Question Answering Systems and Inclusion: Pros and Cons

Victoria Firsanova[a]

[a] *Saint Petersburg State University, 7-9 Universitetskaya Emb., St Petersburg, 199034, Russian Federation*

Abstract

In the inclusion, automated QA might become an effective tool allowing, for example, to ask questions about the interaction between neurotypical and atypical people anonymously and get reliable information immediately. However, the controllability of such systems is challenging. Before the integration of QA in the inclusion, a research is required to prevent the generation of misleading and false answers, and verify that a system is safe and does not misrepresent or alter the information. Although the problem of data misrepresentation is not new, the approach presented in the paper is novel, because it highlights a particular NLP application in the field of social policy and healthcare. The study focuses on extractive and generative QA models based on BERT and GPT-2 pre-trained Transformers, fine-tuned on a Russian dataset for the inclusion of people with autism spectrum disorder.

Keywords

Natural Language Processing, Question Answering, Information Extraction, BERT, GPT-2

1. Introduction

AI-powered question answering systems might find their practical application in the medical and social domain. Question answering (QA) systems take questions in natural language as input and provide (for example, by text generation or data extraction) corresponding answers as outputs. In the healthcare field, automated QA might benefit both patients and medical practitioners by providing immediate access to required extracts from medical knowledge bases. Closed-domain QA can be used as an additional source of information for volunteers or members of a social institution by providing immediate access to the internal information of a certain organization. Based on a rich and reliable database, QA systems can be used as an additional educational source in the processes of gamification and digitalization at schools or higher education institutions.

The idea of the paper came after the first trial of building an informational question answering system. The system aims to give information about inclusive education in the Russian language. The project supports the inclusion of people with autism spectrum disorder (ASD). In the inclusion, automated QA might become an efficient tool. Limited knowledge of the inclusive education process and lack of awareness about the people with special needs raise anxiety among both neurologically typical members of the inclusion and members with developmental characteristics. The information awareness would help to dispel misconceptions and prevent conflicts in classes.

AI-powered QA is a way to provide information fast and playfully. Children and young adults are not likely to read and analyze extensive texts to find the needed information. The ability to ask any question in a free form would not require a high concentration and save a lot of time, making the inclusion more comfortable. Moreover, members of the inclusion would have an opportunity to ask frequent and uncomfortable questions anonymously. For example, if a student needs a tip for communication with a classmate with ASD and is too shy to ask a friend or teacher, or there is no teacher or tutor around, the student will have a chance to ask a QA bot and get reliable information immediately.

IMS 2021 - International Conference "Internet and Modern Society", June 24-26, 2021, St. Petersburg, Russia
EMAIL: vifirsanova@gmail.com
ORCID: 0000-0002-8474-0262

CEUR Workshop Proceedings (CEUR-WS.org)

However, the integration of QA systems into inclusive organizations requires confidence that the built applications are safe. Safe applications involve language models that do not generate false information or mislead. Such models should be bias-resistant. They should interact with a user in a friendly way generating coherent and understandable texts, although they should not entertain a user.

One of the challenges of neural approaches towards natural language processing is their controllability. High scores of perplexity imply coherent text generation but do not exclude the generation of misleading or false responses. Thus, the outputs of uncontrollable models might be generic or factually incorrect, whereas, for neural conversation models, semantic control ensuring is essential [1]. The semantic control provides dialogue specification, ensures model flexibility, and develops the model knowledge grounding [2].

The paper aims to highlight the linguistic features of question answering systems' responses and analyze their strengths and weaknesses from the users' perspective. The study will lead to a broader understanding of the capabilities of the practical efficiency of AI-powered QA. The research focuses on the underlying causes of dialogue system errors and will contribute to the further development of conversational AI.

As a research method, it was chosen to build two question answering systems using two different approaches. The first approach is extractive. This approach is widespread in the reading comprehension task, one of the problems of natural language understanding (NLU) [3]. In the extraction based QA, the answer to a user's question is a specific piece of information from a given database. The answer can be presented in the form of a single word, sentence, or paragraph [4]. The second approach is generative. Generative models learn to exploit correlations in the data by memorizing the information [5]. This can also be a result of zero-shot learning within the ability of a model to learn some generalizations during the training across tasks [6]. Zero-shot learning is a learning method allowing one to solve a task without training on examples of that task. The method allows a model process previously non observed classes by associating knowledge gained during the pre-training on data representing other classes.

For the implementation of two approaches, self-attention Transformer network architecture models were applied. The generative approach was implemented with the Transformer decoder based model GPT-2 [7]. The extractive one was implemented with the Transformer encoder based model BERT [8]. Both models were fine-tuned on a custom question answering dataset. GPT-2 was trained as a traditional language model, which uses zero-shot learning to memorize the structure of a QA dataset and generate answers. BERT was fine-tuned for the downstream question answering task. In recent years, the models based on Transformer architecture showed high efficiency on many NLP tasks, including question answering, due to the self-attention mechanism, which allows attending the focus to specific words and establishing sequence contexts. This allows analyzing texts while training more accurately, memorizing longer sequences, and transferring the gained knowledge to new tasks.

