Two Approaches to Building Dialogue Systems for People on the Spectrum

Victoria Firsanova*
Department of Computer and Applied Linguistics
Saint Petersburg State University
Saint Petersburg, Russia 199034
vifirsanova@gmail.com

Abstract

The paper presents a study on combining model- and data-centric approaches to building a question answering system for inclusion of people with autism spectrum disorder. The study shows that applying sequentially model- and data-centric approaches might allow achieving higher metric scores on closed-domain low-resourced datasets.

1 Introduction

The paper focuses on the development of a dialogue system for inclusive education. The system is designed for high-functioning people on the spectrum and their relatives. It aims to give relevant information about autism spectrum disorder. The mission of this work is to create a user-friendly tool for inclusion through natural language processing. Language-centric tools might one day become highly efficient in making steps towards a tolerant society; however, building such tools may present challenges.

I suppose that safety, as the main challenge, should come first in the development of tools for inclusion. For example, we can consider a dialogue system for inclusion system unsafe when it acts as an uninformative or even "aggressive" model that misleads users or evokes undesirable reactions. Is it hard to avoid building such a system? The answer is yes.

Although there are some inclusive and sociomedical datasets for building informational dialogue systems or teaching tools [1, 2, 3], because they are closed-domain and specific, it is impossible to use them, for example, for building models for the inclusion of people with autism. In turn, collecting a new dataset for such a purpose would require hours-long manual work of knowledgeable crowdworkers, which again complicates the whole process. New closed-domain sociomedical datasets might be very low-resourced, especially at the beginning of the journey. So what can be done while the dataset is small?

2 Data-centric Experiments

In the paper, I focused on improving the performance of a dialogue system trained on my low-resource dataset about autism spectrum disorder [4]. After some model-centric experiments with Transformer[5]-based SOTA models (see subsection 2.2), I have decided to conduct several experiments on the dataset design. My central argument is that experiments on a dataset structure might lead to improvements in the model performance. To prove or disprove it, I have used the results of a model-centric approach (the optimum fine-tuned Transformer-based model) for the data-centric experiments. Emphasis has been placed on low-resource closed-domain systems for inclusive education.

*https://github.com/vifirsanova

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

2.1 Dataset

The basis of the study is Autism Spectrum Disorder and Asperger Syndrome Question Answering Dataset 1.0 (ASD QA) [4] collected by me for building natural language dialogue systems for inclusive education. ASD QA is a Machine Reading Comprehension (MRC) dataset with a structure similar to the Stanford Question Answering Dataset 2.0 (SQuAD 2.0) [6]. MRC is a Natural Language Understanding (NLU) task, the goal of which is to teach a system to read and comprehend texts [7], for example, by learning to answer questions that can be extracted from a given reading passage.

The ASD QA dataset contains information for high-functioning people with autism spectrum disorder and their relatives covering different topics, for example, sport or communication. The dataset is being collected from reliable information resources in Russian, like [8]. One can use the dataset for building generative or extractive question answering systems, MRC models, closed-domain dialogue systems about autism, etc. The dataset is yet very low-resourced. It comprises 1,134 question-answer pairs and 96 informational paragraphs, which is 45,400 symbols or 6,578 words.

Figure 1 illustrates the dataset structure on the example of one topic set. Each topic set consists of several question-answer (QA) blocks, which are presented in Figure 1 as grey rectangles with several coloured rectangles inside. QA-blocks contain reading passages (see green rectangles in Figure 1), which are informational paragraphs containing answers (orange rectangles) to questions (yellow rectangles). Answers are extracts from reading passages annotated with sequence numbers of the first and the last symbols of spans in reading passages. QA-blocks have tags of relevance (blue rectangles) showing if a question has an answer in a reading passage (*relevant*), or not (*irrelevant*). So the dataset is provided with impossible questions, which a trained system should learn to ignore.

