https://doi.org/10.17323/jle.2024.22244

Wrong Answers Only: Distractor Generation for Russian Reading Comprehension Questions Using a Translated Dataset

Nikita Login

HSE University, Moscow, Russia

ABSTRACT

Background: Reading comprehension questions play an important role in language learning. Multiple-choice questions are a convenient form of reading comprehension assessment as they can be easily graded automatically. The availability of large reading comprehension datasets makes it possible to also automatically produce these items, reducing the cost of development of test question banks, by fine-tuning language models on them. While English reading comprehension datasets are common, this is not true for other languages, including Russian. A subtask of distractor generation poses a difficulty, as it requires producing multiple incorrect items.

Purpose: The purpose of this work is to develop an efficient distractor generation solution for Russian exam-style reading comprehension questions and to discover whether a translated English-language distractor dataset can offer a possibility for such solution.

Method: In this paper we fine-tuned two pre-trained Russian large language models, RuT5 and RuGPT3 (Zmitrovich et al, 2024), on distractor generation task for two classes of summarizing questions retrieved from a large multiple-choice question dataset, that was automatically translated from English to Russian. The first class consisted of questions on selection of the best title for the given passage, while the second class included questions on true/false statement selection. The models were assessed automatically on test and development subsets, and true statement distractor models were additionally evaluated on an independent set of questions from Russian state exam USE.

Results: It was observed that the models surpassed the non-fine-tuned baseline, the performance of RuT5 model was better than that of RuGPT3, and that the models handled true statement selection questions much better than title questions. On USE data models fine-tuned on translated dataset have shown better quality than that trained on existing Russian distractor dataset, with T5-based model also beating the baseline established by output of an existing English distractor generation model translated into Russian.

Conclusion: The obtained results show the possibility of a translated dataset to be used in distractor generation and the importance of the domain (language examination) and question type match in the input data.

KEYWORDS

automatic distractor generation, multiple-choice questions, reading comprehension, large language model, dataset translation

Citation: Login, N. (2024). Wrong answers only: Distractor generation for Russian reading comprehension questions using a translated dataset. *Journal of Language and Education, 10*(4), 56-70. https://doi.org/10.17323/jle.2024.22244

Correspondence:
Nikita Login,
nlogin@hse.ru

Received: August 15, 2024
Accepted: December 16, 2024
Published: December 30, 2024

INTRODUCTION

Automatic question generation is a promising sphere for application of natural processing techniques as it can enhance the educational processes in multiple ways. According to (Kurdi et al., 2020), standardised examination usually requires exam organisers to keep large banks of curated test exercises which should be regularly updated to prevent cheating. With automated generation these banks can be populated continuously, ensuring variability of test exercises and reducing the costs of organising exams. Furthermore, automatic exercise generation can help test takers, as it can provide them with almost infinite source of test items for preparation.

The availability of sufficient-quality training data is crucial for automatic question generation. Most of the datasets used for model training in automatic question generation were originally designed around the machine reading comprehension problem – these include RACE (Lai et al., 2017), SciQ (Welbl et al., 2017), SQuAD (Rajpukar et al,. 2016), COQA (Reddy et al., 2019), Natural Questions (Kwiatkowski et al, 2019) and TriviaQA (Joshi et al., 2017). Most of these datasets include items consisting of a text passage for reading, a set of questions accompanying the text, right answer and (optionally) a set of distractors for each question. However, there are datasets designed specifically for question generation, including QGSTEC (Rus et al., 2012) and FairyTaleQA (Xu et al., 2022). Among these datasets, RACE is notable for containing exam-style questions as its items were originally extracted from Chinese websites containing English examination materials. For Russian language there are question datasets, such as DaNetQA (Glushkova et al., 2021), MuSeRC/RuCoS (Fenogenova et al., 2020), SberQUAD (Efimov et al, 2020) and RuBQ (Rybin et al., 2021). DaNetQA and SberQuAD contain crowdsourced questions corresponding to Wikipedia paragraphs, RuBQ is based on quizzes and Wikidata, MuSeRC and RuCoS contain crowdsourced questions corresponding to text paragraphs retrieved from a variety of sources. Among Russian datasets, MuSeRC is notable as the only dataset to contain distractors.

Distractor generation is a particularly important subtask of automatic question generation. The advantage of incorporating distractors in online testing materials is that it allows for immediate automated test grading, while excluding the possibility of unfair judgement (as in case with answer matching for open questions). However, this subtask remains one of the most difficult due to the following reasons:

(1) In distractor generation multiple outputs (different independent distractors) correspond to a single input
(2) There cannot be a closed set of ground-truth distractors for a given question, so it is difficult to estimate the performance of a trained model
(3) The generated outputs need to be incorrect in context of the given question but correct in terms of language (Kurdi et al, 2020, p. 145) and also not be too irrelevant to the question

Due the rapid development of neural networks in 2020–2024 years, the most trending approach to question generation nowadays is neural network-based. It is mostly implemented in one of the three ways:

(1) By training/fine-tuning a sequence-to-sequence model (Lee et al., 2020; Makhnytkina et al., 2020; Xiao et al., 2020; Xu et al., 2022; Hadifar et al., 2022; Manakul et al., 2023; Zhang, 2023)
(2) By fine-tuning an autoregressive (designed for text continuation) large language model (Belyanova et al., 2022)
(3) By prompting a large instruct/chat-based model (Elkins et al., 2023; Wang et al., 2023).

Distractor generation generally implements the same techniques – Seq2Seq (Qiu et al., 2020; Hadifar et al., 2022; De-Fitero-Dominguez et al., 2024; Ghanem & Fyshe, 2024), autoregressive (Chung et al., 2020; Ghanem & Fyshe, 2024) and prompting (Bitew et al., 2023; Maity et al., 2024) approaches.

Question and distractor generation are usually automatically evaluated by metrics originally designed for machine translation and text summarization, such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and ROUGE (Lin, 2004). BLEU is based on the geometric mean of modified n-gram precision values. Modified n-gram precision is calculated as the ratio of words in generated sequence that appear in the ground-truth sequence with respect to the unique word counts in the latter. The maximum word length of n-grams used while calculating BLEU is used as an indicator of a specific variant of this metric (BLEU-1, BLEU-2, …). ROUGE metric can be based on recall, precision or their harmonic mean (F-score) with equal weights and has variants depending on n-gram match (ROUGE-N) as well as on Longest Common Subsequence (ROUGE-L). METEOR was developed to address the found issues of BLEU (lack of recall and noisiness of analysed n-grams) and is based on the F-score of unigram match with greater weight of recall over precision.

Fine-tuning of sequence-to-sequence models remains the most popular solution for tackling question text generation problem. (Lee et al., 2020) implemented a BiLSTM-based question generation model jointly trained on two tasks – right answer prediction and question text prediction. Xiao et al. (2020) trained a custom Multi-Flow Attention Transformer (Vaswani et al., 2017) model on question text prediction task using SQuAD dataset. Xu et al. (2022) fine-tuned BART (Lewis et al., 2020) model on FairyTaleQA data for question text generation and reached ROUGE-L F1 score of 52.7. Hadifar et al. (2022) fine-tuned a T5 (Raffel et al., 2020) model for question text generation task on EduQG and SQuAD data, reaching BLEU-4, METEOR and ROUGE-L scores of 15.41, 29.65 and 34.26 correspondingly. Wang et al. (2023) implemented non-fine-tuned GPT-2 (Radford et al., 2019) prompting for question text generation, using a Beam Search extension named NeuroLogicDecoding (Lu et al., 2021). The technique was evaluated on ClariQ-FKw (Sekulić et al., 2021) dataset, reaching BLEU-4, ROUGE-L and METEOR scores of 21.61, 41.03 and 47.87 correspondingly.