One of the issues of modern NLP is that most of the models are evaluated on the English data. However, the English language is rather weakly inflected. That is not typical for most of the Indo-European languages. Thus, high model evaluation scores might be reached without taking into consideration the facts about linguistic features of other languages. The Russian language, for example, is fusional. That means that the morphological features are crucial for the understanding of the meaning of a sentence. Spans, which represent the answers in extractive QA, are direct citations of the text. Thus, if the wording of the question is not equal to the wording of the context, the rules of conjugation and declension might be broken.

Although the problem of data misrepresentation is not new, the approach presented in the paper is novel, because it highlights a particular NLP application in the field of social policy and healthcare. The development of two QA models and their analysis presented in the paper should shed light on the problems of building social-oriented conversational AI systems. That might help to predict possible issues and solve them before they happen.

2. Related Work

The study focuses on building a conversational AI (ConvAI) system. According to Gao et al. [9], conversational systems usually solve three fundamental tasks: question answering, task-oriented dialogues, and chatbots. Conversational systems aim to imitate human behavior. One of the ways to reach this is to use language patterns that would ensure dialogue credibility. The credibility might be established when human-AI dialogue lines would be considered close enough to real-life human interaction according to some objective criteria. Among such objective criteria, the linguistic features of the text can be considered. For example, dialogue systems should learn to generate coherent, grammatically correct utterances without redundant lexical repetitions. Those elements ensure intuitive dialogue capabilities, such as reasoning, logic inference, and associative properties [10].

The tasks of ConvAI vary, although there are common fundamental tasks that form the basis of the research field. One of the foundational problems of conversational AI is task completion. While solving this type of problem, the dialogue agent should be capable of recognizing the user's needs. After the task recognition, the agent should be able to accomplish it and give an appropriate response in the natural language if necessary. The range of tasks varies from the restaurant and hotel reservations to the meeting scheduling and business planning [9].

Another foundational task is social chat. Social chatbots are designed for human-AI communication, which imitates everyday human interaction. The development of such systems may have the goal of modeling human conversations to pass the Turing test [9]. Apart from that, social chatbots might give recommendations and provide psychological support. Although such systems cannot and should not replace professional therapists, they might become helpful in situations when assistance is needed instantly, and other sources of support are not available [11].

The current study focuses on question answering systems. Question answering is another foundational ConvAI task [9]. QA agents aim to provide a user brief answers to his or her request on a certain topic. The answers of such dialogue systems can be based on knowledge bases, such as text collections, web sources, sets of structured or unstructured data on narrow subjects, for example, on a certain field of medicine.

The spectrum of QA-world represents such systems as Knowledge-Based QA agents, or KB-QA, text-QA, and Machine Reading Comprehension (MRC) models. Question answering systems that use natural language as a part of their interface are more convenient to use than similar systems not based on NLP algorithms. For example, KB-QA agents are often compared to SQL-like systems. KB-QA are considered to be more user-friendly than their predecessors due to their interactiveness [9]. The flexibility of QA systems is reflected, for example, in text-QA agents integrated with mobile virtual assistants. Such systems usually have web access. That allows them to provide answers to simple questions faster and more convenient than traditional search engines [9].

Neural MRC is another important QA related model. The task of MRC is to generate an answer to a user's question posed on a given text. The task aims to evaluate the machine capability of natural language understanding. Theoretically, the ability of a machine to make some conclusions after the reading, for example, to answer text-related questions might lead to a breakthrough in human-AI interaction. MRC might have a broader practical application. For example, MRC algorithms can be integrated into search engines allowing them to give short answers to a user's query instead of providing an unstructured list of possible web-pages with relevant information [12]. In the current study, an MRC algorithm would be used as a basis for the informational extractive QA model.

One of the examples of reading comprehension datasets is Stanford Question Answering Dataset (SQuAD) [13]. SQuAD has the following features. Firstly, the authors and creators of SQuAD paid attention to answer types. They have allocated several categories including, for example, dates, persons, locations, and others. Secondly, the developmental SQuAD set was provided with reasoning labels. For example, they have highlighted such types of reasoning as a lexical and syntactic variation. Besides, some actions were made to ensure that the dataset is diverse. For example, the answers were categorized into numerical and non-numerical ones by means of constituency parsing and POS-tagging. The non-numerical answers were also split into narrower categories, such as persons and locations by using Named Entity Recognition (NER).

SQuAD v2.0 [14] has several differences from its predecessor SQuAD v1.1. The renewed dataset can evaluate the model's capability to ignore the questions that do not have an explicit answer in a given reading passage. The authors of SQuAD v2.0 offer to include some unanswerable questions in their dataset, although these unanswerable questions should be relevant to the corresponding reading passage and have a plausible answer in the text. That complicates the reading comprehension task by inviting the model to learn how to distinguish answerable questions from unanswerable ones and thus achieve higher accuracy in its analysis.