I have developed several dataset modifications for the data-centric experiments. The original dataset and its modifications can be found in [4]. The original version and others except for a "multiple" one contain only one answer to each question due to limited time and human resources for the manual work (in the year 2022, the author is launching a crowdsourcing project to finish the dataset collection). The idea of augmenting the dataset with new questions came after the SQuAD [6] structure investigation. The SQuAD initially contains questions with several possible answers; that is why the author created a version of the ASD QA called "multiple" which contains from one to four answers to each question. According to the diversity of content of a corresponding reading passage, the answers could be of different lengths.

According to [15], the twice larger datasets might lead to significant improvements in the model performance. I have decided to prove or disprove this theory by contradiction; I have created a "half-sized" version of the ASD QA dataset which comprises 50% of the shuffled original data and assumed that models trained on this modification will give twice lower metric scores than the ones trained on the original dataset version.

Then I have created modifications that would be easier for models to process by omitting the dataset elements that do not contain significant information and truncating the compulsory elements. A “no impossible” version does not include any irrelevant (or impossible) questions. A "short" version includes shortened answers from the original version (the shortage was losslessly applied only when possible). All the modifications (excepting "multiple" and "half-sized" versions) were applied to the first third of the dataset to accelerate the experiments. The versions are listed in the Mode column in Table 1 (see subsection 2.2).

2.2 Method

The training was performed over Google Colab [9] with the NVIDIA Tesla K80 graphics processing unit provided by the collaborative environment. The first model-centric part of the experiment was to examine SOTA Transformer-based models for extractive question answering (the ASD QA MRC dataset is perfect for this task). The aim was to find out and fine-tune the most powerful model for the second data-centric stage. Transformer [5] is an architecture based on attention mechanisms that allow weighing the significance of the input tokens. Transformer-based models allow applying transfer learning techniques achieving state-of-the-art results on a wide range of tasks only by transferring knowledge from one task or language to another. For example, Transformer-encoder models, like BERT [10], became quite efficient in MRC while being trained on tasks like Masked Language Modeling (MLM, or fill-in-the-gap task).

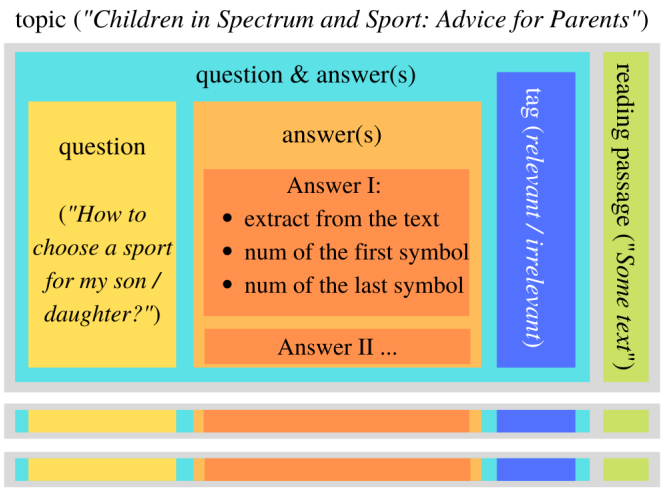

Figure 1: The ASD QA dataset structure.

Table 1: Results obtained on the original ASD QA dataset (model-centric approach) and it's modifications (data-centric approach)

Model-centric (original data)				Data-centric (modified data)			
Model	Precision	Recall	F1	Mode (for XLM-R)	Precision	Recall	F1
mBERT	0.42	0.25	0.31	short	0.37	0.29	0.33
mDBERT	0.51	0.24	0.33	multiple	0.39	0.36	0.38
XLM-R	0.39	0.36	**0.37**	no impossible	0.44	0.40	0.42
ruBERT	0.45	0.28	0.35	half-sized	0.72	0.04	0.07

Table 1 is divided into model- and data-centric parts. The model-centric part shows metric scores for several SOTA models after the hyperparameters optimization. I will not focus on the parameter optimization stage because the data-centric part of the work is more important for the paper. Among the chosen models are multilingual BERT (mBERT) [10], a multilingual distilled (compressed) version of BERT (mDistilBERT or mDBERT) [11], a cross-lingual model based on Facebook's RoBERTa (XLM-R) [12], and a version of BERT fine-tuned for the Russian language by Geotrend (ruBERT) [13] (the ASD QA dataset is in Russian).