As for question text generation for Russian language data, Makhnytkina et al. (2020) used a BiLSTM-based Encode-Decoder model trained on conversational dataset CoQA automatically translated to Russian using Yandex.Translator service. The model reached BLEU-2 score of 12.0. (Belyanova et al., 2022) implemented a RuGPT3 model fine-tuned on DaNetQA and RuBQ corpora. The generation was performed in autoregressive manner, question text was predicted as the continuation of input sequence, right answer text was not used. The model reached BLEU-4 of 4.75 and 1.95 on RuBQ and DaNetQA datasets correspondingly.

In distractor generation sequence-to-sequence approach is also popular. Qiu et al. (2020) used a Seq2Seq model consisting of an Attention-based encoder and a BiLSTM-based decoder for fine-tuning on distractors of RACE-DG dataset, a version of RACE specially pre-processed and filtered for distractor generation by (Gao et al., 2019). They used disjoint decoding in form of Beam Search algorithm on top of model-predicted word probability distributions to get multiple distractors from one input, using Jaccard score to obtain diverse option sets. Their model showed BLEU-4 scores of 7.57/6.27/5.27 for each of the three distractor options correspondingly. Chung et al. (2020) fine-tuned autoregressive BERT (Devlin et al, 2019) language model in a joint learning scheme on two tasks: sequential and parallel prediction of each token of the distractors. They used the same disjoint generation scheme as (Qiu et al., 2020) but applied Maximum Entropy criterion instead of Jaccard score. RACE-DG dataset was used for training and evaluation, the BLEU-4 and ROUGE-L scores on test subset were 13.56 and 34.01 correspondingly. However, later they released[1] enhanced versions of their models based on sequence-to-sequence BART architecture, that reached maximum BLEU-4/ROUGE-L of 16.33/37.5 correspondingly.

In more recent distractor generation works sequence-to-sequence T5 architecture is widely used. Hadifar et al. (2022) implemented distractor generation using a T5 model trained on both RACE and their own new dataset EduQG. The whole set of distractors was predicted at once. Obtained BLEU/METEOR/ROUGE-L scores on EduQG consisted 17.73/21.54/34.13 correspondingly. Ghanem & Fyshe (2024) fine-tuned GPT-2 and T5 models on distractor generation task as a part of work on their prediction-based distractor generation quality metric DISTO. They used RACE dataset for fine-tuning and evaluation and implemented two versions of T5 – with joint and disjoint distractor generation. Their best solution, a disjoint T5 model, reached 2.3 in terms of BLEU-4 while GPT-2 and joint T5 reached only 0.3 and 0.9 BLEU-4 scores correspondingly. De-Fitero-Dominguez et al. (2024) implemented distractor generation using mT5 (Xue et al, 2020) model, a multilingual version of T5, on a combined translated distractor dataset. Their dataset included items from RACE-DG, CosmosQA (Huang et al., 2019) and SciQ, translated with Opus-MT (Tiedemann & Thottingal, 2020) model. Their implementation reached 7.21 and 21.76 on test subset of RACE-DG in terms of BLEU-4 and ROUGE-L metrics.

Distractor generation was also treated as a ranking problem, as it was implemented by Bitew et al. (2022) – the models were trained to select the most appropriate distractors for the given question and a right answer to it. Two ranking solutions were implemented – one using feature engineering and logistic regression and another one using a multilingual BERT model. Three BERT-based models were used – based on distractor-right answer similarity, based on distractor-question similarity and a joint model combining the two beforementioned. Average precision and recall of ranking were used as quality metrics, and the highest scores (57.3 and 62.8 respectively) were obtained by a joint BERT-based model.

The most recent works also experimented with prompting approach to distractor generation. Bitew et al. (2023) addressed distractor generation by using a T5 model trained on Televic dataset and ChatGPT prompting in zero-shot (using a prompt without distractor examples) and few-shot (using a prompt with examples) configurations. All models were evaluated manually by experts. Maity et al. (2024) used a multi-level pipeline based on ChatGPT and DaVinci, consisting of input text paraphrase generation, keyword extraction from paraphrase, question generation and distractor generation itself. The best BLEU-4 and ROUGE-L scores (2.49 and 13.54 accordingly) were obtained by a Davinci-based multilevel model.

From all the reviewed works only Qiu et al. (2020) specifically address the issue of potential triviality of distractors in design of their solution. In their work they view triviality as irrelevance to the given question and reading passage, and claim to solve this issue by incorporating blocks that combine information from the reading passage and question text (referred as ‘Reforming Passage’ and ‘Reforming Question’ modules) to their original Transformer-based model. However, no clear conclusions on how the exclusion of both reforming modules affects the metrics (only exclusion of each module separately is analysed) are made and situations where triviality is not connected with the relatedness to input data are not accounted.

When there are numerous reading comprehension item generation solutions for English language, only a few were developed for Russian (Makhnytkina et al., 2020; Belyanova et al., 2022), and there was no evidence found of solutions that tackled distractor generation for Russian-language questions. Also, only one of existing Russian-language reading comprehension question datasets contains distractors, and none of these datasets is tackled specifically for language examination in reading comprehension. Another issue is that the parameters such as type and structure of questions are not utilised in distractor generation in the previous works, whereas accounting for these parameters

[1] https://github.com/voidful/BDG

in DG model design may have a potential of making the task easier for NN models. The importance of these parameters can by demonstrated by findings of Xu et al. (2022), who implemented question categorization in the design of their question dataset FairyTaleQA using a system of narrative elements and relations described in Paris & Paris (2003). They found that nature of the answer can depend on the narrative category of question on the example of "Feeling" question type.

Taking into account the importance of automatic exam-style reading comprehension test generation and the lack of solutions for distractor generation on Russian data, the purpose of this Paper is to develop an automatic distractor generation solution for Russian reading comprehension exam-style questions. Due to the lack of Russian exam-style distractor datasets, we also aimed to explore the possibility of using a translated high-quality English dataset for Russian distractor generation, as it was done by Makhnytkina et al. (2020) for question generation and by De-Fitero-Dominguez et al. (2024) for Spanish data. Additionally, we aimed to investigate the prospects of fine-tuning distractor generation models on specific categories of questions. We expected that a rich and thoroughly curated English dataset would serve as an efficient source of training data and that training on a specific category of questions would allow for better transferability of DG model intelligence to standardised examination questions. We formed our research questions as follows:

RQ#1: Whether a distractor generation model can be effectively fine-tuned on an English dataset, that was automatically translated to Russian?

RQ#2: Is there a need for a specific Russian multiple-choice reading comprehension question dataset for efficient exam-style distractor generation or is an existing non-exam-style dataset MuSeRC appropriate for this task?