3. Data

The models built for the experiments were trained on a custom question answering dataset. The dataset was collected by the author of the paper. It is available online (see Online Resources). The dataset is called Autism Spectrum Disorder Question Answering (ASD QA). ASD QA is based on the data from the informational websites about autism spectrum disorder and Asperger syndrome in children and adults, inclusion and support of people with Asperger syndrome and ASD, their health, and communication with neurologically typical people. ASD QA is a long-term project. For the year 2021, it has the status of active, which means that the dataset is in the process of collection and development.

The data for the ASD QA was collected from the informational website about ASD and Asperger syndrome http://aspergers.ru/ with the agreement of the website administration. The data from the website represent a collection of articles and texts of related genres (blog entries, messages to readers, etc.). The texts were created by neurologically typical people and people with Asperger's syndrome or ASD, created in Russian or translated into Russian from foreign languages. The authors are native or fluent speakers of the Russian language.

According to the website categories, the publications from the informational source cover the following topics: basic information about Asperger's syndrome and ASD, diagnostics of Asperger's syndrome and ASD, symptomatic of Asperger's syndrome and ASD, problems of people with Asperger's syndrome and ASD, social skills and communication issues of people Asperger's syndrome and ASD, recommendations for parents of children with Asperger's syndrome and ASD, education, and training, work and employment, relationships, love and family, discussions about ASD, myths and facts about ASD, etc.

Figure 1 presents a topical data distribution in the ASD QA dataset as at May 2021. The topics were extracted from the website http://aspergers.ru/ which served as a source for the ASD QA dataset. Each article on the website has one or several tags indicating its topics. After we had extracted those tags we built a bar chart showing the number of articles covering each topic. One article could cover several topics.

The data was collected with an HTML parser built with Beautiful Soup 4 [15] on Python. Beautiful Soup is a library that is often used for web data extraction. For the data extraction from the chosen for the dataset collection website, the following steps were made. Firstly, HTML content from pages of the website was obtained with the "get" method from the "Requests" Python library. Secondly, the text data was analyzed and parsed with "findAll" and "find" basic Beautiful Soup methods. Finally, the extracted texts were saved as text data for further processing and dataset development.

After the data was collected, it was important to structurize it. Insofar as the dataset was being designed for the question answering models training and evaluation, it was decided to develop it like a reading comprehension one. In contrast with traditional question answering datasets, which contain only sets of QA-pairs, the format of reading comprehension datasets also implies the presence of reading passages. Reading passages are sets of sentences or paragraphs, which an MRC model should learn to "understand" or answer the questions about the information contained in each passage.

Another important aspect is the question acquisition. The reading passages were split into sentences separated by periods, ellipses, question or exclamation marks. We strove to ask one or several questions to each sentence, but some of the text pieces (for example, some introductory remarks or personal reflections) did not contain significant information, so we had to ignore them. We have asked 2-3 questions on average to each sentence containing significant information, using different types of questions. We have chosen the type of a question based on the structure of its possible answer (excerpt from a reading passage). For example, we have asked closed questions to sentences containing affirmative or negative constructions, and we have asked open questions to sentences containing factual information. This was done manually because the ASD QA dataset is being designed for ”safety-first" systems which require the best available training data.

