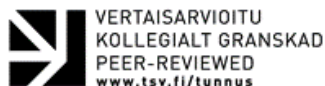Article

Supporting the Inclusion of People with Asperger Syndrome: Building a Customizable Chatbot with Transfer Learning

Victoria Firsanova[1]

[1] St. Petersburg University, St Petersburg, Russia

Abstract: The study focuses on building an informational Russian language chatbot, which aims to answer neurotypical and atypical people's questions about the inclusion of people with autism spectrum disorder and Asperger syndrome, in particular. Assuming that lack of awareness about the inclusion process and characteristics of people with special needs might cause communication difficulties or even conflicts between pupils, university and college students, or co-workers, a chatbot, which is based on reliable sources, provides the information informally and allows asking uncomfortable questions, could perhaps reduce stress levels during the inclusion. The paper describes two conceptual models of the chatbot. The first one is based on traditional language modeling with GPT-2, and the second one is based on BERT applied to question answering. The training data is collected from the informational websites about ASD, and its usage is agreed with the administration. For training BERT for question answering, the dataset structure was transformed according to the Stanford Question Answering Dataset (SQuAD). F1-score and perplexity metrics were used to evaluate the systems. The study shows the effectiveness of building conceptual models in tracking weaknesses and making significant adjustments at the design stage.

Keywords: Conversational AI, Chatbot, Question Answering, BERT, GPT-2

1 Introduction

In Russia, the inclusion of people with special needs is gradually developing, and tools aimed at making the process of inclusive studying or inclusive work more comfortable for both neurotypical and atypical people might become more relevant for the Russian-speaking community. One of the possible inclusion challenges might be a lack of awareness among pupils, students, or co-workers of inclusive organizations about their schoolmates, fellow students, or colleagues with special needs. The lack of awareness might cause communication difficulties, conflicts, and be the reason for bullying. A chatbot built on a dataset based on the information from reliable sources that answers a wide range of questions in an informal way, customizes according to the user's

age, and allows asking uncomfortable questions, could perhaps reduce stress levels during the inclusion.

The paper describes the process of building two conceptual models of an informational Russian language chatbot, which aims to answer various questions about the inclusion of people with Asperger syndrome. I have chosen this specification because I am the creator and leader of a non-commercial inclusive Natural Language Processing Discussion Club for students and support the inclusion of people with Asperger syndrome. The information about the Club is available at the following web address: https://vifirsanova.wixsite.com/nlp-discussion-club.

The first conceptual chatbot model is based on BERT (Devlin et al., 2019) applied to question answering, and the second one is based on traditional language modeling with GPT-2 (Radford et al., 2019). The final model of the informational chabot should not misrepresent or alter the information while generating unique answers in different styles according to the user's age. Thus, the goal of the study is to build two conceptual models based on multilingual BERT and GPT-2 to discover the capabilities of both models, evaluate and analyse their strengths and weaknesses, choose the best or decide to combine them and make dataset adjustments at the design stage.

Transfer learning techniques that allow extending the capabilities of a pre-trained model, for example, by fine-tuning it on custom datasets and downstream tasks, show strong performance on a wide range of NLP tasks (Pan et al., 2010, Ruder et al., 2019, Houlsby et al., 2019). According to the working principles of two selected transformer networks, I made a hypothesis. Firstly, the question answering system based on BERT (Devlin et al., 2019) might give accurate and informative answers, although it might not be able to interact with a user. Secondly, the language model based on GPT-2 (Radford et al., 2019) trained to take questions as input to generate answers might be entertaining and customizable according to the user's age. However, it might misrepresent the data and generate fake facts. The development of a conceptual model should help to define whether it will be better to choose one of two models or combine them to get the best of two worlds.

2 Related work

Conversational AI is a field that targets issues of human-AI interaction. To solve Conversational AI tasks, like building question answering systems and developing virtual assistants and chatbots, various natural language processing and machine learning approaches and their combinations can be applied. The challenges of conversational AI systems are usually task-dependent. Nevertheless, it should be strived to develop systems that generate coherent, topically relevant responses without word or content repetitions. These issues are significant and still require further investigation (Gao et al., 2019).

