# A Russian Question Answering System for Inclusive Education

Victoria Firsanova

December 2020

**Abstract**

The project report contains a description of a closed domain model for question answering in Russian built with transfer learning techniques. The data and the solution's code are distributed through GitHub in the following repository: `https://github.com/vifirsanova/nlp-huawei-project`.

## 1 Introduction

The project aims to develop two base model for a question answering system for inclusive education in Russian. The study highlights two approaches to building QA systems: retrieval-based and generative. Two models described in the project allows discovering the weaknesses and strengths of both, using BERT [Devlin et al., 2019] for the implementation of retrieval based QA model, and GPT-2 [Radford et al., 2019] for the implementation of a generative QA model.

The motivation for conducting the study is that inclusion of people with special needs becomes more and more widespread in Russia, although the problem of lack of information and false information spreading is acute and might cause misunderstandings between members of inclusive organizations.

The result of the study is a model that will base a chatbot for inclusive education, an AI-powered tool. Furthermore, the study describes building a non-English NLP tool, supporting the development of non-English language modeling. The models in the study were trained on a custom closed-domain dataset for Russian collected by its author and focuses on an acute social issue, which makes this study on dialogue systems socially significant and unique.

### 1.1 Team

The author of the project is **Victoria Firsanova**, Saint Petersburg State University master student in computational linguistics.

## 2 Related Work

The study focuses on building a question answering agent. Question answering agents along with task-oriented dialogue agents and social bots form the Conversational AI field [Gao et al., 2018]. ConvAI applications might solve problems of human-machine interaction or some narrow closed-domain tasks. Apart from task-dependent challenges, conversational agents, or dialogue systems, should generate coherent, grammatically correct sentences without redundant repetitions. These are intuitive dialogue capabilities, which enables reasoning, logic inference, and associative properties [Vassallo et al., 2010]. Such elements and features of conversational agents still need further investigation.

DialoGPT [Zhang et al., 2020] is an example of a conversational system based on a generative pre-trained transformer GPT-2 [Radford et al., 2019]. The model pre-trained on Reddit data can be fine-tuned for building neural dialogue systems. The training data for AI-powered conversational agents is another issue. Such data should represent or, at least, resemble lines of dialogue or question-answer queries. Apart from using the data from social media, like Reddit or Twitter, one can adapt TV shows transcripts for the modeling task [Li et al., 2016], or collect the dataset manually with some additional manipulations like paraphrasing [Zhang et al., 2018].

AI-powered dialogue agents, hypothetically, might be used as a tool for providing psychological support. Although such systems cannot and should not replace professional therapists, they might become helpful in situations when some assistance is needed immediately, and other sources are not available [Ta et al., 2020]. Replika is one example of such systems. This generative GPT-3 based chatbot uses various strategies to provide a user with emotional support in everyday conversations. For example, Replika can generate compliments to form assertions or satisfy users' ones [Hakim et al., 2019].

## 3 Model Description

For the experiments, two Transformer based models were chosen, BERT (Bidirectional Encoder Representations from Transformers [Devlin et al., 2019]) and GPT-2 (Generative Pre-trained Transformer [Radford et al., 2019]). The key feature of Transformer architecture described in [Vaswani et al., 2017] is that it is based solely on attention mechanisms with a fully connected feed-forward network. The position-wise feed-forward network in the Transformer takes a vector $x$ and passes it through the matrices *W1* and *W2* and bias vectors *b1* and *b2*:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

The unit of a multi-head self-attention mechanism $A$ in the Transformer takes key-value pairs ($K$, $V$) as input. The keys are of dimension $d_k$, and the values are of dimension $d_v$. For the computation of the scaled dot-product attention described in [Vaswani et al., 2017], the attention function is multiplied by a matrix $Q$, and a softmax function is applied:

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

During the experiments, two different approaches will be applied for question answering, retrieval-based and generative. For the implementation of a retrieval based question answering, BERT will be used. BERT is a model pre-trained for the masked language modeling (MLM) task. In this task, a model analyses the context surrounding masked tokens to predict them. The knowledge acquired through MLM task solving can be successfully transferred to tasks connected to information retrieval and extraction [Shazeer et al., 2020]. For example, BERT [Devlin et al., 2019] showed significant improvements in questions answering with SQuAD v1.1 [Rajpurkar et al., 2016] and SQuAD v2.0 [Rajpurkar et al., 2018]. For the implementation of a generative question answering, GPT-2 will be used. GPT-2 is a model for the traditional language modeling, which analyzes only left-to-right context to predict the next token in a given sequence. Nevertheless, the model shows high zero-shot performance on various tasks, including question answering [Radford et al., 2019]. Fig. 1 represents a conceptual comparison between the two models.