RQ#3: Can fine-tuning on a specific type of questions result in better performance of distractor generation model on standardised exam data?

METHOD

Research Design

In this work we have performed large language model fine-tuning experiments on the task of distractor generation for reading comprehension questions using different datasets. Our primary focus was on fine-tuning on a translated English-language dataset. For that purpose, we have used RACE, as it contains reading comprehension multiple-choice questions in language examination style and have been used in many distractor generation works (Chung et al., 2020; Qiu et al., 2020; De-Fitero-Dominguez et al., 2024; Ghanem & Fyshe, 2024). Also, we have included an originally Russian multiple-choice reading comprehension dataset (MuSeRC) to see if a translated dataset was necessary and if it was possible to reach appropriate generation quality by using already available Russian data.

In our experiments we have fine-tuned two large language models made available (Zmitrovich et al., 2024) by AI-Forever team – RuGPT3 and RuT5, which are Russian-language implementations of GPT3 (Brown et al., 2020) and T5 (Raffel et al., 2020) models correspondingly. RuGPT3 employs autoregressive text generation and consists only of Transformer decoder blocks, while RuT5 is based on Sequence-to-Sequence approach and contains both encoder and decoder. For evaluation of our generated output, we have used conventional automatic Sequence-to-Sequence generation quality metrics (see "Assessment" subsection of "Methods").

To compare models trained on different datasets we have also used a small set of original Russian examination data – USE, obtained from open-access Internet sources. We have also included baselines in our evaluation, including a non-fine-tuned version of RuGPT3 and enhanced versions of models from Chung et al. (2020), which outputs were translated to Russian automatically.

Datasets

RACE

RACE is a dataset consisting of 98,000 questions on English reading comprehension, designed for Chinese middle- and high-schoolers as a part of the national exam. Each text of RACE was accompanied by several multiple-choice questions, and each of the questions was accompanied by 5 answer options – 1 correct and 4 incorrect. We translated RACE dataset using Opus-MT English-to-Russian translation model available from EasyNMT[2] Python package. Each question and set of distractors were translated in concatenation with the reading text in order for the translation model to not lose context. After a descriptive analysis of RACE questions performed in Microsoft Excel and Python environments, we have found two distinct question categories suitable for distractor generation:

- Questions asking the participant to select the best title for the given passage (TITLE);
- Questions asking the participant to select TRUE or FALSE sentences from the given set (TF).

[2] https://github.com/UKPLab/EasyNMT

We collected questions of these types using regular expression search. Datasets obtained using this technique – Ru-RACE-TITLE and Ru-RACE-TF – contained 4892 and 3799 items correspondingly.

For Ru-RACE-TITLE we selected 805 unique question texts from RACE that matched the regular expression ***\Wtitle\W*** (contained the word ***title***). Then we manually filtered out 53 irrelevant question texts (e.g. containing word ***title*** referring to a person's social status or asking about the title of some item referenced in the reading text). The resulted dataset was split into train/test/dev subsets using original subset labels from RACE, which resulted in 4575/219/242 split.

For Ru-RACE-TF we selected question texts which lowercased variants matched the regular expression ***which of the following .+(true|false)***. This way 693 unique question texts were retrieved. 143 question texts were filtered out manually. Applying the same split logic as in Ru-RACE-TITLE, we got 3288/175/187 split. The formulation of Ru-RACE-TF task is identical to Task 18 in USE exam in Russian language, which allowed us to use USE data, as described in "USE-TF" section of this paper.

MuSeRC

To compare performance on the translated dataset with performance on the original Russian data, we have also used MuSeRC. MuSeRC is a dataset created by (Fenogenova et al, 2020) as a part of the RussianSuperGLUE benchmark. It contains 12,805 Russian multiple choice reading comprehension questions made by crowdsource workers from texts of different domains. Each text is accompanied by a set of questions, each question includes mostly 1-2 right answers and 2-3 distractors.

USE-TF

USE (Unified State Examination*, Единый Государственный Экзамен*, *Edinyi Gosudarstvennyi Examen*,) is a compulsory Russian state exam which is used for assessment of knowledge of high school graduates and as an entrance test for higher education institutions. Format of USE in subject of Russian language contains Task 18, which is a multiple-choice reading comprehension question asking the participants to select either TRUE or FALSE sentences from the given set. The data for this task contained test items collected by Shavrina et al. (2020), as well as obtained from other openly available Internet sources. This set contained 55 unique questions with 5 answer options for each. Some questions contained more than one correct option, so we have preprocessed them as described in "Data Preprocessing" section of this paper.

Methods

Data Preprocessing

In Ru-RACE-TITLE all question texts were replaced with *Какое название лучше всего подойдёт для этого текста?* ("Which is the best title for this text?"). In Ru-RACE-TF question texts were replaced with either *Какое высказывание СООТВЕТСТВУЕТ тексту?* ("Which statement is TRUE according to the text?") or *Какое высказывание НЕ СООТВЕТСТВУЕТ тексту?* ("Which statement is NOT TRUE according to the text?"). MuSeRC question and option texts were left unchanged as we wanted to perform training and evaluation on the whole original dataset splits.

For USE-TF we have applied the same preprocessing procedure as for RuRACE, with enhancements addressing having more than one correct option. For items that had more correct options than incorrect, we changed the question text to the opposite (*Какое высказывание СООТВЕТСТВУЕТ тексту?* was changed to *Какое высказывание НЕ СООТВЕТСТВУЕТ тексту?* and vice versa). Then first of the original distractors was used as the correct answer and original correct options were used as distractors. If there were more incorrect options than correct, the question was unchanged and the first right option was used as a right answer, while the distractors were used without changes.

Model Training

The models were trained on a remote private server with a Nvidia Tesla V100 GPU. All models were trained for 20 epochs with ADAM optimizer, initial learning rate of 5e-5 and weight decay factor of 0.01. We defined the maximum output length for training and inference as 0.99 quantile of input length on the training set. Training subsets were used for model fine-tuning, while testing and development subsets were exploited for evaluation.

For training phase, we constructed input examples for RuGPT3 as concatenations of reading passage, question text, right answer and a line-separated set of distractors, interrogated by Russian phrases indicating the parts of an input example (*ВОПРОС*, *ПРАВИЛЬНЫЙ ОТВЕТ* and *НЕПРАВИЛЬНЫЕ ВАРИАНТЫ ОТВЕТА*). For training of RuT5 we constructed separate input and output examples, as the model (as opposed to RuGPT3) worked not in an autoregressive but in a Sequence-to-Sequence way. The input example for RuT5 included a reading passage, a question and a right answer, interrogated by the same phrases as RuGPT3 examples, while the output example consisted of distractors enclosed in double quotes and separated by semicolons.

Model Inference

At the inference phase, input examples for RuGPT3 had the same structure as at the training phase, but included only reading text section, question text and right answer. For those models we generated text until our maximum defined length was reached. After that we split the predicted continuation of input by line breaks. After that, we filtered out distractors that were either non-unique or identical to the right answer. Then we sorted the retrieved set of distractors by alphabet and kept the first 3 results. For RuT5 we used maximum length as well as end-of-sequence token as stopping points for generation, while retrieving the distractors by splitting the output by semicolons and removing the enclosing quotes.

