime information are provided. For this task we tested three different approaches:

1. zero-shot perplexity-based approach: unsupervised news classification based on model perplexity;
2. supervised classification: standard supervised approaches for pair classification;
3. combined approach: supervised classification using perplexity features.

Below each approach is described in detail.

Unsupervised zero-shot perplexity-based approach First group of methods uses the concept of zero-shot approach introduced in [12], which proved to be very effective for GPT-3 models. In zero-shot setting the prediction is generated without any supervision based solely on a natural language prompt constructed from the test example. For the competition task we used a modification of the original method in which the decision is made based on the model perplexity and the pre-defined threshold. Namely, for each test sample we first calculate the perplexity of the model on the prompt constructed by uniting the two news:

$$PPL(t) = exp(-\frac{1}{|t|}\sum_{i=0}^{|t|} log_{p_\theta}(x_i|x_{<i})) \quad (1)$$

[6] https://github.com/IlyaGusev/purano

Combination	Threshold	F1 score
1 title + 2 title	35	**0.7**
1 title + “ , ”+ 2 title	56	0.672
1 title + 2 text	22	0.657
2 title + 1 text	20	0.651

Table 1: Zero-shot perplexity results

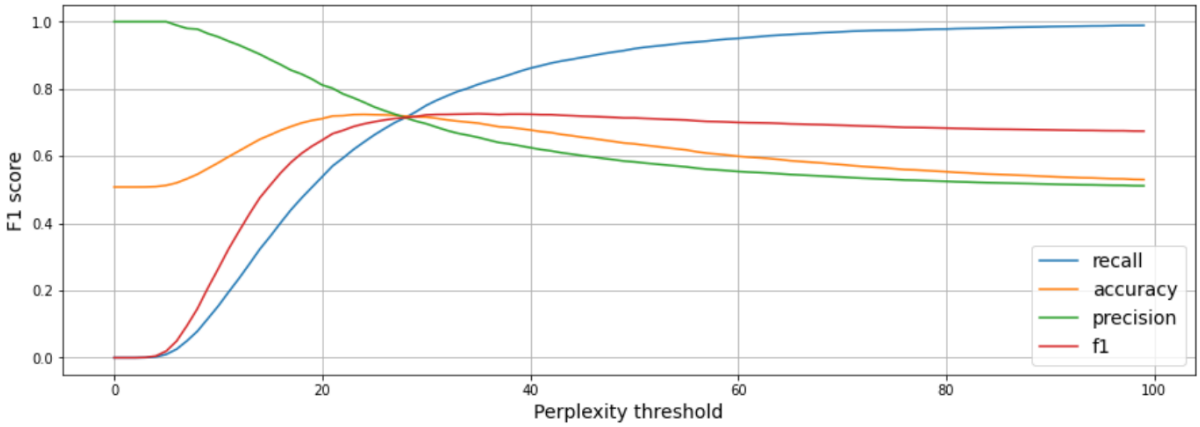

Figure 2: Metrics in dependence on the threshold on train set. Perplexity is calculated for 2 joint titles.

where t is an input text, $|t|$ is the length of the text in tokens, and $log_{p_\theta}(x_i|x_{<i})$ is the log-likelihood of the i-th token in t conditioned on the preceding ones.

The final classification is then performed based on the threshold selected on the subset of training data of 100 examples (the small subset size is chosen in order to maximally preserve the unsupervised setting of the experiment). The idea of the method is based on the hypothesis that for news on different events their joint perplexity is higher than for news on the same event.

All the experiments were carried out using RuGPT-3-XL[7]. We set the perplexity threshold equal to 35.

The advantage of the proposed method is that it does not require any model training, fine-tuning, or cauterization, which can be computationally difficult and time-consuming. Basically, such a method allows performing classification using any pretrained language model. And we only require a reasonable number of examples for threshold selection, not a large training set. Thus, such an approach can be regarded as an unsupervised zero-shot classification method.

For perplexity calculation we tested 4 different news aggregations:

1. title 1 + title 2;
2. title 1 + text 2;
3. title 2 + text 1;
4. title 1 + “ , ”+ title 2.

Results on the public test set and perplexity thresholds are presented in Table 1. The best result (*F1 for positive examples = 0.7*) was obtained for the perplexity of two titles combined together. The metric curves for different perplexity thresholds on the train set for this aggregation are presented in Fig. 2.

Supervised classification The next group of methods includes standard supervised classification methods based on word embeddings. We embedded titles joint with texts. Thus, for each news pair, we obtained 2 word vectors that were concatenated and used as features in CatBoost[8]. We tested Fast-

[7] https://huggingface.co/sberbank-ai/rugpt3xl

[8] https://catboost.ai/

Model	Public test set F1 score	Private test set F1 score
GPT-3 emdeddings + Perplexity	0.843	0.841
GPT-3 emdeddings	0.867	0.850
FastText emdeddings	0.861	0.856

Table 2: Supervised and combined approach results

Text and RuGPT-3 embeddings (see Table 2).