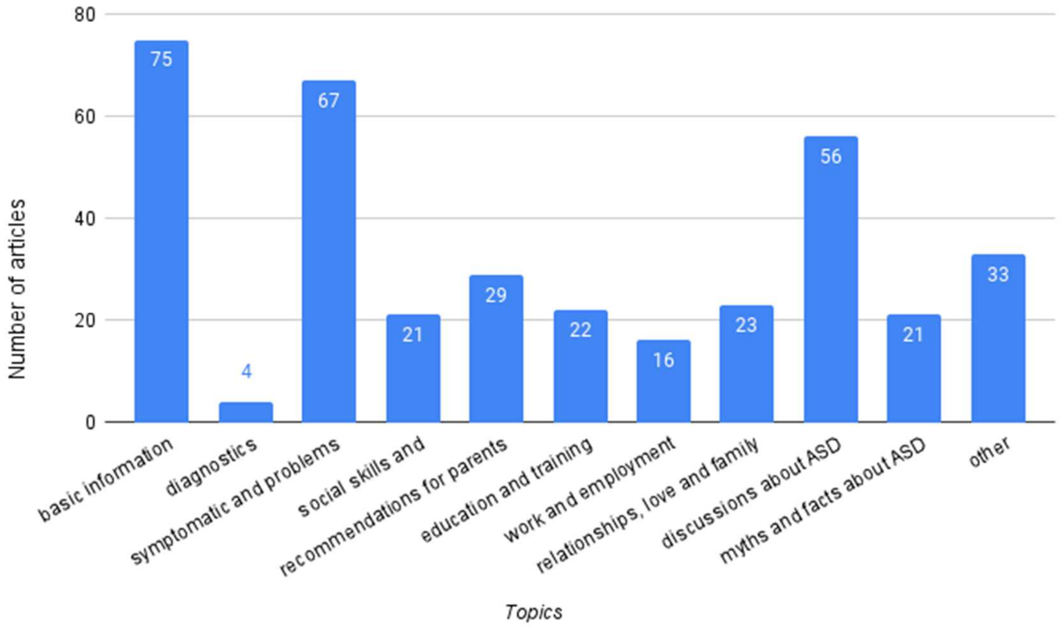


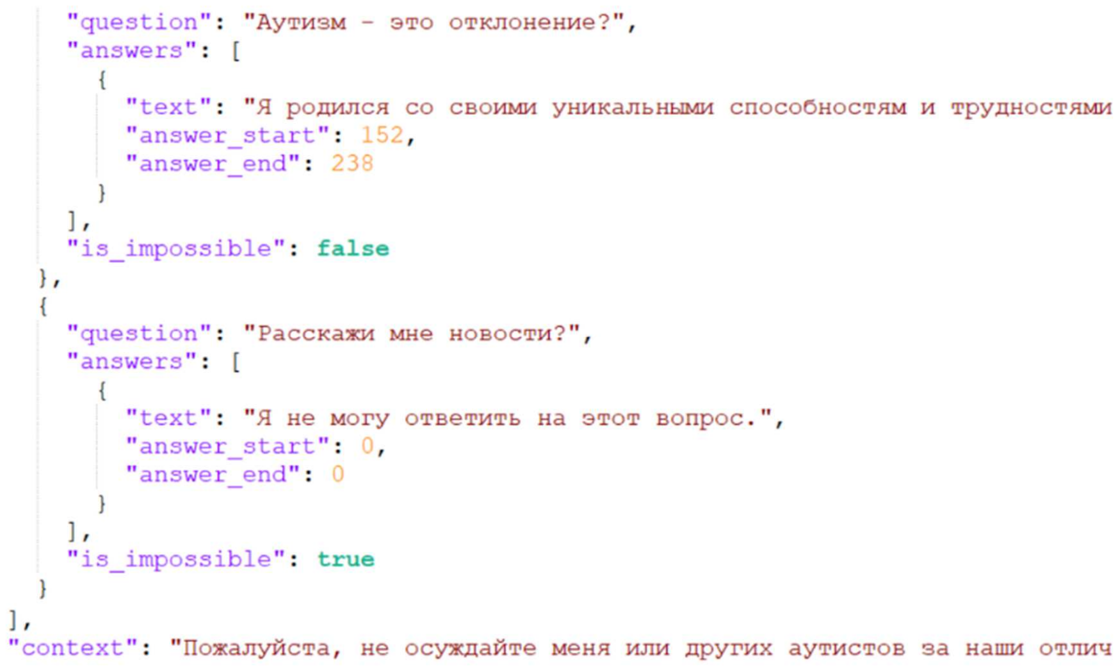
Figure 1: ASD QA topical data distribution. May 2021

Figure 2 presents an ASD QA dataset sample. The dataset structure was inspired by SQuAD v2.0 [14]. During the development of the ASD QA dataset, it was decided to provide it with several unanswerable questions too. However, after the first training trials on a new dataset, it was noticed that the aim of unanswerable questions in ASD QA should differ from the aim of those in SQuAD v2.0.

During the ASD QA development, the dataset was provided with 5% of unanswerable questions on the principle of SQuAD v2.0. Unanswerable questions in the ASD QA dataset are deliberately irrelevant, which means that there are no answers to these questions in the reading passages, and also there are no answers in the dataset at all. Among such questions, there are ones that aim to set an entertaining tone in a human-AI dialogue. For example, some questions ask a system to tell a joke or a fairy tale, some are about artificial intelligence misconceptions, some contain complaints about boredom, etc. Presumably, users can ask such questions for entertainment purposes. However, the systems, for training and evaluation of which the ASD QA dataset is developed, should avoid such questions. These systems aim to consult and give accurate information. They do not have an aim to entertain a user.

For the unanswerable questions, the system includes a label ”is unanswerable”. The JSON object containing the ASD QA data includes the label with a Boolean for each QA-pair. Thus, if a question has a piece of information in a corresponding reading passage, the label ”is unanswerable” is False. Otherwise, the label is True. For example, in Figure 3 two QA-pairs are presented. The question of the first pair is translated from Russian into English as “Is autism a deviation?”. This question has an answer in a corresponding reading passage, which is marked as a “context” in the dataset. The label “is unanswerable” is False. Labels “answer start” and “answer end” mark the answer span, serial numbers of the first and last characters position of answers in the passage.