Models based on GPT-2 find its implementation in solving tasks of the conversational AI field. For example, DialoGPT (Zhang et al., 2020) is a dialogue generative pre-trained transformer based on GPT-2 (Radford et al., 2019). The model can be fine-tuned. It is designed for neural conversational systems research and development. The model was trained on Reddit commentaries, the structure of which is close to dialogue lines. Such type of data extracted from social media is considered to be quite efficient because it

represents the language of the real world. The development of the model opened the discussion of the further usage of reinforcement learning for further improvement and control of the generated responses.

BERT (Devlin et al., 2019) is widely used to create models for language understanding fine-tuned and customized for specific languages. One of the examples is the Italian language understanding model AlBERTo (Polignano et al., 2019), which is focused on the processing of language of social media and can be efficient, for example, in sentiment analysis. Another example is CamemBERT (Martin et al., 2019), a language model based on RoBERTa (Liu et al., 2019) fine-tuned for French. The performance of CamemBERT shows the importance of the use of web crawled data and proves that a relatively small dataset can lead to results as high as obtained with large corpora. The motivation for the development of such monolingual models is that many State-of-the-art (SOTA) models are trained for English data. As a result, the practical use of such models in languages except English is limited. This makes the development of monolingual Transformer-based NLP models for other languages, except English, highly relevant.

Creating captivating personalized chatbots is another issue that implies unique approaches. Although the goal of this study is to create an informational chatbot, not an entertaining one, the age customization requires some variations of personalization too. One of the possible solutions is to collect a dataset properly. For example, the dataset can be based on social media data, or even TV shows transcripts (Li et al., 2016). Another way is to collect conversational lines manually and paraphrase them additionally (Zhang et al., 2018).

It is also possible to create a psychologically safe personalized chatbot using conversational strategies. For example, Replika chatbot uses compliment strategies to provide users emotional support (Ta et al., 2020). According to some research, Replika generates compliments as initiative and reactive acts. Initiative acts allow the establishment of a dialogic assertion, and reactive acts allow it to satisfy; compliments can also be used as a form of greeting, response, or mean of user-machine relationship maintenance (Hakim et al., 2019). The study of Replika's user experience showed that AI dialogue agents have the potential to provide both emotional and informational support, which might become especially helpful for users in situations when traditional sources are unavailable (Ta et al., 2020).

3 Dataset

At the current stage, the dataset is being collected from informational websites about autism spectrum disorder in Russian. The dataset version 1.0 is already available online as an open-source: https://doi.org/10.6084/m9.figshare.13295831 (Firsanova 2020). The dataset used for building the conceptual chatbot model is extracted from the website of a project that supports people with Asperger syndrome and ASD https://www.aspergers.ru/. The usage of the information is agreed with the website administration. The data from the website was extracted with an HTML parser built with Beautiful Soup 4 library (Richardson 2020). Raw data from the website include texts of publications, which can be grouped into four thematic clusters shown in Figure 1.
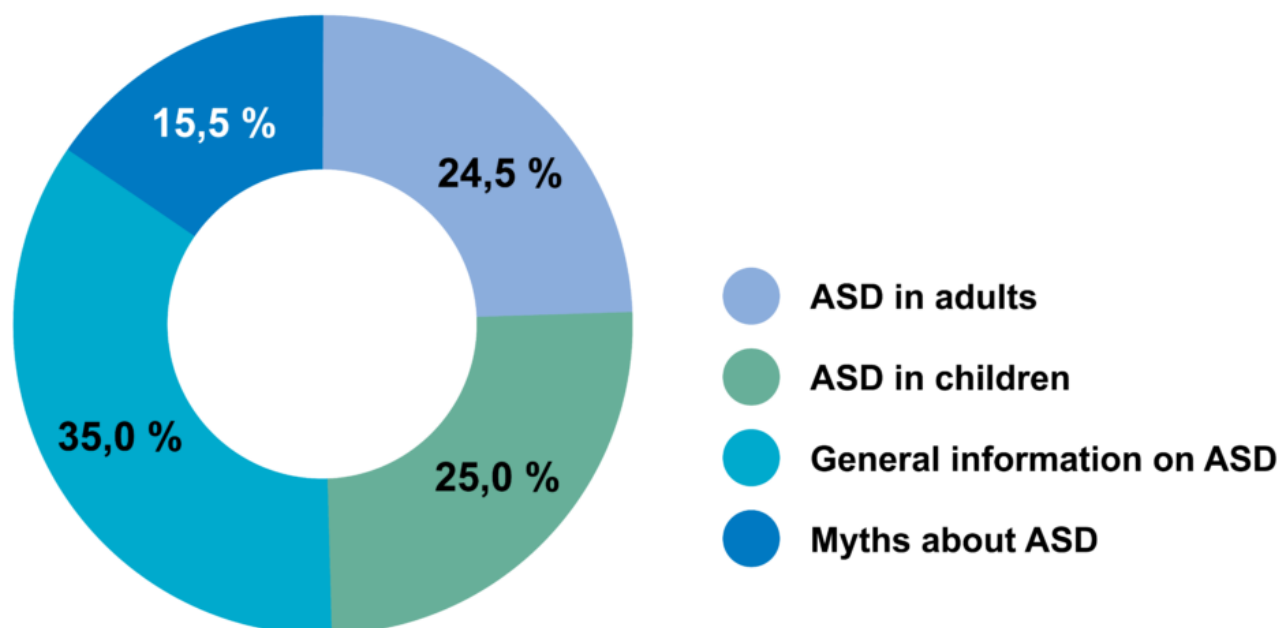
Figure 1. Dataset thematic clusters