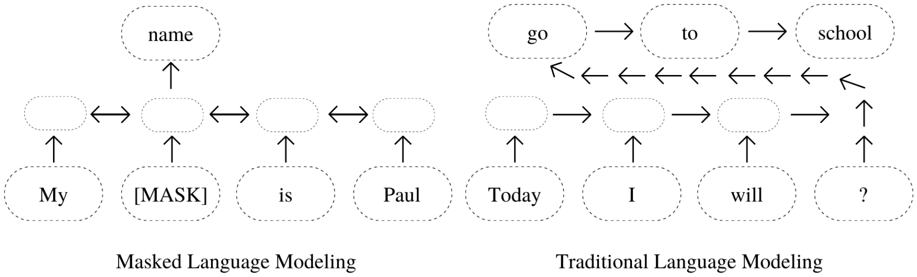


Figure 1: Masked and traditional language modeling concepts.

# 4 Dataset

The dataset used for the model training is presented in [Firsanova, 2020]. The name of the dataset is Autism Spectrum Disorder and Asperger Syndrome Question Answering Dataset, it was collected for the current study by the author. It is available online for research purposes as an open source[1]. The dataset is based on the publications from the informational online-source on Autism spectrum disorder (ASD) and Asperger syndrome[2]. The usage of the data from

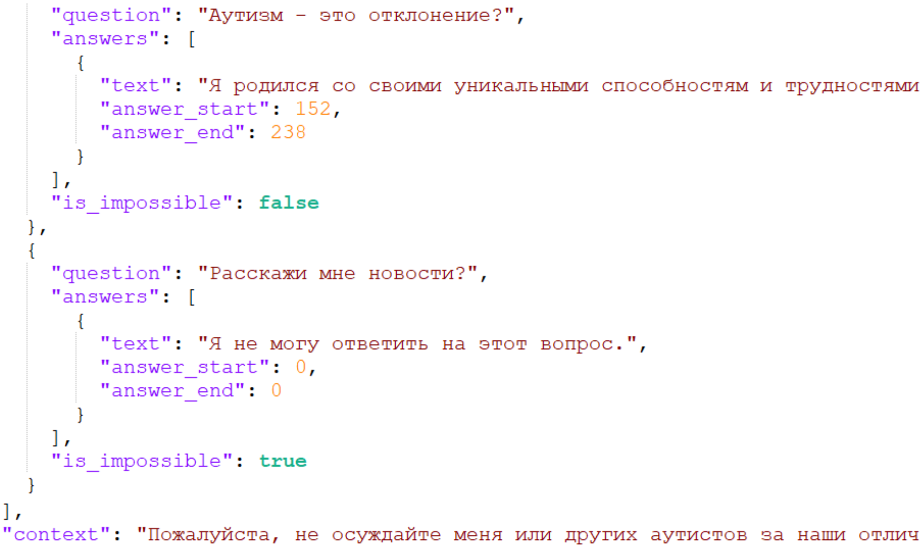
[1] The dataset can be downloaded for free by doi.org/10.6084/m9.figshare.13295831.

[2] The web-address of the informational source is aspergers.ru.

the website is agreed. The data from the website was collected with an HTML parser built with Beautiful Soup 4 [Richardson, 2020] on Python.

The dataset structure was inspired by Stanford Question Answering Dataset [Rajpurkar et al., 2016]. This is a reading comprehension dataset that includes context paragraphs, questions to them, and answers, which represent spans of the corresponding paragraphs. The dataset used in the study contains a collection of paragraphs extracted from the informational website on ASD and Asperger syndrome, a set of questions posed to each, and a set of answers. The set of answers contains textual citations from the paragraphs giving the information on posed questions. The texts are provided with tags of start and end positions of corresponding spans as integer indexes representing symbols' positions in the paragraphs.