The question of the second pair is translated from Russian into English as “Tell me the news?”. This question has no answer in the dataset reading passages, it is added in the dataset to complicate the task. The label “is unanswerable” for this question is True. The values of “answer start” and “answer end” are both 0. Despite the fact that the question is unanswerable and irrelevant, the dataset is provided with a plausible answer, which is translated from Russian into English as “I cannot answer this question”. This makes the dataset also suitable for the training of generative QA models. Such models instead of answering irrelevant questions can learn to generate this phrase.



```
      "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: An ASD QA dataset sample

Table 1 presents the ASD QA data statistics in the context of the paper research. For the implementation of the experiments, the dataset was split with the “train_test_split” method from the Scikit-learn library [16] for machine learning in Python. The set of 756 QA-pairs (including the corresponding context and the metadata: spans, and labels of answerableness) was randomly shuffled and split into a train set including 69% of the data (523 QA-pairs), a validation set including 17% of the data (126 QA-pairs), and a test set including 14% of the data (107 QA-pairs). The size of the vocabulary created and used for the question answering models’ training was 30 522 tokens on a word level. According to the frequency vocabulary built during the pre-processing, 4.47% Out-of-Vocabulary (OOV) tokens were replaced by an (meaning “unknown”) token. During the data processing, each OOV-token was split into sub-words greedily using byte pair encoding (consecutive bytes are steadily replaced with a new byte). This allows allocating frequently used pieces of words, such as prefixes and suffixes, as well as roots, and conducting a lossless analysis.

Table 1
Statistics of the ASD QA data used for the paper research

Parameter	Train	Valid	Test	Sum
QA pairs	523	126	107	756
Tokens (word level)	12 264	3 694	2 936	18 894
Unanswerable questions	5.8%	3.95%	4.35%	
Vocabulary size (sub-words)		30 522		
Out of Vocab rate		4.47%		

4. Approaches

4.1. Extractive Approach

The extractive approach, which is closely related to machine reading comprehension (MRC), was implemented using pre-trained Transformer Bidirectional Encoder Representations from Transformers (BERT) [8]. BERT is a model that was pre-trained for the masked language modeling (MLM) task. MLM is a task of predicting a masked token (for example, a word) according to its context surrounding. BERT was the first model that used MLM as a training task. The BERT performance shows that knowledge acquired through MLM solving can be successfully transferred to information retrieval and information extraction tasks. That makes BERT based models suitable for MRC and extractive QA [17]. BERT showed significant improvements in MRC performance obtained with SQuAD v1.1 [13] and SQuAD v2.0 [14] in comparison to architectures which previously showed State-of-the-Art results, such as models based on Bidirectional Long Short-Term Memory (BiLSTM), Gated Recurrent Unit (GRU) or Convolutional Neural Network (CNN).

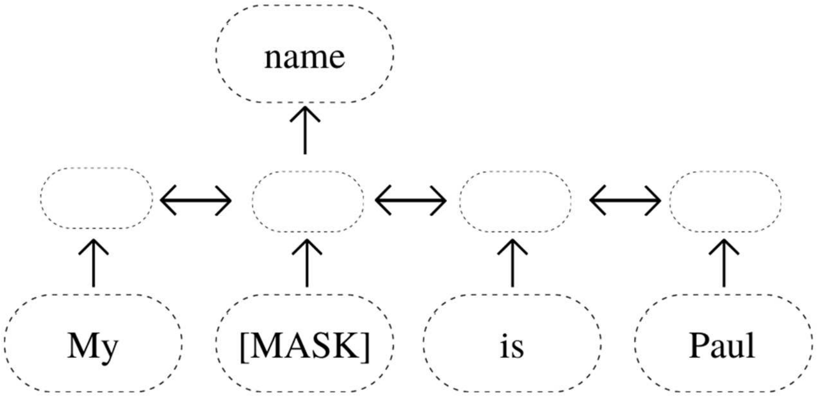


Figure 3: A concept of the MLM task

Figure 3 represents a concept of the MLM task. Bidirectional arrows in Figure 4 show bidirectional BERT processing. [MASK] illustrates a masked token that a model should predict. The sequence "My [MASK] is Paul" is input data. The word "name" is a model output, which is a result of model processing.

For the model training, a Russian dataset containing the QA-pairs on ASD and Asperger's syndrome was used (see Online Resources). The structure of the dataset represents a traditional MRC dataset structure (see Figure 3) containing three key MRC elements. Those elements are a reading passage, a set of questions posed on the passage, and a corresponding answer or a set of answers. Apart from that, the dataset includes a label, which indicates whether the question has an answer in the reading passage. Thus, if a user's question is irrelevant, a system should ignore it.

One can find different pre-trained BERT models. In this paper, the Multilingual BERT (MBERT) was used, BERT-Base, Multilingual Cased model [18]. M-BERT is a language model for 104 languages. The model was trained on Wikipedia data. The capabilities of the model allow using Transfer Learning techniques. Transfer Learning allows transferring knowledge from a general task to a specific one, or downstream one, by fine-tuning a pre-trained model or adding some layers to the original model architecture [19].

4.2. Generative Approach