Raw texts were cut into paragraphs no longer than 512 symbols to compile the dataset. Hypothetically, the chosen size should be enough to pose several questions on a text passage without causing difficulties during the training of a small conceptual model. Each paragraph was truncated until the last full stop, and a truncated sentence was replaced with the subsequent passage to avoid the loss of information. Then, a set of questions was posed on each dataset paragraph. The answer to each question was a segment of text from a corresponding passage. The compiled dataset consists of sets of question-answer pairs and text passages. This dataset structure is inspired by the Stanford Question Answering Dataset (SQuAD) for reading comprehension (Rajpurkar et al., 2016).

The dataset for the conceptual model has several features. Firstly, it includes paraphrased questions with identical answers. Another feature I find practical is unanswerable questions described in the original SQuAD 2.0 paper (Rajpurkar et al., 2018). Nevertheless, unanswerable questions of SQuAD 2.0 should be relevant to the paragraph and contain plausible answers to provide efficiency in reading comprehension, which gives no guarantees that a chatbot based on such a system would handle irrelevant user's questions. Consequently, I have decided to add some unanswerable irrelevant questions to the dataset so that a model could recognize inputs, which it should ignore. The reason for that is that the informational chatbot should give a user relevant information, and it should not entertain him or her like a toy.

After the transformations, the volume of the dataset for the conceptual model is 500 question-answer pairs, in total 10 711 words. 10% of all the questions are unanswerable. For BERT (Devlin et al., 2019) training for question answering, the dataset was provided with tags of answer start and end tokens positions in a corresponding text passage and with Boolean tags that indicate if a question is unanswerable. To train GPT-2 (Radford et al., 2019) to implement a language modeling approach to question answering, for each unanswerable question in

the dataset, a sample answer "I cannot answer this question" was appended. Figure 2 presents the dataset samples for both models.

GPT-2 Dataset Sample

Pair A. A standard question
Autizm lechitsya?*
*Can autism be treated?***
Autizm yavlyaetsya pozhiznennym sostoyaniem.*
*Autism is a lifelong condition.***

Pair B. An unanswerable question
Kak rabotayut nejroseti?*
*How neural networks work?***
Ya ne mogu otvetit' na etot vopros.*
*I cannot answer this question.***

* Russian transliteration
**Translation from Russian

```
Аспергерные аутисты - социопаты?
Аутичные люди - это не социопаты.

Все аспи - социопаты?
Аутичные люди - это не социопаты.

Люди с РАС на самом деле социопаты.
Аутичные люди - это не социопаты.

Аспи наплевать на других людей.
Нам не плевать на других людей.

Аутисты склонны игнорировать других людей?
Аутизм часто связывают с образом человека,
которому абсолютно плевать на существование
других людей, в то время как в реальности,
это просто проблема коммуникации.
Нам не плевать на других людей.

В чем секрет жизни?
Я не могу ответить на этот вопрос.
```