Like in SQuAD 2.0 [Rajpurkar et al., 2018], some of the questions from the used dataset are unanswerable, however, the purpose of those unanswerable units is different. In SQuAD 2.0 paper, the authors offer to include some unanswerable but relevant questions to robust the question answering, although the used dataset contains 5% of unanswerable and irrelevant questions to allow QA models distinguishing user's questions unrelated to the topic of ASD and inclusion, which are better to ignore. Fig. 2 presents a dataset sample. The dataset was split into train, validation and test sets before the model training. Tab. 1 shows the dataset statistics.



```
    "question": "Аутизм - это отклонение?",
    "answers": [
      {
        "text": "Я родился со своими уникальными способностям и трудностями
        "answer_start": 152,
        "answer_end": 238
      }
    ],
    "is_impossible": false
  },
  {
    "question": "Расскажи мне новости?",
    "answers": [
      {
        "text": "Я не могу ответить на этот вопрос.",
        "answer_start": 0,
        "answer_end": 0
      }
    ],
    "is_impossible": true
  }
],
"context": "Пожалуйста, не осуждайте меня или других аутистов за наши отлич
```

Figure 2: A dataset sample.

|  | Train | Valid | Test |
|---|---|---|---|
| QA pairs | 523 | 126 | 107 |
| Tokens (word level) | 12,264 | 3,694 | 2,936 |
| Vocabulary size |  | 30,522 |  |
| Out of Vocab rate |  | 4.47% |  |

Table 1: Statistics of the Autism Spectrum Disorder and Asperger Syndrome Question Answering Dataset. The out of vocabulary tokens have been replaced by an $\langle unk \rangle$ token.

# 5 Experiments

## 5.1 Metrics

For the question answering evaluation F-Score was used as proposed in [Gillard et al., 2006]. It is calculated as the harmonic mean of the precision $P$ and recall $R$, where precision is the fraction of relevant (true positive) model answers among the retrieved (true positive and false positive), and recall is the fraction of the total amount of relevant model (true positive) answers among all the samples (true positive and false negative):

$$F = \frac{2PR}{P+R}$$

$$P = \frac{tp}{tp+fp}$$

$$R = \frac{tp}{tp+fn}$$

In question answering task, true positive answers are the tokens *shared* between the correct (or *gold*) tokens and the all *predicted* tokens, false positives are the *predicted* tokens absent in the correct *gold*) answer, and false negatives are the tokens from the correct (*gold*) answer absent in the *predicted*.[3] With this correction, the formula is the following:

$$F = \frac{2PR}{P+R}$$

$$P = \frac{shared}{shared + (predicted - shared)}$$

$$R = \frac{shared}{shared + (gold - shared)},$$

[3] See Evaluation Script on SQuAD web-page.

### 5.2 Experiment Setup

The model training was performed in Google Colaboratory with the Tesla T4 GPU. The code was implemented in Python. The data from the dataset was shuffled and split into train (65% of the data), validation (20% of the data) and test (15% of the data) sets with *train_ test_ split* method from the Scikit-learn library. The PyTorch library was used to train the models.

The configuration of the BERT based model is the following. The activation function is GELU [Hendrycks and Gimpel, 2016]. The maximum number of tokens in an input sequence is 512. The size of the encoder layers is 768. The size of the feed-forward layer is 3,072. The number of attention heads is 12. The number of hidden layers is 12. The dropout ratio is 0.1. The learning rate is 5e-5. The model ran for 10 epochs.

The configuration of the GPT-2 based model is the following. The activation function is GELU [Hendrycks and Gimpel, 2016]. The maximum number of tokens in an input sequence is 512. The number of embeddings is 1,280. The number of attention heads is 20. The number of hidden layers is 36. The temperature (the value that controls output randomness) is 0.7. The top K (the value that controls the output diversity) is 40. The model ran for 3000 steps.