Assessment

BLEU and METEOR metric values were used for automatic assessment of the generated distractors. The implementations of BLEU and METEOR accessible from Evaluate[3] Python package were used. In order to make our results comparable with the previous and forthcoming works we have also included ROUGE-L metric in our evaluation. As the official implementation of ROUGE, accessible from Evaluate package, cannot process Russian-language data, we have used an unofficial implementation of it[4]. However, the authors of this implementation admit that the values obtained from it may differ from the official variant. As most of the previous works (Chung et al, 2020; Qiu et al, 2020; Belyanova et al, 2022; Wang et al, 2023; Maity et al, 2024) utilise the 4-gram version of BLEU, this was the BLEU configuration used by default in our paper. We have also used BERTScore (Zhang et al, 2023) for semantic assessment of the generated distractors. BERTScore is a metric based on similarities of word embeddings from the BERT model instead of exact word/n-gram matches. To enhance the convenience of interpreting the results, all metric values (defined from 0 to 1) were presented as percentages, ranging from 0 to 100.

Baselines

As, during the work on this paper, we have curated USE-TF dataset and our original modifications of RACE dataset, we tested baseline models on our data instead of just reporting scores from previous works. This was done in order to allow for fair comparison, as metrics used during our evaluation cannot be directly compared across different languages. The implementation of baseline models is described in this section, whereas their results are reported and analysed in comparison with implemented models in “Results” and “Discussion” sections.

As a first baseline in our experiments, we have used a non-fine-tuned version of RuGPT3. Along with a zero-shot RuGPT3, we have also used BART-DG models, enhanced versions of models introduced by Chung et al. (2020), that hold the state-of-the-art results in terms of BLEU in distractor generation on RACE data. In order to produce Russian outputs from these models, the same translation pipeline that was used in compiling of Ru-RACE was used to translate their outputs to Russian. USE inputs were translated to English using the same multilingual translation model (Opus-MT) before feeding them to BART-DG models.

RESULTS

Ru-RACE

Table 1 illustrates results of models trained on translated RACE subsets. In both tasks the best performance was

Table 1

Results of Models on Translated RACE Subsets

	BLEU-4		METEOR		ROUGE-L		BERTScore	
	dev	**test**	**dev**	**test**	**dev**	**test**	**dev**	**test**
Ru-RACE-TITLE								
RuGPT3-RACE-TITLE	3.83	3.19	12.78	12.41	12.32	12.60	68.72	68.68
RuT5-RACE-TITLE	**25.17**	**22.96**	**46.09**	**45.35**	**16.79**	**16.21**	**79.09**	**78.72**
Baseline RuGPT3	0.46	0.53	5.37	5.57	4.31	4.47	62.72	62.46
Ru-RACE-TF								
RuGPT3-RACE-TF	8.75	4.89	18.92	16.84	16.16	13.80	71.01	70.23
RuT5-RACE-TF	**26.36**	**22.43**	**44.84**	**42.75**	**28.36**	**25.30**	**77.07**	**76.24**
Baseline RuGPT3	1.23	1.73	9.54	9.44	8.30	8.29	63.64	64.04

[3] https://pypi.org/project/evaluate/

[4] https://github.com/pltrdy/rouge

demonstrated by T5-based models (RuT5-RACE-TITLE and RuT5-RACE-TF). This can be attributed to the Sequence-to-Sequence nature of T5 that allows it to transform inputs to outputs that have slightly different structure. Both models have surpassed the baseline established by non-fine-tuned RuGPT3 in both tasks, so we can conclude that fine-tuning allowed them to successfully adapt to the structure of our translated datasets.

For Ru-RACE-TITLE, the highest quality in terms of BLEU, METEOR and BERTScore on both dev and test subsets was reached by fine-tuned RuT5 model. Fine-tuned RuGPT3 demonstrated definitely lower results, with BLEU-4 reaching only 3.83 and 3.19 for dev and test subsets correspondingly (compared to 25.17 and 22.96 of RuT5-RACE-TITLE). Even higher absolute difference can be spotted in METEOR, with 12.78/12.41 on dev/test sets for RuGPT3 against 46.09/45.35 for RuT5. These differences indicate that RuT5-RACE-TITLE model greatly surpasses RuGPT3-RACE-TITLE both in terms of precision and recall. As for BERTScore, the difference between the two fine-tuned models on test set (10.04) is higher than the difference between the least scoring model and the baseline (4.24), which indicates that the RuT5-RACE-TITLE's title ability to produce semantically coherent distractors highly surpasses that of RuGPT3-RACE-TITLE. In terms of ROUGE-L values, the two fine-tuned models are not so far apart (4.47/3.61 on dev/test subsets) but both of them greatly surpass the baseline. The performance of all models on dev and test subsets is quite close, which proves that the models were not overfitted on the validation sets during hyper-parameter tuning.

For Ru-RACE-TF, the highest quality in terms of BLEU-4/METEOR/BERTScore was also reached by fine-tuned RuT5 model and the scores of fine-tuned RuGPT3 were also substantially lower (by 17.54/25.91/6.01 points on test set correspondingly). BLEU-4 and METEOR scores of RuGPT3-RACE-TITLE lie closer to the baseline than to the values of RuT5-RACE-TITLE. However, BERTScore differences between the baseline and the second-scoring model and between the second-scoring model and the baseline are quite close (7.37/6.19 against 6.06/6.01 on dev/test subsets), which suggests that the gap in semantic coherence between the two fine-tuned models might be not so broad. The difference in ROUGE-L appears to be equally broad both between the second-scoring model and the baseline and between the first- and second-scoring models. The difference in distractor generation quality between development and test subsets for Ru-RACE-TF is not enough for an overfit to be spotted. It can be seen that all model scores in this task are higher than in Ru-RACE-TITLE.

MuSeRC

Table 2 illustrates results of models trained on MuSeRC dataset. Due to MuSeRC test subset not being available at the dataset's developer website, all evaluation was performed on the development set. Both models beat the attested zero-shot GPT3 baseline in all of the three metrics. BLEU-4 score of RuT5-MuSeRC-DG is nearly twice as better as that of RuGPT3-MuSeRC-DG (23.62 against 12.48), while METEOR score of RuT5-MuSeRC-DG is only slightly higher (45.78 against 40.87). According to BERTScore values (76.02 and 76.02 for RuT5-MuSeRC-DG and RuGPT3-MuSeRC-DG correspondingly), distractors generated by the two models are nearly equally semantically similar to the gold standard distractors. In terms of ROUGE-L the values of two fine-tuned models are quite close and both greatly surpass the zero-shot baseline. As both fine-tuned models produce results that beat the non-fine-tuned baseline, they were used for evaluation on USE data.

USE-TF

Table 3 illustrates results on the USE-TF dataset. The highest values of metrics were reached by RuT5-RACE-TF model, with BART-DG-PM model holding the second place. However, we can see that the scores of translated BART-DG-PM outputs (11.02/28.47/70.90 in terms of BLEU-4/METEOR/BERTScore) are quite close to scores of our best model (11.64/29.61/71.06 correspondingly).