BERT Dataset Sample

Pair A. A standard question
Question: Autizm lechitsya?*
*Can autism be treated?***
Answer: Autizm yavlyaetsya pozhiznennym sostoyaniem.*
*Autism is a lifelong condition.***
Is unanswerable = False   **Start** = N   **End** = N + L
(N is the position of the first symbol of an answer in a corresponding text passage, L is the answer length)

Pair B. An unanswerable question
Question: Kak rabotayut nejroseti?*
*How neural networks work?***
Is unanswerable = True

```
"data": [
    {
        "title": "синдром Аспергера и РАС — общие сведения",
        "paragraphs": [
            {
                "qa_pairs": [
                    {
                        "question": "Достаточно ли информации об аути
                        "answer": [
                            {
                                "text": "Статьи для родителей дают неточн
                                "answer_start": 11,
                                "answer_end": 152
                            }
                        ],
                        "context": "Этот текст подготовила группа аутичны
```

Figure 2. The dataset samples

4 Method

4.1   Chatbot Architecture

Figure 3 presents two chatbot architectures. The model based on BERT, Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), takes a user's question as input and then extracts a corresponding text passage with the highest probability (a context from the database, which contains a possible answer) with beam search. Beam or greedy search can be used for the next word prediction in chatbots (Li et al., 2016), although I have decided to apply the algorithm to another task. To reach the goal, I have decided to transfer the ability to extract the data from comparable corpora for machine translation with beam or greedy search (Tillmann, 2009) to a new task, the extraction of a text passage corresponding to a given question. The solution is to implement the beam search extraction based on a corpus, where each question from the database is aligned with the corresponding paragraph. Then, the question answering system fine-tuned on the dataset described in section 2 takes a user's question and an extracted text passage (context) as input. The question answering system generates an answer, which is a target output. Hypothetically, such a system should give accurate, informative answers citing the corresponding segments of the source from the database, although it would not be flexible enough for the age customization.

The model based on GPT-2, Generative Pre-trained Transformer (Radford et al., 2019), uses traditional language modeling with attention (Vaswani et al., 2017) to take a question of a user as input and predict an answer. A simple convolutional neural network for multi-label sentence classification (Kim, 2014) will be used to implement the age customization. The CNN classifier should recognize the age group of a user by analyzing the text of his or her question. Then, according to the recognized age group, a fine-tuned GPT-2 modification will be chosen. The model will take the question as a prefix to generate an answer. As the GPT-2 modification would be fine-tuned according to the recognized age group, the generated answer will be customized according to the user's age. The hypothesis is the following. On the one hand, such a system might be interactive and entertaining due to the customization. On the other hand, the model might generate fake facts, misrepresent or alter the original data, and generate new words due to byte pair encoding (Gage, 1994).

BERT

User's question

↓

Beam search (context extraction)