We got RuGPT-3-L[9] pooling embeddings for each text (taking all sentences of a text).

Combined approach Finally, we combined 4 perplexity results from the first approach with embedding features and used CatBoost (10000 iterations) on top. This yielded a better result than perplexity alone, achieving the score of 0.843 on the public test set and 0.841 on the private test set with RuGPT-3 embeddings (see Table 2).

We also experimented with RuGPT-3 embeddings from previous layers, as well as with RuBERT embeddings from various layers [10], in order to find out which embedding features might be better combined with perplexity features, but did not get better results. These experiments should be conducted further.

We also experimented with adding different thresholds for the intersection of dates and time in a text pair. As all text were from the same day, it did not improve the results, as well as in [11]. Adding various thresholds for named entities intersections, namely persons and locations, extracted with Natasha library [11], also did not help to improve the results.

4.3 Error analysis and next steps

For the model that combines perplexity results with RuGPT-3-L embedding features, we performed the error analysis.

Classification errors for news texts pairs from the same cluster may be caused by different text lengths and, therefore, different keywords. Another reason is that one of the texts in a pair may contain more detailed information, or provide some background about the context of the news story. For instance, both texts are about the possible economic impact of the coronavirus pandemic in Russia, but the latter one contains a more detailed forecast: Title 1 “Handelsblatt (Германия): крупнейший кризис для Путина — России грозит глубокая депрессия” [12] and Title 2 “России предрекли глубокий экономический кризис” [13] Both texts are devoted to the same English football club that is not popular among football fans, but the first one provides a more detailed statistics about other English teams, and the second one contains background information about the coronavirus pandemic impact on football championships in England: Title 1 ““Лидс»” – самая ненавистная команда Англии по рейтингу English Football Statistician” [14] and Title 2 “Назван самый ненавистный клуб Англии”[15] .

Misclassification examples of texts from different clusters, that were labelled as texts from one cluster, can be also mentioned. Such errors might occur if topics and keywords of both texts in a pair are similar, despite that they describe different events. For example, both texts are about the coronavirus pandemic trends in Russia, but the first one is focused on a definite region in Russia, and the second one is about the situation in the whole country: Title 1 “В Ростовской области под медицинским наблюдением из-за коронавируса находятся почти 9 тысяч человек” [16] and Title 2 “В России

[9]https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2

[10]https://huggingface.co/DeepPavlov/rubert-base-cased

[11]https://github.com/natasha/natasha

[12]https://inosmi.ru/politic/20200527/247504317.html

[13]https://news.rambler.ru/community/44248748-handelsblatt-germaniya-krupneyshiy-krizis-dlya-putina-rossii-grozit-glubokaya-depressiya/

[14]https://bombardir.ru/news/604644-Lids-samaya-nenavistnaya-komanda-Anglii-po-reytingu-English-Football-Statistician

[15]https://www.euro-football.ru/article/31/1004367954_nazvan_samyiy_nenavistnyiy_klub_anglii

[16]https://donday.ru/v-rostovskoj-oblasti-pod-medicinskim-nabljudeniem-iz-za-koronavirusa-nahodjatsja-pochti-devjat-tysjach-chelovek.html

под медицинским наблюдением из-за коронавируса находятся почти 300 тысяч человек" [17]. Both texts are about football teams, contain interviews with their trainers and similar lexics about perspectives of a team, but teams and trainers are different: Title 1 "Алексей МИХАЙЛИЧЕНКО: "Возможно, мы будем делить "Олимпийский" с "Шахтером", а, может быть, будем проводить наши матчи на "Динамо"" [18] and Title 2 "Маркевич: Когда Днепр на жилах добрался до 1/16 Лиги Европы, я понял, что эта команда может пошуметь в Европе" [19].

As to the next steps, in order to improve the results, RuGPT-3 embedding features could be obtained not for a whole text, but a title and N first sentences. Further experiments might be conducted, aiming to reveal better aggregation of news aggregation (for example, title 1 + title 2 + text 1 + text 2) in zero-shot perplexity based classification as well as a better combination of the zero-shot method with various embeddings (such as LaBSE and ELMO) and different supervised classification techniques. Experiments for several document similarity metrics with different thresholds might be performed as well.

5 Title Generation

Headline generation track is aimed to evaluate modern ways of automatically generating headlines from the text. Taking into account the experience of past competitions, this task can be considered similar to the summation or isolation of the main meaning of the text - the organizers did not set any restrictions on the generative methods of solution. Therefore, as a baseline solution, a method of a random choice of the first or second sentence of the news text was proposed.