We can see that RuT5-RACE-TF model displays robustness when dealing with data of USE, as its BLEU, METEOR and BERTScore values still greatly exceed the unsupervised baseline. However, this is not true for other models trained on Russian data, as their metric values degrade closer to baseline values attested by a zero-shot RuGPT3. The BLEU scores of all models, excluding RuT5-RACE-TF and BART-DG-based, rapidly decrease to zero with increasing the rank of BLEU, which indicates the lack of robustness in these models. This is especially true for models trained on MuSeRC, which means that existing Russian datasets cannot offer data that is suitable for distractor generation for complex general reading comprehension tasks, which can be found in language exam materials. We can conclude that training on the translated dataset can offer robustness of results while training on existing Russian-language dataset cannot. This can be attributed to MuSeRC dataset containing more trivial texts than USE, as MuSeRC consists mostly of news reports, while USE texts are usually extracts from high school-level literary works posing ethical problems worth of discussion. Due to the RuT5-RACE-TF being our only robust model in comparison, it is planned to use only its predictions in the future manual evaluation of data produced on the basis of USE-TF.

DISCUSSION

The results (Tables 1-3) demonstrate that the performance on translated datasets in distractor generation task is on par with the existing works, with BLEU-4 values reaching a maximum of around 25 for test subsets of datasets they were originally fine-tuned on. RuGPT3-based fine-tuned models

Table 2

Results of Models on MuSeRC Dataset

	BLEU-4	METEOR	ROUGE-L	BERTScore
RuGPT3-MuSeRC-DG	12.48	40.87	21.77	**76.04**
RuT5-MuSeRC-DG	**23.62**	**45.78**	**25.97**	76.02
Baseline RuGPT3	5.16	11.25	6.81	62.91

Table 3

Results of Models on USE-TF Dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	BERTScore
RuGPT3-RACE-TF	15.11	3.57	0.08	0.00	9.22	6.83	65.84
RuT5-RACE-TF	**29.11**	**20.93**	**15.66**	**11.64**	**29.61**	**13.55**	**71.06**
RuGPT3-MuSeRC	9.56	1.69	0.48	0.00	6.65	4.35	61.62
RuT5-MuSeRC	10.30	2.20	0.55	0.00	7.77	4.66	62.63
Baseline RuGPT3	11.22	2.00	0.53	0.00	7.57	4.74	55.72
BART-DG	26.66	19.44	14.64	10.78	27.77	12.57	70.83
BART-DG-PM	28.52	20.24	15.06	11.02	28.47	12.40	70.90
BART-DG-ANPM	27.39	19.71	14.64	10.71	27.78	11.75	70.62

have shown generally better performance at producing consistent distractor outputs than RuT5-based, which can be attributed to Sequence-to-Sequence RuT5 being pre-trained on a text reconstruction rather than text generation task and thus being more prone to fine-tuning. From our results we can see that only RuT5 was able to produce coherent outputs both on its original dataset and on the independent dataset of USE questions, while our other fine-tuned models were able to do so only on the data from test subset of the dataset they were fine-tuned on. In this section we will explain the relationships between scores of different models on different datasets and propose ways to improve our results, while comparing our findings with that from the previous works on the subject.

The inability of models trained on MuSeRC to produce coherent distractor outputs for USE can be explained by the nature of MuSeRC dataset and relatively lower complexity of its items compared to real-word reading comprehension examination tasks. The fact of BART-DG (enhanced versions of models from Chung et al., 2020) holding a strong baseline against our results on USE data can be explained by the complex structure of BDG models, that features additional engineering techniques applied to base model. These techniques include entropy maximization-based decoding on different generation paths to produce multiple distractors independently (while our models produce them consecutively), parallel multi-task training (PM) and answer-negative (AN) regularisation.

The ability of T5-based models to be successfully fine-tuned on distractors from RACE questions compared to models of different architectures, found in our study, is supported by works of Hadifar et al. (2022), Ghanem & Fyshe (2024) & De-Fitero-Dominguez et al. (2024). The gap in generation quality metrics between decoder-only and encoder-decoder models is also found in Ghanem & Fyshe (2024), where fine-tuned T5-base models significantly outperforms fine-tuned GPT2-small (8.4 and 13.7 in terms of BLEU-2 for joint and disjoint T5 correspondingly against 3.9 for GPT2).

The results of comparison against BDG on the same set of data are contrastive to those of Ghanem & Fyshe (2024), who found that their best model outperforms BDG only in terms of BLEU-1 (32.0 against 30.2). However, authors of the referenced work fine-tuned BDG model on their data, whereas in our work we use readily fine-tuned models. Also, they implemented disjoint generation of distractors (as was also done by Chung et al., 2020) and succeeded to achieve a performance gain by it, whereas in our work only joint approach is implemented.

Better generation quality on the translated dataset than on an original-language one, found in our study, was also encountered by De-Fitero-Dominguez et al. (2024). However, it may be caused by reduction of lexical space imposed by the lexical knowledge of the translation model. However, the results on translated data have not always been especially high. For example, in Makhnytkina et al. (2020) question generation model demonstrated poor performance on test subset of dataset it was fine-tuned on in terms of formal metrics. Nevertheless, these results can be explained by the use of older model architecture (BiLSTM) and the lesser development of translation models in 2020.

The results of our experiments, where METEOR score in all settings is higher than ROUGE-L, contradict the findings of Hadifar et al. (2022), where METEOR score was much lower than ROUGE-L. Taking into account differences in metrics calculation (much higher weight of recall in METEOR than in ROUGE and use of longest common subsequence length in ROUGE-L instead of unigram match in METEOR), it can be deducted that while their solution better captures patterns from distractor data than preserves the lexical content, the opposite is true for our solutions.

Taking into account underperformance of GPT3-based models in our experiments, it is worth noting that we have used "small" version of RuGPT3 for fine-tuning due to the lack of computational resources needed to fine-tune larger versions of RuGPT3. Considering the closeness of our results on independent USE-TF dataset to BART-DG models, it is worth noting that they implemented disjoint decoding of output distractors, while our solutions use joint decoding. Taking into account the advantage in metric values of disjoint generation over joint, found Ghanem & Fyshe (2024), it is possible that our models will outperform BART-DG by higher value if we implement disjoint decoding.

Our expectations about the results, described in the "Introduction" section (efficiency of fine-tuning on translated high-quality dataset and the advantage of fine-tuning on specific type of questions), are met by performance of RuT5-RACE-TF model on USE-TF data, which beats both baselines and performance of models trained on Russian-language dataset containing questions of different types. However, these expectations are not met by RuGPT3-RACE-TF model, that surpasses results of MuSeRC-trained models, but not the baselines attested by models trained on full RACE-DG dataset.

It is worth noting that the reported formal quantitative metrics are based on the similarity between the generated and original distractors and the overall plausibility of distractors should be estimated by human evaluation, which is described in the "Future Work" subsection. While our models utilise base implementations of T5 and GPT3, further engineering enhancements can be applied to them.