↓

Question Answering System takes user's question and extracted context as input

↓

Question Answering System generates the answer

GPT-2

User's question

↓

CNN multi-label sentence classification to recognize user's age

↓

A model chosen according to a recognized age class takes user's question as prefix
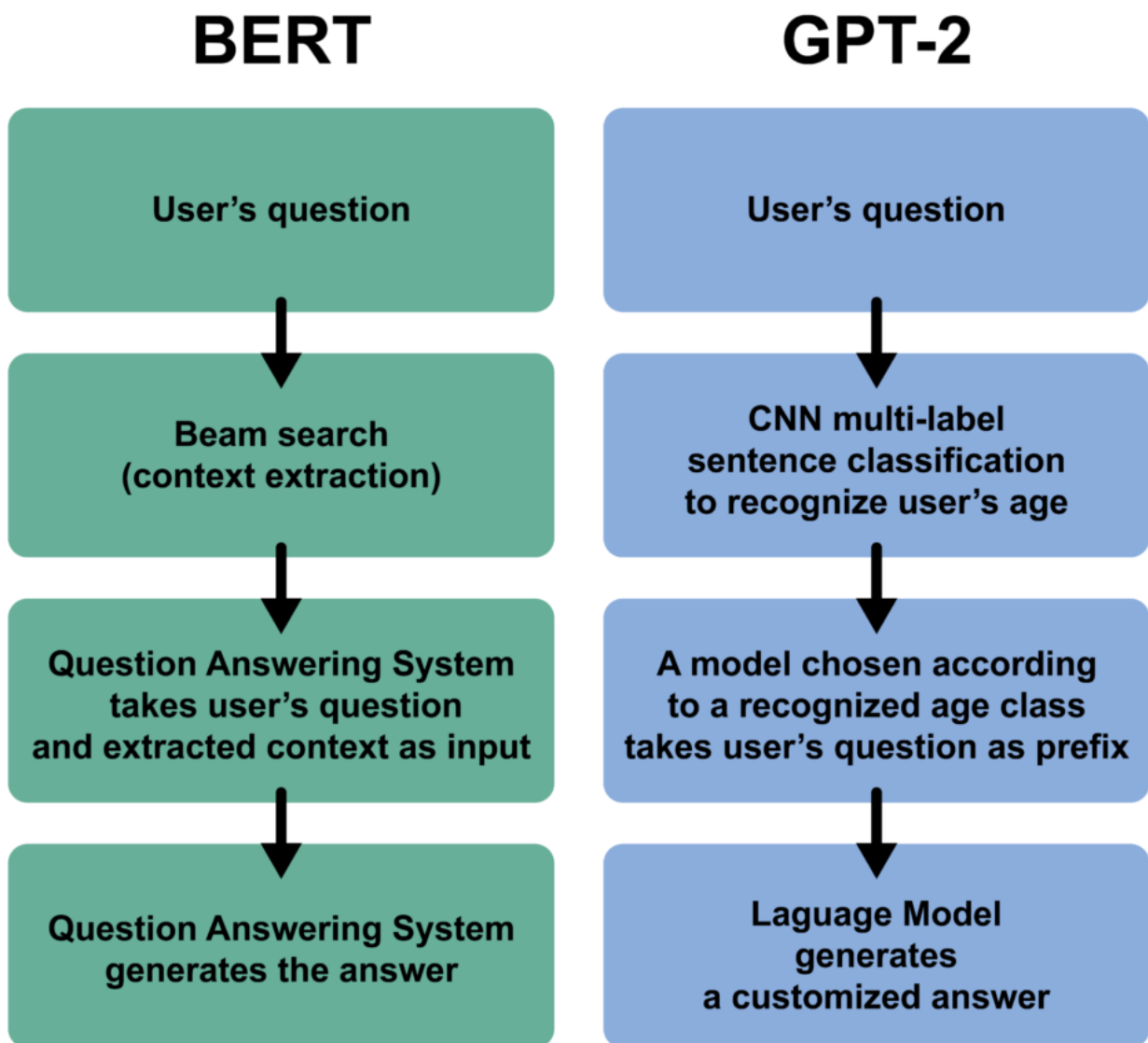
↓

Laguage Model generates a customized answer

Figure 3. The conceptual chatbot architecture

4.2   Learning Configuration

The training has been performed in Google Colaboratory (Google Colab, 2020) using Tesla P4 GPU (Google Cloud, 2020). During the experiments the following models were used: "BERT Base Multilingual Cased" (Google Research, 2019), GPT-2 774 million parameter model (OpenAI, 2019). The implementation is based on the open-source HuggingFace Transformer repository (HuggingFace, 2019). All the experiments were implemented in PyTorch (PyTorch, 2020).

The BERT model applied for question answering was trained using the following configuration. The dropout ratio for the attention probabilities and all fully connected layers in the embeddings, encoder, and pooler is 0.1, the activation function is GELU, the size of the encoder layer is 768, the size of the feed-forward layer is 3 072, the maximum number of tokens in an input sequence is 512, the number of attention heads is 12, the number of hidden layers is 12, and the vocabulary size is 30 522.

The GPT-2 model was trained using the following configuration. The activation function is GELU, the number of embeddings is 1 280, the maximum number of tokens in an input sequence is 1 024, the number of attention heads is 20, the number of hidden layers is 36, and the vocabulary size is 50 257. For the output generation, the temperature (the value that controls output randomness) is 0.7, and the top K (the value that controls the output diversity) is 40.

5 Evaluation and results

To evaluate the performance of a span-based question answering system, F1-Score metrics, the harmonic mean of the precision and recall, can be used (Celikyilmaz et al., 2020). The precision was calculated as the number of tokens common to the correct answer and the prediction (true positives) divided by the number of tokens presented in the prediction but absent in the correct answer (false positives). The recall was calculated as the number of tokens common to the correct answer and the prediction (true positives) divided by the number of tokens presented in the correct answer but absent in the prediction (false negatives). For GPT-2, the perplexity is also calculated to evaluate the language model performance by assessing the quality of a sample prediction (Brown et al., 1992). Table 1 presents the evaluation results.