In the current research, the generative approach was implemented with a Generative Pre-trained Transformer (GPT-2) [7]. We have used the original GPT-2 Large with 774 million parameters also known as 774M GPT-2. GPT-2 is a model for traditional language modeling. The model is unidirectional. GPT-2 analyzes only left-to-right context to predict the next token in a given sequence. Apart from showing high perplexity scores on the language modeling task, GPT-2 model shows high zero-shot performance on a wide range of other tasks. Zero-shot learning allows achieving high performance on domain-specific tasks without fine-tuning. Zero-shot learning capabilities can be revealed after evaluating a model on tasks, which it did not learn to solve during the training.

Among GPT-2 zero-shot learning achievements are solving question answering tasks and MRC, summarization, and translation without fine-tuning, and others. All this is achieved only by pre-training the model for traditional language modeling. Figure 4 represents the concept of traditional language modeling. Unidirectional arrows in Figure 4 show unidirectional GPT-2 processing. The question mark illustrates the model task to complete a given sequence. The sequence "Today I will" is input data, or prefix, which a model should continue. The sequence "go to school" is a model output.

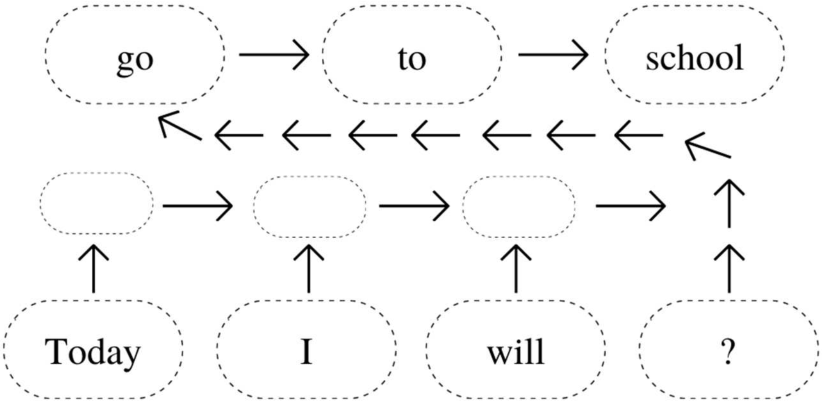


Figure 4: A concept of traditional language modeling

Both BERT based and GPT-2 based models were trained with a Russian dataset for the inclusion of people with autism spectrum disorder (ASD) (see Online Resources), although, for the generative model training, some changes were required. The dataset includes a special label indicating whether a question has an answer in a corresponding reading passage. If the value of this label is True, the answer presented in the dataset is special (see Figure 2). It is translated from Russian into English as "I cannot answer this question". This dataset feature was provided for generative models training, so they could learn to answer irrelevant questions politely.

The original version of the dataset is designed for MRC, so it had to be changed for the generative GPT-2 based model training. Firstly, all the answers and questions were extracted from the original dataset. Pairs of questions and answers, or QA-pairs, were located sequentially, separated by an empty row. All the QA-pairs were randomly shuffled. Secondly, the spans metadata was removed. Thirdly, the reading passages were not removed for the model failsafe. That was intended for cases when a possible answer to a user's question was contained in reading passages but absent in the training QA-pairs. Finally, the meta-information on answerable and unanswerable questions was removed, but the answer" I cannot answer this question" was saved for each unanswerable question.

5. Methodology

Transfer Learning techniques were used to fine-tune the models for the experiments. Transfer Learning allows using the knowledge gained while solving one general task to solve another similar one. The model is first trained on a large amount of data. Then, the pre-trained model is trained on the target dataset to solve a downstream problem. There are different Transfer Learning techniques. In this study, a fine-tuning strategy is used. The network trains end-to-end on a new custom dataset to adjust and adapt for the downstream task.

5.1. Metrics

For the question answering evaluation, F1-Score was used as proposed in [20]. F1 is the harmonic mean of the precision P and recall R. P is the fraction of relevant (true positive) model answers among the retrieved (true positive and false positive) ones. R is the fraction of the total amount of relevant model (true positive) answers among all the samples (true positive and false negative):

$$F = \frac{2PR}{P + R} \quad (1)$$

$$P = \frac{tp}{tp + fp} \quad (2)$$

$$R = \frac{tp}{tp + fn} \quad (3)$$

In question answering, true positive answers are the tokens shared between the correct (gold) tokens and all the predicted tokens. False positives are the predicted tokens absent in the correct (gold) answers, and false negatives are the tokens from the correct (gold) answer absent in the predicted ones. With this correction, the formula is the following as presented in the SQuAD evaluation script [21]:

$$F = \frac{2PR}{P + R} \quad (4)$$

$$P = \frac{shared}{shared\ +\ (predicted\ -\ shared)} \quad (5)$$

$$R = \frac{shared}{shared\ +\ (gold\ -\ shared)} \quad (6)$$

5.2. Experiment Setup