Future Work

For future manual evaluation of our models fine-tuned on Ru-RACE-TF, it is planned to use USE-TF data (as a professionally curated set of originally Russian multiple-choice questions) and include predictions from RuT5-RACE-TF and BART-DG-PM models. It is planned to equip each question of USE-TF with 4 answer options – the original right answer, a "filler" distractor, one of the original distractors, one prediction from our fine-tuned model and one prediction from BART-DG-PM. A filler distractor can be a sentence extracted from an existing Russian corpus, that is semantically close (as attested by a formal metrics, such as BERTScore) to the reading text passage. The plan is to attract Russian-speaking participants with higher education, so the test takers will be able to actually distinguish distractors from the real right answers. The participants will be asked to rate each of the examples on a scale from 1 to 5, with 1 indicating the most unsuitable option and 5 indicating the option most likely to be the right answer. The hypothesis is that the distractors from our models will be on average rated higher than filler distractors but lower than the original right answers.

For manual evaluation of our best model trained on Ru-RACE-TITLE (RuT5-RACE-TITLE) it is planned to use arbitrary Russian texts as inputs for the distractor generation models. This may include extracts from newspaper sources and stories for children, as most of the texts from RACE are of narrative nature. The design of questions will be the same as for Ru-RACE-TF evaluation set, the original right answer will be implemented as the article's original title or created manually. The same findings about relationships between average rank values as from USE-TF dataset are expected hypothetically. As metric values for fine-tuned models on the Ru-RACE-TF were higher than on Ru-RACE-TITLE, it is expected that during proposed manual evaluation average ranks of distractors for Ru-RACE-TF will also be higher than ranks for Ru-RACE-TITLE.

While accessing the overall plausibility of generated options, this method can also help determine how often trivial distractors are generated, as the trivial options are expected to be averagely rated on par with "filler" distractors. Also, it may be beneficial annotate a set of model distractor predictions in terms of plausibility and triviality. The annotation of distractor characteristics gained through this procedure can be used for future training of a distractor assessment ML model. Although triviality is not usually tackled specifically in recent works, as modern generative NNs are able to extract patterns from the presented non-structured data without the need for additional engineering-based output conditioning, this model can be used during future training of a new distractor generation pipeline by penalising outputs that would be predicted to be too trivial and rewarding outputs which prediction of plausibility would be high. Additional enhancements may include disjoint output decoding, the use of larger model variants and the implementation of prompting approach as an alternative.

CONCLUSION

In this paper automatic distractor generation was implemented for Russian data. 6 large language models of two types (GPT-3 and T5) were fine-tuned on distractor generation tasks on 3 datasets – 2 machine-translated English-to-Russian question datasets containing only specific types of questions (title selection and true/untrue fact selection) and an originally Russian dataset. RuT5-based models demonstrated generally better results than RuGPT3-based. Both model types surpassed the unsupervised baseline attested by non-fine-tuned RuGPT3 model, proving the possibility of effective fine-tuning on distractor generation task on English-to-Russian translated data addressed in RQ1.

During the experimentation on Russian examination data, it was found out that translated English reading comprehension examination dataset is more efficient in terms of use in model fine-tuning than an existing Russian non-examination reading comprehension dataset, as the models trained on the latter dataset demonstrated poor performance compared to models trained on the former. This highlights the importance of domain and complexity level of questions in distractor generation task and proves the need of a comprehensive Russian exam-style multiple-choice reading comprehension question dataset addressed in RQ2. T5-based model fine-tuned on true statement selection distractors demonstrated better performance on USE data than both MuSeRC-trained models and the state-of-the-art exam-style distractor generation solution, demonstrating the advantage of fine-tuning on a specific type of questions, which possibility was addressed in RQ3.

The value of work lies in training distractor generation models for Russian-language data, which has not been done in the previous works. The explored possibility of transferring intelligence learned on specific categories of questions, found in large-scale datasets, to distractor generation for exam questions of a specific standard presents an additional value. Our findings can be beneficial for exam preparation platform creators, who can include models trained in the described settings into their products, allowing for the automatic item bank replenishment.

Our findings can also be helpful for reading comprehension dataset creators. The annotation of distinct common question types can be implemented by them in dataset design. The lack of exam-style Russian question datasets needed for successful model learning can present an opportunity for them.

The future work in this direction should include manual evaluation of the generated data and the development of a Russian reading comprehension dataset designed specifically for examination. Another direction of future work lies in exploring the possibilities of different alternative generation techniques not covered in described experiments. A detailed comparison of Russian-language distractors obtained from models trained on distractor generation task and retrieved from chat- and instruct-based large language models via prompting can present another interesting direction for future research.

ACKNOWLEDGMENTS

The author wants to thank his academic advisor Prof. Olga Lyashevskaya for providing beneficial advice during the preparation of revised manuscript. This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

DECLARATION OF COMPETITING INTEREST

None declared.

REFERENCES

Alsubait, T. M. (2015). *Ontology-based multiple-choice question generation* [Unpublished PhD thesis]. University of Manchester.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgements. In J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds.), *Proceedings of the ACL Workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). Association for Computational Linguistics.

Belyanova, M. A., Andreev, A. M., & Gapanyuk, Y. E. (2022). Neural text question generation for Russian language using hybrid intelligent information systems approach. In B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, Y. Tiumentsev, & V. V. Klimov (Eds.), *Advances in neural computation, machine learning, and cognitive research V* (vol. 1008, pp. 217–223). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-91581-0_29

Bitew, S. K., Hadifar, A., Sterckx, L., Deleu, J., Develder, & C., Demeester, T. (2022) Learning to reuse distractors to support multiple choice question generation in education. *IEEE Transactions on Learning Technologies, 17,* 375–390. IEEE Computer Society Press. http://dx.doi.org/10.1109/TLT.2022.3226523

Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023) *Distractor generation for multiple-choice questions with predictive prompting and large language models* (Version 1). arXiv. http://dx.doi.org/10.48550/arXiv.2307.16338

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (vol. 33, pp. 1877–1901). Curran Associates, Inc. http://dx.doi.org/10.48550/arXiv.2005.14165

Chung, H.-L., Chan, Y.-H., & Fan, Y.-C. (2020). A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4390–4400). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.393

De-Fitero-Dominguez, D., Garcia-Lopez, E., Garcia-Cabot, A., Del-Hoyo-Gabaldon, J.-A., & Moreno-Cediel, A. (2024). Distractor generation through text-to-text transformer models. *IEEE Access*, *12*, 25580–25589. http://dx.doi.org/10.1109/ACCESS.2024.3361673

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (vol. 1: Long and Short Paper, pp. 4171–4186). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1423

Efimov, P., Chertok, A., Boytsov, L., & Braslavski, P. (2020). SberQuAD – Russian reading comprehension dataset: Description and analysis. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, & N. Ferro (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction* (vol. 12260, pp. 3–15). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-58219-7_1

Elkins, S., Kochmar, E., Serban, I., & Cheung, J. C. K. (2023). How useful are educational questions generated by large language models? In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (vol. 1831, pp. 536–542). Springer Nature Switzerland. http://dx.doi.org/10.1007/978-3-031-36336-8_83

Fenogenova, A., Mikhailov, V., & Shevelev, D. (2020). Read and reason with MuSeRC and RuCoS: Datasets for machine reading comprehension for Russian. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6481–6497). International Committee on Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.coling-main.570

Gao, Y., Bing, L., Li, P., King, I., & Lyu, M. R. (2019). Generating distractors for reading comprehension questions from real examinations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6423–6430. http://dx.doi.org/10.1609/aaai.v33i01.33016423