XLM-R showed the best performance according to its F1-Score calculated as in SQuAD evaluation script [14]. As a result, XLM-R was chosen for the further data-centric stage. During the data-centric stage, I have retrained the XLM-R-based model with the optimum parameters (the learning rate is 3e-5, the batch size is 1, the number of epochs is 10, the dropout rate is 0.1) using different modifications of the ASD QA dataset described in subsection 2.1. The analysis of the results is presented in subsection 2.3.

2.3 Discussion

Before the experiments, I made several hypotheses about the model performance after the data-centric modifications. Firstly, I supposed that the shortage of answers would lead to higher performance because that would make the task more focused. During the first model-centric experiments, I noticed a shift in models' outputs. Mostly, they were correct, but the extracted spans were shifted to the right or left on several symbols (for example, *"by. Autichnye ljudi - jeto ne sociop..."*[2] instead of *"Autichnye ljudi - jeto ne sociopaty."*[3]). This phenomenon could be related to the WordPiece tokenization that breaks tokens into subwords. That also could be connected to the length of answers in the training dataset. Nevertheless, my hypothesis was not confirmed. Shortened answers lead to the decrease of

[2]Russian transliteration. Translation into English is as follows: "would. Autistic people are not sociop..."
[3]Russian transliteration. Translation into English is as follows: "Autistic people are not sociopaths."

F1-Score by 4% (see "short" in Table 1). Supposedly, the shorter answers prediction task became harder for the system because it decreased the probability of random guesses.

My second hypothesis was that the dataset with several answers would be less challenging for the system because, in some cases, it would have one or several alternatives for the output. The hypothesis was confirmed, but the increase was not significant, only by 1% (see "multiple" in Table 1). There were still a lot of incomplete answers consisting of one letter or one word (for example, *"A..."*[4], *"Ne..."*[5], *"Kak..."*[6]).

The third hypothesis was that the exclusion of irrelevant or impossible question-answer pairs would increase the model performance. The ASD QA dataset was provided with irrelevant questions (for example, *"Kak issledovat' iskusstvennyj intellekt?"*[7]), which have no answer in given reading passages. The model should learn to ignore such questions to be strictly informative and not entertaining. Nevertheless, learning to distinguish such questions significantly complicates the training process. The hypothesis was confirmed, (see "no impossible" in Table 1) the model performance without irrelevant questions increased by 5%.

The last hypothesis was that a half-sized dataset would decrease the model performance twice. This hypothesis was not confirmed. The model performance decreased by 30%. F1-Score became 0.07 instead of predicted 0.19 (see "half-sized" in Table 1). However, precision became unpredictably high and achieved 0.72. Within three previous dataset modifications, precision and recall were changing proportionally, whereas in this case, precision increased by 27%, while recall decreased by 24%. That means that the system gave more accurate but rare answers. Most outputs of the system were empty.

3 Conclusion

The paper presents a combined approach to building a question-answering system with model- and data-centric methods applied sequentially to achieve the best metric scores. The study focuses on a low-resource dataset about autism spectrum disorder. The study describes new empirical data-centric techniques that will be generalized to broader contexts in perspective.

The results achieved on a dataset version with shortened answers were unsuccessful. The experiment on a dataset with multiple answers did not bring significant improvement as well. Supposedly, a "multiple" dataset would allow achieving higher metric scores on a larger dataset.

The experiment without impossible answers allowed achieving the highest metric scores. However, I cannot reject that these results brought some losses. Such a system now cannot learn to distinguish irrelevant questions to ignore them. That provides much food for thought. Can we develop a model that would recognize irrelevant questions with some other methods? For example, can we use rule-based methods or engage generative algorithms? To answer this question, I need further investigation.