The model training was performed in Google Colaboratory with the Tesla T4 GPU. The code was implemented in Python [22] with the PyTorch library [23]. The configuration of the BERT based model and the GPT-2 based model is presented in Table 2. For the BERT base model, the HuggingFace Transformers repository [24] was used. For the GPT-2 based model, the Gpt-2- simple package was used [25]. We have also used the HuggingFace Transformers repository for the data pre-processing. During the pre-processing, we have not removed the stop words, because this might influence the structure of the utterances in the training data. Transformations of the structure of questions and answers might cause difficulties in natural language understanding during the question answering. However, this hypothesis needs verification with additional experiments.

Table 2
BERT based and GPT-2 based models configurations. General settings

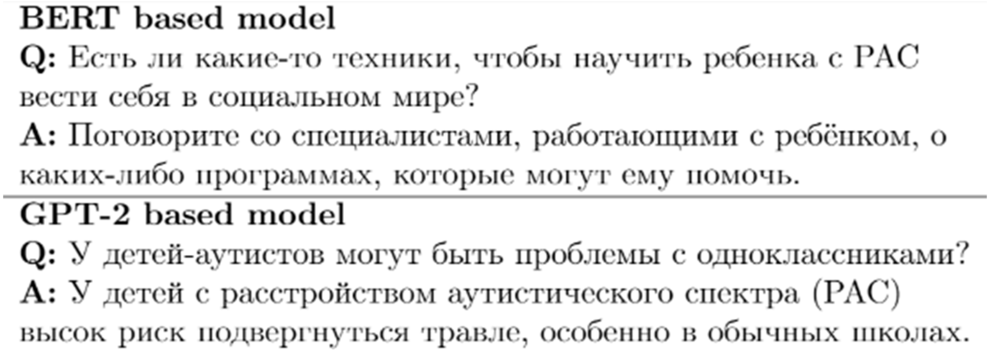
Parameter	BERT based model	GPT-2 based model
The base pre-trained model	BERT base multilingual	774 M GPT-2
The activation function	GELU	GELU
Tokens in an input sequence	512	1 024
Attention heads	12	20
The dropout ratio	0.1	0.1
The learning rate	5e-5	1e-4
The number of runs	10 epochs	3000 steps
Hidden layers	12	36
Vocabulary size	30 522 (word level)	50 257 (byte pair encoding level)
Other parameters	The size of encoder layers: 768 The size of feed-forward layer: 3 072	The number of embeddings: 1 280 The temperature (the output randomness): 0.7 The top K (the output diversity): 40

Table 3
Results obtained on the test data

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

6. Results and Analysis

Table 3 shows the results of both models obtained on the test set. 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. Table 4 shows the results of the linguistic analysis of both models outputs. The analysis allowed to reveal strengths and weaknesses of both models and make conclusions about their possible integration in inclusive education. Figure 5 shows a sample output.



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.

Figure 5: An output sample

Table 4
The results of the linguistic analysis

Language	BERT based model		GPT-2 based model	
level	Strength	Weakness	Strength	Weakness
Syntax	Complete answers if copes with a question	One-word or one-letter answers if do not cope	Often gives a complete answer	Frequent syntactic violations
Morphology	No or rare morphological violations	Truncates words if do not cope	Rare morphological violations	Might generate new words or word forms
Grammar	No or rare grammar mistakes	Unknown words might cause grammar mistakes	Can generate grammatically correct original utterances	Frequent grammar mistakes
Lexical diversity	Extracts single answer without lexical repetitions	Cannot generate unique utterances	Generate unique utterances without topical violation	Creates words that do not exist, repeats lexical constructions

7. Conclusion

After the linguistic analysis, the author of the paper defines four criteria of the models’ outputs evaluation. The criteria were determined according to the language levels that the author of the paper found essential for the analysis. The analysis focused on the evaluation of the QA models’ safety for their further integration into inclusive education. The criteria and language levels are the following: syntax level, morphology level, grammar correctness, lexical diversity.

On the syntax level, it was found that the extractive BERT based model can give full, syntactically correct sentences, but only if it copes with a user’s question. If the model cannot correctly recognize a user’s question, it would output a single word or a single letter, the first token from a corresponding context, with a high probability. For example, during the research, prepositions were very frequent in the model outputs. The generative GPT-2 based model, in turn, tends to give complete answers more often. However, its outputs contain frequent syntactic violations. That is inappropriate and is yet to be improved.

On the morphology level, the BERT based model did not show significant violations due to the extraction properties, although it truncated words in cases when it did not cope with a question. The GPT-2 based model could generate new words or word forms, which is worse because it might create unexisting lexical units. Grammar mistakes in the extractive model could only be caused by the presence of unknown words (due to the size of the training vocabulary) in the dataset. In the generative model, grammar mistakes were more frequent.

The extractive model did not make lexical repetitions extracting single answers. That makes the model clear and informative. However, this model cannot generate unique utterances. The generative model, in turn, could generate lexically diverse unique sentences. However, it also could create words that do not exist and repeat lexical constructions.