Ghanem, B. & Fyshe, A. (2024). DISTO: Textual distractors for multiple choice reading comprehension questions using negative sampling. In M. Marras, M. Ueno (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 23–34). International Educational Data Mining Society. http://dx.doi.org/10.5281/ZENODO.12729766

Glushkova, T., Machnev, A., Fenogenova, A., Shavrina, T., Artemova, E., & Ignatov, D. I. (2021). DaNetQA: A yes/no question answering dataset for the Russian language. In W. M. P. Van Der Aalst, V. Batagelj, D. I. Ignatov, M. Khachay, O. Koltsova, A. Kutuzov, S. O. Kuznetsov, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, M. Pelillo, A. V. Savchenko, & E. Tutubalina (Eds.), *Analysis of Images, Social Networks and Texts* (vol. 12602, pp. 57–68). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-72610-2_4

Hadifar, A., Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023). EduQG: A multi-format multiple-choice dataset for the educational domain. *IEEE Access, 11*, 20885–20896. http://dx.doi.org/10.1109/ACCESS.2023.3248790

Huang, L., Le Bras, R., Bhagavatula, C., & Choi, Y. (2019). CosmosQA: Machine reading comprehension with contextual commonsense reasoning. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2391–2401). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D19-1243

Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In R Barzilay., & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational linguistics* (vol. 1: Long Papers, pp. 1601–1611). Association for Computational linguistics. http://dx.doi.org/10.18653/v1/P17-1147

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. http://dx.doi.org/10.1007/s40593-019-00186-y

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, *7*, 453–466. http://dx.doi.org/10.1162/tacl_a_00276

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale reading comprehension dataset from examinations. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 785–794). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D17-1082

Lee, D. B., Lee, S., Jeong, W. T., Kim, D., & Hwang, S. J. (2020). Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 208–224). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.20

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.703

Lin, C.-Y. (2004). ROUGE: A Package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013

Lu, X., West, P., Zellers, R., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). NeuroLogic decoding: (Un)supervised neural text generation with predicate logic constraints. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4288–4299). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2021.naacl-main.339

Maity, S., Deroy, A., & Sarkar, S. (2024). A novel multi-stage prompting approach for language agnostic MCQ generation using GPT. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in information retrieval* (vol. 14610, pp. 268–277). Springer Nature Switzerland. http://dx.doi.org/10.1007/978-3-031-56063-7_18

Makhnytkina, O., Matveev, A., Svischev, A., Korobova, P., Zubok, D., Mamaev, N., & Tchirkovskii, A. (2020). Conversational question generation in Russian. In S. Balandin, L. Turchet, & T. Tyutina (Eds.), *2020 27th Conference of Open Innovations Association (FRUCT)* (pp. 1–8). IEEE. http://dx.doi.org/10.23919/FRUCT49677.2020.9211056

Manakul, P., Liusie, A., & Gales, M. (2023). MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, & A. A. Krisnadhi (Eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific chapter of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 39–53). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2023.ijcnlp-main.4

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *The 40th Annual Meeting on Association for Computational Linguistics—ACL '02* (pp. 311–318). Association for Computational Linguistics. http://dx.doi.org/10.3115/1073083.1073135

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly*, *38*(1), 36–76. http://dx.doi.org/10.1598/RRQ.38.1.3

Qiu, Z., Wu, X., & Fan, W. (2020). Automatic distractor generation for multiple choice questions in standard tests. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 2096–2106). International Committee on Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.coling-main.189

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research, 21, 1*, 5485–5551. https://dl.acm.org/doi/abs/10.5555/3455716.3455856

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D16-1264

Reddy, S., Chen, D., & Manning, C. D. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics, 7*, 249–266. http://dx.doi.org/10.1162/tacl_a_00266

Rybin, I., Korablinov, V., Efimov, P., & Braslavski, P. (2021). RuBQ 2.0: An innovated Russian question answering dataset. In R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, & M. Alam (Eds.), *The Semantic Web* (vol. 12731, pp. 532–547). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-77385-4_32

Sekulić, I., Aliannejadi, M., & Crestani, F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 167–175). Association for Computing Machinery. http://dx.doi.org/10.1145/3471158.3472257

Shavrina, T., Emelyanov, A., Fenogenova, A., Fomin, V., Mikhailov, V., Evlampiev, A., Malykh, V., Larin, V., Natekin, A., Vatulin, A., Romov, P., Anastasiev, D., Zinov, N., & Chertok, A. (2020, May). Humans keep it one hundred: An overview of AI Journey. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A.

Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2276–2284). European Language Resources Association. https://aclanthology.org/2020.lrec-1.277/

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – Building open translation services for the world. In A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, & M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 479–480). European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.61/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (vol. 30, 6000–6010). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3295222.3295349

Welbl, J., Liu, N. F., & Gardner, M. (2017). Crowdsourcing multiple choice science questions. In L. Derczynski, W. Xu, A. Ritter, & T. Baldwin (Eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text* (pp. 94–106). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W17-4413

Xiao, D., Zhang, H., Li, Y., Sun, Y., Tian, H., Wu, H., & Wang, H. (2020). ERNIE-GEN: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In C. Bessiere (Ed.) *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (pp. 3997–4003). International Joint Conferences on Artificial Intelligence Organization. http://dx.doi.org/10.24963/ijcai.2020/553

Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., Zhang, Z., Li, T., Bradford, N., Sun, B., Hoang, T., Sang, Y., Hou, Y., Ma, X., Yang, D., Peng, N., Yu, Z., & Warschauer, M. (2022). Fantastic questions and where to find them: FairytaleQA – An authentic dataset for narrative comprehension. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 447–460). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.acl-long.34

Xue, L., Constant, N., Roberts, A., Kale, N., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). *MT5: A massively multilingual pre-trained text-to-text transformer* (Version 3). arXiv. http://dx.doi.org/10.48550/arXiv.2010.11934

Zhang, C. (2023). *Automatic generation of multiple-choice questions* (Version 1). arXiv. http://dx.doi.org/10.48550/ARXIV.2303.14576

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). *BERTScore: Evaluating text generation with BERT* (Version 3). arXiv. http://dx.doi.org/10.48550/ARXIV.1904.09675

Zmitrovich, D., Abramov, A., Kalmykov, A., Tikhonova, M., Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., Kadulin, V., Markov, S., Shavrina, T., Mikhailov, V., & Fenogenova, A. (2024). *A family of pretrained transformer language models for Russian*. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 507–524). ELRA Language Resource Association. http://dx.doi.org/10.48550/arXiv.2309.10931