The rules of the subtrack also included the following items:

1. participants should generate only one title per cluster and then the script compares the generated title with every original title from a cluster. The best score among these comparisons is used.
2. the texts for testing the models are provided in advance and are real news articles with titles can be found online;
3. to train the systems, it is possible to use any data, including the training data of the competition and additional existing news corpora, open news sources, etc., however, the participant is responsible not to search for real headings of test texts on the Internet, and also to exclude a test leak when training on external data.

5.1 Data and Metrics

The main metric that determines the success of a headline is ROUGE f-measure from python package[20], including rouge-1, rouge-2 and rouge-l between the hypothesis and the real headline.

$$(ROUGE-1 + ROUGE-2 + ROUGE-L)/3 \tag{2}$$

BLEU score from nltk.translate package is also considered as part of the evaluation.

Evaluation script[21] also includes text preprocessing pipeline:

- tokenization with razdel python library[22];
- deletion of the excessive spaces;
- transfering the text into lower case.

The pipeline was preserved during the data preparation process. The data being the same as in the other evaluation subtracks has resulted in a format adapted to title generation task: *special beginning token (BOS) + ': ' + the article text itself + ': ' + the title + special ending token (EOS).*

Training data used contains 29676 news articles with titles.

[17] https://russian.rt.com/russia/news/750040-koronavirus-rossiya-nablyudenie

[18] http://dynamomania.com/news/531289-aleksej-mihajlichenko-vozmozhno-my-budem-delit-olimpijskij-s-shahterom-a-mozhet-byt-budem-provodit-nashi-matchi-na-dinamo

[19] https://isport.ua/football/1849289-markevich-kogda-dnepr-na-zhilakh-dobralsya-do-1-16-ligi-evropy-ya-ponyal-chto-eta-komanda-mozhet-poshumet-v-evrope

[20] https://pypi.org/project/rouge-metric/

[21] https://github.com/IlyaGusev/purano/blob/master/purano/util/hg_evaluate.py

[22] https://github.com/natasha/razdel

5.2 Methods and Results

As the main method of the subtrack, ruGPT-3 Large with 760 millions parameters was fine-tuned on the training data. During 5 epochs of the training, the 2.68 loss was obtained.

The inference mode of interaction with the fine-tuned model included regex-based splitting of the generated continuation of the input article text + special token added after the text. This basic method has provided the initial overall 0.2687 ROUGE score with top-k sampling and 0.2686 ROUGE score with greedy generation (beam search, no sampling).

Additional +3% to the score has been added with a simple trick: sentence tokenization of the article texts was performed with rusenttokenize package[23], and then only the first 7 sentences of the original article text were passed as an input to the tuned model. This heuristic was dictated by the fact that news texts for the most part contain all the most basic information at the very beginning, and then the story is filled with additional details, comments and, at the very end, a call to action and links to other materials. This principle of writing a news article is also known to the editors as the principle of the inverted pyramid: a news article, which is written according to the scheme of an inverted pyramid, contains answers to six questions: "five Ws and one H":

1. Who is the main character of this news?
2. What exactly happened?
3. When did it happen?
4. Where did it happen?
5. Why did this happen?
6. How did it happen that this happened?

According to our observations, this principle is maintained in the training data of the competition and can be followed in the test data - although some examples are clearly difficult to title in essence, since they include information such as an indication of the source, an indication of a comment on an event in the title, and not the event itself. The example below attaches headings to the same text, generated by the network, and by successively decreasing the number of first sentences of the text of a news article[24]:

Gold title: Володин отреагировал на блокировку Facebook публикаций российских СМИ
10+ sentences: Facebook заблокировал статью о расследовании в Воронеже, в которой утверждалось, что в РФ есть больные COVID-19
9 sentences: Facebook заблокировал статью о расследовании в Воронеже, в которой утверждалось, что в РФ есть больные COVID-19
8 sentences: Facebook заблокировал статью о расследовании дела Вороновского
7 sentences: Володин заявил о нарушении Facebook российскими СМИ
6 sentences: Facebook заблокировал статью о российских военных в Сирии
5 sentences: Facebook удалил десятки публикаций российских СМИ
4 sentences: Facebook заблокировал статью о врачах, заразившихся коронавирусом
3 sentences: Facebook удалил десятки публикаций российских СМИ
2 sentences: Facebook удалил десятки публикаций о коронавирусе
1 sentences: Facebook удалил десятки публикаций о коронавирусе

This example clearly illustrates the excess and lack of information for the generative model that makes up the heading: if there are 10+ sentences, the title matches the actual content of the text as much as possible, but does not correspond to the generalization in the gold heading; with 7-8 first sentences, the generalization is optimal; and with the first 5 sentences and less the quality of the heading gradually degrades, becoming too general or drifting to the side, getting lost on accidentally touched upon or statistically probable, but unmentioned topics.