	GPT-2	
	F1-Score	Perplexity
Original Dataset	0,43	24,86
Modified Dataset	0,67	24,79
	BERT	
	F1-Score	
Original Dataset	0,35	
Modified Dataset	0,56	

Table 1. Results obtained on the test set

Building and evaluating the conceptual chatbot model made it possible to discover the capabilities of both chosen architectures, track the dataset disadvantages, and make adjustments at the design stage. The initial BERT based model F1-Score is 0.35, and the initial GPT-2 based model F1-Score is

0.43. Although the initial F1-Score is considerably low, the perplexity of the GPT-2 based model is 24.86, which looks promising.

During the analysis of the model performance on the validation data, several issues were identified. In some cases, BERT based model generated correct but perceptually complicated answers. The problem might be caused by syntactical stylistic features, like parcellations, in a context from the database. It was decided to review the dataset and correct it so that each answer in the training dataset complies with the standard syntactic rules of Russian to solve the issue. One of the challenges of Russian is that it is syntactically flexible, which allows using a wide range of constructions to express thoughts but also complicates solving NLP tasks.

GPT-2 based model could recognize a question topic. For example, it could link the word "help" in the prefix with something connected with problems of people with ASD and generate collocations, like "high sensitivity" or "difficulties with social interaction". Unfortunately, the efficiency of such recognition is in doubt, as the model is not capable of generating explicit answers. Then, sets of paraphrased questions with identical answers were appended to the dataset to solve the issue.

After the transformations, the BERT based model F1-Score is 0.56, and the GPT-2 based model F1-Score is 0.67. Examples (1) and (2) present sample answers of the BERT and GPT-2 based models before the transformations, respectively. Examples (3) and (4) present sample answers of the BERT and GPT-2 based models after the transformations, respectively. Questions, answers, and contexts are presented in transliterated Russian, and the English translations of the texts are presented in brackets. In example (1) the answer is correct, however, the preposition is omitted, and example (2) shows a grammatically correct output, however, it does not contain the explicit answer to the question. Examples (3) and (4) show the outputs for the same questions. The performance of both models became significantly better, and the issues described earlier were solved.

(1) Question: Travyat li detej-autistov? (Are autistic children being bullied?) Answer: detej s rasstrojstvom autisticheskogo spektra (RAS) vysok risk podvergnut'sya travle, osobenno v obychnyh shkolah (*children with autism spectrum disorder (ASD) are at high risk of bullying, especially in mainstream schools)

Context: Travlya — eto vsegda nenormal'no! U detej s rasstrojstvom autisticheskogo spektra (RAS) vysok risk podvergnut'sya travle, osobenno v obychnyh shkolah… (Bullying is always abnormal! Children with autism spectrum disorder (ASD) are at high risk of bullying, especially in mainstream schools…)

(2) Question: Kak obshchat'sya s aspi? (How to communicate with aspies?) Answer: Sushchestvuet ogromnoe kolichestvo zabluzhdenij, vazhnyh faktov, o kotoryh lyudi dazhe ne podozrevayut, i neskol'ko universal'nyh istin, kotorye vsegda ignoriruyutsya. (There are a lot of misconceptions, important facts that people are not even aware of, and several universal truths that everyone always ignores.)

(3) Question: Travyat li detej-autistov? (Are autistic children being bullied?) Answer: Deti s rasstrojstvami autisticheskogo spektra (RAS) podvergayutsya

izdevatel'stvam chashche, chem tipichno razvivayushchiesya deti. (Children with autism spectrum disorder (ASD) are bullied more often than neurotypical children.)

Context: Travlya — eto vsegda nenormal'no! Deti s rasstrojstvami autisticheskogo spektra (RAS) podvergayutsya izdevatel'stvam chashche, chem tipichno razvivayushchiesya deti. (Bullying is always abnormal! Children with autism spectrum disorder (ASD) are bullied more often than neurotypical children.)

(4) Question: Kak obshchat'sya s aspi? (How to communicate with aspies?) Answer: Eto mozhet stat' trudnoj zadachej, no eto vpolne vozmozhno. (It can be a difficult task, but it is quite possible.)