The experiment with a half-sized dataset was conducted to ensure that a twice larger dataset allows achieving much higher metric scores. This case illustrates the robustness of enlarging the volume of the training data. The still low-resource dataset collected manually by one person allows achieving significantly better results just by enlarging. However, such a high precision (0.72) was unexpected, and it is yet hard for me to interpret such a result.

Some of the answers of the developed system allow concluding that despite low metric scores, we can already use a system based on the ASD QA dataset in inclusive education for testing. For example, the outputs like *"Nalichie autizma ne delaet vashu zhizn' bessmyslennoj."*[8] or *"Ubedites', chto vash rebjonok znaet, chto takoe travlja."*[9] inspire. Apart from that, one of the main advantages of such extractive models is that they do not create occasional answers like generative ones. When they cannot output a correct answer, they usually keep silent.

[4]Russian transliteration. This can be a part of a word or a conjunction in Russian.

[5]Russian transliteration. Negative particle in Russian. This can be a part of a word or a negative particle in Russian.

[6]Russian transliteration. This can be a part of a word or a word "How" in Russian.

[7]Russian transliteration. Translation into English is as follows: "How to research artificial intelligence?"

[8]Russian transliteration. Translation into English: "Having autism does not make your life meaningless."

[9]Russian transliteration. Translation into English: "Please make sure that your child knows what bullying is."

References

[1] Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wang, N.X.R., Wilhelm, C., Xie, B., Raymond, D., Weld, D.S., Etzioni, O. & Kohlmeier, S. (2020) CORD-19: The COVID-19 Open Research Dataset. *ACL NLP-COVID Workshop 2020.*

[2] Chakravarthi, B.R. (2020) HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion. In: *Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media*, pp. 41–53. Barcelona, Spain (Online).

[3] Mamas, C. (2019) Learn to Conduct Descriptive Whole Social Network Analysis Within an Educational Setting in Ucinet With Data From the Inclusive Education Project (2015–2018) *SAGE Research Methods Datasets Part 2.*

[4] Firsanova, V. (2021) Autism Spectrum Disorder and Asperger Syndrome Question Answering Dataset 1.0 `https://figshare.com/articles/dataset/Autism_Spectrum_Disorder_and_Asperger_Syndrome_Question_Answering_Dataset_1_0/13295831`. Last accessed 25 Jul 2021.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. Kaiser, D., Polosukhin, I. (2017) Attention is All You Need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 6000–6010. New York: Curran Associates Inc.

[6] Rajpurkar, P., Jia, R., 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)*, pp. 784–789. Melbourne: Association for Computational Linguistics.

[7] Zhang, Z., Zhao, H., Wang, R. (2020) Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond. *Computational Linguistics Volume 1, Number 1,* pp. 1–51. Association for Computational Linguistics.

[8] Autistic City. `https://aspergers.ru/`. Last accessed 30 Sep 2021.

[9] Google Colab. `https://colab.research.google.com/`. Last accessed 30 Sep 2021.

[10] Devlin, J., Chang, M-W., Lee, K., 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),* pp. 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

[11] Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS*, pp. 1–5.

[12] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020) Unsupervised Cross-lingual Representation Learning at Scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. Association for Computational Linguistics.

[13] Geotrend - One click for intelligent data, `https://www.geotrend.fr/`. Last accessed 30 Sep 2021.

[14] Official evaluation script for SQuAD version 2.0., `https://worksheets.codalab.org/rest/bundles/0x6b567e1cf2e041ec80d7098f031c5c9e/contents/blob/`. Last accessed 30 Sep 2021.'

[15] A Chat with Andrew on MLOps: From Model-centric to Data-centric AI, `https://www.youtube.com/watch?v=06-AZXmwHjo&ab_channel=DeepLearningAI`. Last accessed 22 Nov 2021.