The final method brought us the following metrics presented in Table 3.

[23] https://pypi.org/project/rusenttokenize/

[24] Source: https://russian.rt.com/russia/news/840380-volodin-facebook-smi

2*#	2*User	2*ROUGE	2*BLEU	2*ROUGE-1	2*ROUGE-2	2*ROUGE-L
1	LOLKEK	0.387 (1)	0.695 (1)	0.463 (1)	0.264 (1)	0.433 (1)
2	Our Team	0.292 (2)	0.596 (2)	0.365 (2)	0.176 (2)	0.335 (2)

Table 3: Results of the Title Generation Track

6 Conclusion

We present the results of our participation in the DE2021: Russian News Clustering and Headline Generation shared task. The implemented methods in both subtracks are based on the Generative Pretrained Transformer-3 architecture for Russian, requiring minimum data and computing power.

We show that zero-shot perplexity based classification without any additional model fine-tuning yields a reasonable F1 score of 0.7. The advantage of this method that it does not require any model training or clusterisation. Moreover, this zero-shot method can be used in combination with other standard supervised classification techniques like word embedding, for example. Fine-tuned ruGPT-3 Large model is used for headline generation task based on the part of the training data: model tuning with special tokens helps to quickly train the language model to produce headline correctly based on the input text.

All the methods presented in the paper are available open-source. We hope that our developments will be useful to the community since all the presented prototypes are easily portable to new domains and tasks: a zero-shot based on perplexity is able to show itself in various classification problems without a training sample, only with the selection of seeds; model fine-tuning for thematic generation tasks and the generation of a certain type of metainformation is also much less demanding in terms of computational costs than training models from scratch, and also requires a small training sample for a quick start.

References

[1] Ivanin V. A., Artemova E. L. et al. Rurebus-2020 shared task: Russian relation extraction for business // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2020”. - Moscow, Russia, 2020. - P. 416-431.

[2] Starostin A. S., Bocharov V. V. 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”. - Moscow, Russia, 2016. - P. 702-720.

[3] Malykh V. A., Kalaidin P. S. Headline Generation Shared Task on Dialogue’2019 // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2019): Issue 18: Supplementary volume. - Moscow, Russia, 2019. - P. 93-100.

[4] Gusev Ilya, Smurov Ivan. Russian News Clustering and Headline Selection Shared Task // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”. - 2021.

[5] Stankevičius Lukas, Lukoševičius Mantas. Testing pre-trained Transformer models for Lithuanian news clustering. - 2020. - Vol. arXiv:2004.03461. Access mode: https://arxiv.org/abs/2004.03461.

[6] Laban Philippe, Hearst Marti. newsLens: building and visualizing long-ranging news stories // Proceedings of the Events and Stories in the News Workshop. - Vancouver, Canada, 2017. - P. 1-9.

[7] Miranda Sebastiao, Artūrs Znotiņš et al. Multilingual clustering of streaming news // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. - Brussels, Belgium, 2018. - P. 4535–4544.

[8] Örs Faik Kerem, Yeniterzi Süveyda, Yeniterzi Reyyan. Event Clustering within News Articles // Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020. AESPEN. Language Resources and Evaluation Conference (LREC 2020). - Marseille, France, 2020. - P. 63-68.

[9] Linger Mathis, Hajaiej Mhamed. Batch Clustering for Multilingual News Streaming. - 2020. - Vol. arXiv::2004.08123. Access mode: https://arxiv.org/abs/2004.08123.

[10] Piskorski Jakub, Jacquet Guillaume. TF-IDF Character N-grams versus Word Embedding-based Models for Fine-grained Event Classification: A Preliminary Study // Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020. AESPEN. Language Resources and Evaluation Conference (LREC 2020). - Marseille, France, 2020. - P. 26-34.

[11] Воропаев П. М., Сопильняк О. А. Сравнение методов сюжетной кластеризации новостей // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2020”. Дополнительные материалы конференции: Студенческие статьи. - Moscow, Russia, 2020.

[12] Brown Tom B., Mannet Benjamin et al. Language models are few-shot learners. - 2020. - Vol. arXiv:2005.14165. Access mode: https://arxiv.org/abs/2005.14165.

[13] Shavrina Tatiana, Fenogenova Alena et al. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). - Online, 2020. - P. 4717–4726.

[14] Zhao Tony Z., Wallace Eric et al. Calibrate Before Use: Improving Few-Shot Performance of Language Models. - 2021. - Vol. arXiv:2102.09690. Access mode: https://arxiv.org/abs/2102.09690.