6 Discussion

During the study discussion, two major discussion points were distinguished. The first point is the age customization, and the second one is the human-centric chatbot evaluation. The chatbot's target audience is schoolchildren, college and university students, young adults who work in inclusive organizations, and parents whose children attend inclusive classes. The chatbot should be suitable for both neurotypical and atypical people. This means that psycholinguistic research is as essential as AI research to build the system psychologically secure, reliable, and practically useful.

Language is one of the ways of information representation. It is flexible enough to describe an unlimited number of objects and phenomena using the rules of the combination and recombination of signs and codes, for example, grammar rules. Language allows describing even those phenomena that are beyond direct perception. With age, the processes of human perception and representation transform. For example, the role of visual images in children and adults differs. Children's imagery is rather specific than general, more vivid, and detailed (Slobin, 1979). Such aspects should be taken into account while building a system that supports age customization. Such a system should select utterances and expressions suitable for a user of a specific age group accurately to ensure a comfortable and psychologically secure interaction.

Furthermore, it is essential to take into consideration that people with autism and Asperger syndrome might have language features not intrinsic to people of the same age (Thomas & Fraser, 1994). For example, according to some research, autistic children might break rules of grammar and pragmatics, and even high-functioning people with ASD might break generally accepted rules of politeness because of the low sensitivity of situational cues (Simmons & Baltaxe, 1975). Ideally, the psychologically secure interaction with a chatbot also implies the respect of the politeness principle (Leech, 1983) and the maxims of tact, generosity, application, and modesty (Grice, 1975), which will require further research in the long run.

The human-centric evaluation of the chatbot is another significant point (Celikyilmaz et al., 2020). Such a form implies the organization of focus groups that represent different age groups of neurotypical and atypical people. Participants of the evaluation experiment will get a task to ask the chatbot some questions and then will take part in an assessment survey. The weaknesses of

the system will be identified according to the survey results, and then the system will be complemented.

7 Conclusion

In this work, the conceptual model of an informational chatbot, which answers basic questions about the inclusion of people with autism spectrum disorder and Asperger syndrome, in particular, is described. The model was trained using the official BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) source code, HuggingFace Transformer repository (HuggingFace, 2019), and PyTorch (PyTorch, 2020) on a Tesla P4 GPU performed over Google Colab (Google Colab, 2020). Two pre-trained models were fine-tuned on the dataset based on publications from the informational website https://www.aspergers.ru/. The BERT based model was fine-tuned for question answering. The dataset structure was transformed according to the Stanford Question Answering Dataset (Rajpurkar et al., 2016). The GPT-2 (Radford et al., 2019) based language model was trained to take questions as input to generate appropriate answers. After the analysis of the performance of both models, several issues were found and fixed. As a result, the F1-Score of both models increased by more than 20%. This proves that building small conceptual models allows making significant adjustments at the design stage, which might prevent future problems during the development of larger models with bigger volumes of data.

The results of the study will form the basis of the future chatbot, which will give answers to basic questions on autism spectrum disorder, Asperger syndrome, and the inclusion process, and provide emotional support. After the analysis of the performance of two conceptual models (BERT based and GPT-2 based), the capabilities of both were discovered. It was found that the bidirectional encoder used in BERT allows retrieving precise, accurate answers. However, to reach high performance, large volumes of training data are needed, and to transform a question answering system into a robust chatbot, additional algorithms, like beam search, are needed. In turn, the left-to-right decoder used in GPT-2 allows generating interactive detailed responses. GPT-2 based models do not require large volumes of data and can be customized according to the user's age just by fine-tuning and using simple classifiers. However, the specifics of the training dataset, which contains myths about autism and Asperger syndrome that the chatbot should dispel, might cause bias and, as a result, the chatbot might generate fake facts. Thus, it was found relevant to try combining them or fine-tune a model like BART (Lewis et al., 2019), which uses a sequence-to-sequence architecture with a bidirectional encoder and a left-to-right decoder, in the same task and dataset, in perspective. It is also planned to expand the dataset, provide it with more paraphrased questions, and create an additional corpus that includes age group labeling to develop the age customization.