APPENDIX A

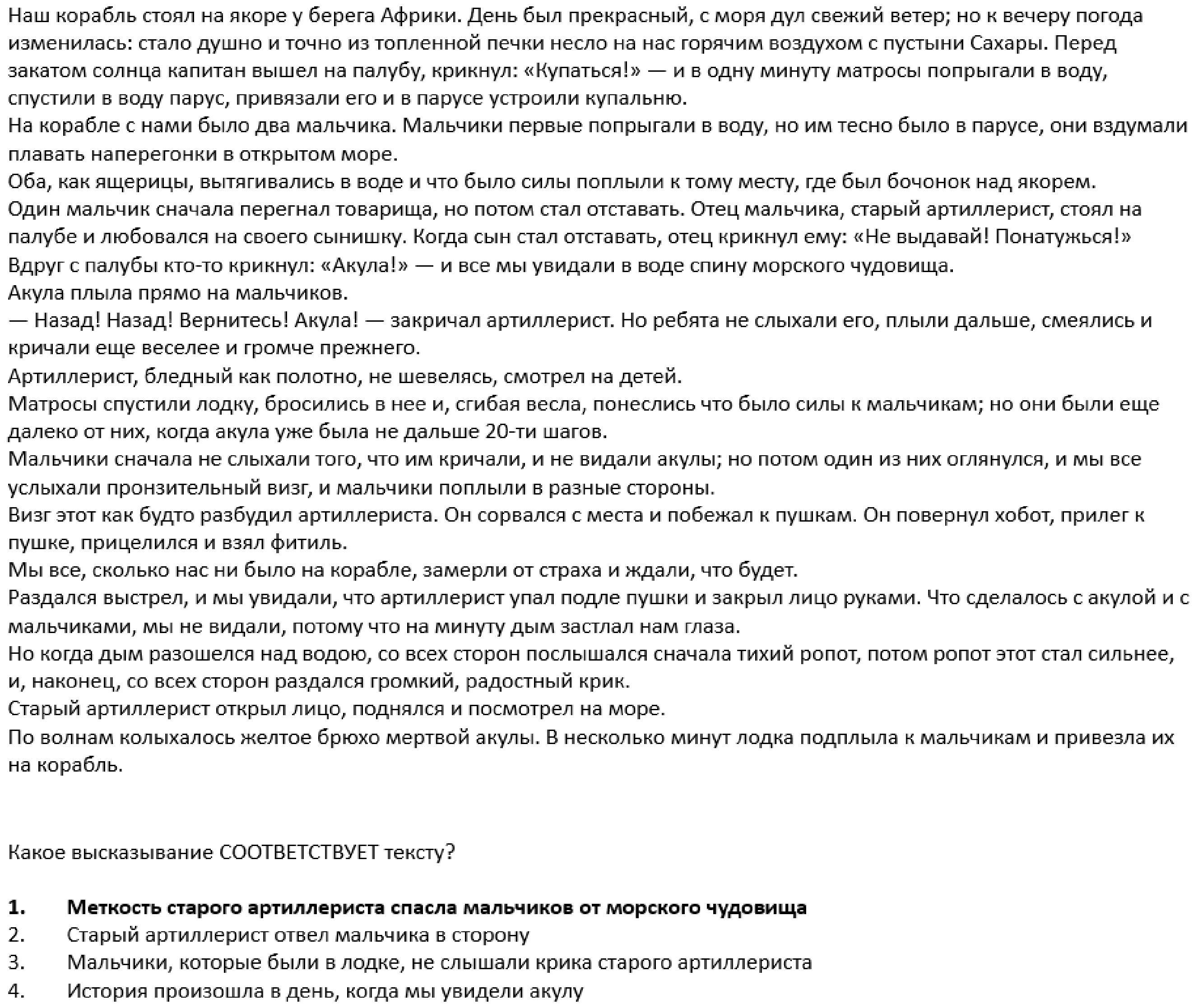
Example of model generation

To better illustrate the performance of our best model, RuT5-RACE-TF, we will analyse an example of distractors generated by our best model, RuT5-RACE-TF, on an arbitrary Russian text – "The Shark", a children's short story by Leo Tolstoy, taken from Russian WikiSource. The story tells us about how a ship's cannoneer saved two boys who went swimming in the open sea from a shark by shooting it with a cannon. The right answer to the "Which statement is TRUE according to the text?" question ("Old cannoneer's marksmanship saved the boys from the sea monster") was created manually.

From the Figure 1 it can be seen that the options 2 and 3 can indeed work as distractors as they contradict the text as they state facts that are not present in the story ("Old cannoneer took the boy aside"; "The boys who were in the boat did not hear the old cannoneer's cry"). However, option 4 ("The story happened the day we saw a shark") is, although being very trivial, true to the text and therefore cannot serve as a distractor in this context. It is worth noting that in this example the language of the generated distractors is consistent and does not break the rules of Russian grammar.

Figure 1

Result of Ru-T5-RACE distractor generation on a Leo Tolstoy's chilfren's story



Наш корабль стоял на якоре у берега Африки. День был прекрасный, с моря дул свежий ветер; но к вечеру погода изменилась: стало душно и точно из топленной печки несло на нас горячим воздухом с пустыни Сахары. Перед закатом солнца капитан вышел на палубу, крикнул: «Купаться!» — и в одну минуту матросы попрыгали в воду, спустили в воду парус, привязали его и в парусе устроили купальню.

На корабле с нами было два мальчика. Мальчики первые попрыгали в воду, но им тесно было в парусе, они вздумали плавать наперегонки в открытом море.

Оба, как ящерицы, вытягивались в воде и что было силы поплыли к тому месту, где был бочонок над якорем.

Один мальчик сначала перегнал товарища, но потом стал отставать. Отец мальчика, старый артиллерист, стоял на палубе и любовался на своего сынишку. Когда сын стал отставать, отец крикнул ему: «Не выдавай! Понатужься!»

Вдруг с палубы кто-то крикнул: «Акула!» — и все мы увидали в воде спину морского чудовища.

Акула плыла прямо на мальчиков.

— Назад! Назад! Вернитесь! Акула! — закричал артиллерист. Но ребята не слыхали его, плыли дальше, смеялись и кричали еще веселее и громче прежнего.

Артиллерист, бледный как полотно, не шевелясь, смотрел на детей.

Матросы спустили лодку, бросились в нее и, сгибая весла, понеслись что было силы к мальчикам; но они были еще далеко от них, когда акула уже была не дальше 20-ти шагов.

Мальчики сначала не слыхали того, что им кричали, и не видали акулы; но потом один из них оглянулся, и мы все услыхали пронзительный визг, и мальчики поплыли в разные стороны.

Визг этот как будто разбудил артиллериста. Он сорвался с места и побежал к пушкам. Он повернул хобот, прилег к пушке, прицелился и взял фитиль.

Мы все, сколько нас ни было на корабле, замерли от страха и ждали, что будет.

Раздался выстрел, и мы увидали, что артиллерист упал подле пушки и закрыл лицо руками. Что сделалось с акулой и с мальчиками, мы не видали, потому что на минуту дым застлал нам глаза.

Но когда дым разошелся над водою, со всех сторон послышался сначала тихий ропот, потом ропот этот стал сильнее, и, наконец, со всех сторон раздался громкий, радостный крик.

Старый артиллерист открыл лицо, поднялся и посмотрел на море.

По волнам колыхалось желтое брюхо мертвой акулы. В несколько минут лодка подплыла к мальчикам и привезла их на корабль.

Какое высказывание СООТВЕТСТВУЕТ тексту?

1. **Меткость старого артиллериста спасла мальчиков от морского чудовища**
2. Старый артиллерист отвел мальчика в сторону
3. Мальчики, которые были в лодке, не слышали крика старого артиллериста
4. История произошла в день, когда мы увидели акулу

Note. Manually crafted right answer is given in **bold***.*

APPENDIX B

The source code and data files for this paper are available at the online repository: https://github.com/nicklogin/Ru-RC-DG