According to the conclusion of the study, extractive question answering is more reliable than generative question answering. The QA chatbot systems integration into inclusive education requires high alertness to its outputs. Thus, generative systems can be unsafe, as they might turn a tool for the information support or consultations into a toy, which is inappropriate.

Nevertheless, the capabilities of generative systems allow them to generate unique answers without grammar mistakes, lexical repetitions, and syntactic violations while maintaining factual accuracy. That makes them efficient. Although the score and errors point are yet far from optimal solution, the solutions presented in the paper provide future directions for improvement. For example, we can build models based on the extractive approach to extract accurate information containing the answer to the user’s question and use generative algorithms as part of the natural language interface to arrange the answer.

8. References

[1] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, in: NeurIPS, 2019.

[2] Z. Wu, M. Galley, C. Brockett, Y. Zhang, X. Gao, C. Quirk, R. Koncel-Kedziorski, J. Gao, H. Hajishirzi, M. Ostendorf, B. Dolan, A controllable model of grounded response generation, arXiv preprint arXiv:2005.00613 (2020).

[3] K. Lee, T. Kwiatkowski, P. A. Parikh, D. Das, Learning recurrent span representations for extractive question answering, arXiv preprint arXiv:1611.01436 (2017).

[4] O. Kolomiyets, M.-F. Moens, A survey on question answering technology from an information retrieval perspective, Inf. Sci. 181 (2011) 5412–5434. URL: https://doi.org/10.1016/j.ins.2011.07.047. doi:10.1016/j.ins.2011.07.047

[5] M. Lewis, A. Fan, Generative question answering - learning to answer the whole question, ICLR (2019).

[6] V. Shwartz, P. West, R. Le Bras, C. Bhagavatula, Y. Choi, Unsupervised commonsense question answering with self-talk, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online,

2020, pp. 4615–4629. URL: https://www.aclweb.org/anthology/2020.emnlp-main.373. doi:10.18653/v1/2020.emnlp-main.373

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/ anthology/N19-1423. doi:10.18653/v1/N19-1423

[9] J. Gao, M. Galley, L. Li, Neural approaches to conversational AI, Foundations and Trends® in Information Retrieval 13 (2019) 127–298. URL: http://dx.doi.org/10.1561/1500000074. doi:10.1561/1500000074

[10] G. Vassallo, G. Pilato, A. Augello, S. Gaglio, Phase Coherence in Conceptual Spaces for Conversational Agents, John Wiley Sons, Ltd, 2010, pp. 357–371. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470588222.ch18. doi:https://doi.org/10.1002/9780470588222.ch18

[11] V. Ta, C. Griffith, C. Boatfield, X. Wang, M. Civitello, H. Bader, E. DeCero, A. Loggarakis, User experiences of social support from companion chatbots in everyday contexts: Thematic analysis, Journal of Medical Internet Research 22 (2020) e16235. doi:10.2196/16235

[12] S. Liu, X. Zhang, S. Zhang, H. Wang, W. Zhang, Neural machine reading comprehension: Methods and trends, Applied Sciences 9 (2019) 3698. doi:10.3390/app9183698

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: https://www.aclweb.org/anthology/D16-1264. doi:10.18653/v1/D16-1264

[14] P. Rajpurkar, R. Jia, P. Liang, 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), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. URL: https://www.aclweb.org/anthology/P18-2124. doi:10.18653/v1/P18-2124

[15] Beautiful Soup documentation, 2020. URL: https://www.crummy.com/software/ BeautifulSoup/bs4/doc/

[16] Scikit-learn, 2021. URL: https://scikit-learn.org/

[17] N. Shazeer, Z. Lan, Y. Cheng, N. Ding, L. Hou, Talking-heads attention, arXiv preprint arXiv:2003.02436 (2020).

[18] Google Research, BERT, multilingual models, 2021. URL: https://github.com/ google-research/bert/blob/master/multilingual.md

[19] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer Learning in Natural Language Processing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 15–18. URL: https://www.aclweb.org/ anthology/N19-5004. doi:10.18653/v1/N19-5004

[20] L. Gillard, P. Bellot, M. El-Bèze, Question answering evaluation survey, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www. lrec-conf.org/proceedings/lrec2006/pdf/515_pdf.pdf

[21] The Stanford Question Answering Dataset, 2021. URL: https://rajpurkar.github.io/ SQuAD-explorer/

[22] Python, 2021. URL: https://www.python.org/

[23] PyTorch, 2021. URL: https://pytorch.org

[24] HuggingFace Transformers, 2021. URL: https://github.com/huggingface/transformers

[25] GPT-2-simple, 2021. URL: https://github.com/minimaxir/gpt-2-simple