The mission of the project is to support the inclusion of people with ASD, and Asperger syndrome in particular. This and further results of the study should not be used for commercial purposes. This is a charitable project that I hope will find a practical application and serve as a step towards equal work, education, and open science for all the people who are passionate about their field of activity, regardless of their special needs or social status. All people are different, all people are equal.

References

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, C. J., & Mercer, R. L. (1992). An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics*, *18*(1):31–40.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of Text Generation: A Survey. arXiv preprint arXiv:2006.14799

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *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), 4171–4186. doi:10.18653/v1/N19-1423

Firsanova, V. (2020). Autism Spectrum Disorder and Asperger Syndrome Question Answering Dataset 1.0. figshare. Dataset. https://doi.org/10.6084/m9.figshare.13295831.v1

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal*, *12*(2), 23–38.

Gao, J., Galley, M., & Li, L. (2019), Neural Approaches to Conversational AI. *Foundations and Trends® in Information Retrieval*, *12*(2–3), 127–298. doi:10.1561/1500000074

Google Colab. (2020). Retrieved October 16, 2020, from https://colab.research.google.com/

Google Cloud. Cloud GPUs (Graphics Processing Units). (2020). Retrieved October 16, 2020, from https://cloud.google.com/gpu

Google Research. BERT repository. (2019). Retrieved October 16, 2020, from https://github.com/google-research/bert

Grice, H. P. (1975). Logic and Conversation. *Speech Acts*, 41–58. doi:10.1163/9789004368811_003

Hakim, F., Indrayani, L., & Amalia, R. (2019). A Dialogic Analysis of Compliment Strategies Employed by Replika Chatbot. *Proceedings of the Third International Conference of Arts, Language and Culture (ICALC 2018)*. doi:10.2991/icalc-18.2019.38

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. arXiv preprint arXiv:1902.00751

HuggingFace. PyTorch transformer repository. (2019). Retrieved October 16, 2020, from https://github.com/huggingface/transformers

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi:10.3115/v1/d14-1181

Leech, G. (1983). *Principles of pragmatics*. London: Longman.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., & Dolan, B. (2016). A Persona-Based Neural Conversation Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 994–1003. doi:10.18653/v1/P16-1094

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692

Martin, L., Muller, B., Ortiz S. P., Dupont, Y., Romary, L., De la Clergerie, E., Seddah, D., & Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* July 2020, Online. doi:10.18653/v1/2020.acl-main.645

OpenAI. GPT-2 repository. (2019). Retrieved October 16, 2020, from https://github.com/openai/gpt-2

Pan, S. & Yang, Q. (2010). A Survey on Transfer Learning. Knowledge and Data Engineering. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. doi:10.1109/TKDE.2009.191

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019). ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. 6th Italian Conference on Computational Linguistics (CliC-it 2019).

PyTorch. (2020). Retrieved October 16, 2020, from https://pytorch.org/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Technical Report, OpenAI.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers). doi:10.18653/v1/p18-2124

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. doi:10.18653/v1/d16-1264

Richardson, L. (2020). *Beautiful Soup 4*. Retrieved November 27, 2020, from https://pypi.org/project/beautifulsoup4/

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer Learning in Natural Language Processing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. doi:10.18653/v1/N19-5004

Simmons, J. Q., & Baltaxe, C. (1975). Language patterns of adolescent autistics. *Journal of Autism and Childhood Schizophrenia*, *5*(4), 333–351. doi:10.1007/bf01540680

Slobin, D. I. (1979). *Psycholinguistics*. Scott, Foresman.

Ta, V., Griffith, C., Boatfield. C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. J *Med Internet Res*, *22*(3). doi: 10.2196/16235

Thomas, P., & Fraser, W. (1994). Linguistics, Human Communication and Psychiatry. *British Journal of Psychiatry*, *165*(5), 585–592. doi:10.1192/bjp.165.5.585

Tillmann, C. (2009). A Beam-Search Extraction Algorithm for Comparable Data. *ACLShort '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 225–228. doi:10.3115/1667583.1667653

Vaswani, A., Shazeer, N.,. Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need. Advances in Neural Information Processing Systems*, 6000–6010. doi: 10.18653/v1/W18-64076

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). doi:10.18653/v1/p18-1205

Zhang, Y., Sun, S., Galley, M., Chen, Y., Brockett, C., Gao, X., & Dolan, B. (2020). DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* doi:10.18653/v1/2020.acl-demos.30