### 5.3 Baselines

For the experiments, two models were built on the basis of the multilingual base cased BERT[4] and GPT-2 with 774 million parameters[5] pre-trained models. For the implementation of BERT for QA, the base model was applied for question answering using HuggingFace Transformers[6] and fine-tuned on a custom SQuAD-like dataset[7]. For the implementation of GPT-2 for QA, the base model was fine-tuned to take questions from the custom dataset as prefixes and generate answers to them using the Gpt-2-simple package[8]. The code is available on the author's GitHub repository[9].

## 6 Results

Tab. 2 shows the results of both models obtained on the training data. The GPT-2 based model showed higher results. However, the model outputs contain some inconsistent, irrelevant answers generated by the language model, whereas the BERT based model showed more coherent and full answers, but the recall of the model was significantly lower because the training data is yet to be supplemented. Results for both models' inference are provided in Tab. 3.

[4] https://github.com/google-research/bert/blob/master/multilingual.md

[5] https://github.com/openai/gpt-2

[6] https://github.com/huggingface/transformers

[7] https://doi.org/10.6084/m9.figshare.13295831

[8] https://github.com/minimaxir/gpt-2-simple

[9] https://github.com/vifirsanova/nlp-huawei-project

| **Base model** | **F-Score** |
|---|---|
| Multilingual BERT | 0.55 |
| 774M GPT-2 | 0.63 |

Table 2: Results obtained on the test data.

| **BERT based model** |
|---|
| **Q:** Есть ли какие-то техники, чтобы научить ребенка с РАС вести себя в социальном мире? |
| **A:** Поговорите со специалистами, работающими с ребёнком, о каких-либо программах, которые могут ему помочь. |
| **GPT-2 based model** |
| **Q:** У детей-аутистов могут быть проблемы с одноклассниками? |
| **A:** У детей с расстройством аутистического спектра (РАС) высок риск подвергнуться травле, особенно в обычных школах. |

Table 3: Output samples.

# 7 Conclusion

During the study, two approaches, retrieval-based and generative, were performed to build a closed-domain question answering model. The results of the study include an open-source dataset for the question answering, and two models, which will further form the basis of a chatbot for inclusive education. Despite the significant differences of the approaches, retrieval-based and generative, after the analysis of the study results several common conclusions on making the models more robust were made. Firstly, the dataset needs to be supplemented and expanded. That would allow the models to give more full and diverse answers. Secondly, the structure of the training data can be edited by adding some paraphrased answers, so that one question might have several possible answers. Finally, building additional algorithms for the data preprocessing and informational retrieval might make the model more robust.

# References

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Firsanova, 2020] Firsanova, V. (2020). Autism spectrum disorder and asperger syndrome question answering dataset 1.0. https://doi.org/10.6084/m9.figshare.13295831.

[Gao et al., 2018] Gao, J., Galley, M., and Li, L. (2018). Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, SIGIR '18, page 1371–1374, New York, NY, USA. Association for Computing Machinery.

[Gillard et al., 2006] Gillard, L., Bellot, P., and El-Bèze, M. (2006). Question answering evaluation survey. In *LREC*.

[Hakim et al., 2019] Hakim, F., Indrayani, L., and Amalia, R. (2019). A dialogic analysis of compliment strategies employed by replika chatbot.

[Hendrycks and Gimpel, 2016] Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv: Learning*.

[Li et al., 2016] Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., and Dolan, B. (2016). A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[Rajpurkar et al., 2018] Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

[Richardson, 2020] Richardson, L. (2020). Beautiful soup. `https://www.crummy.com/software/BeautifulSoup/`.

[Shazeer et al., 2020] Shazeer, N., Lan, Z., Cheng, Y., Ding, N., and Hou, L. (2020). Talking-heads attention.

[Ta et al., 2020] Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., and Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *J Med Internet Res*, 22(3):e16235.

[Vassallo et al., 2010] Vassallo, G., Pilato, G., Augello, A., and Gaglio, S. (2010). Phrase coherence in conceptual spaces for conversational agents. In *Semantic Computing*, IEEE, pages 357–371.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017).

[Zhang et al., 2018] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

[Zhang et al., 2020] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.