Conversational Intelligence Challenge: Accelerating Research with Crowd Science and Open Source

Mikhail Burtsev, Varvara Logacheva

■ *Development of conversational systems is one of the most challenging tasks in natural language processing, and it is especially hard in the case of open-domain dialogue. The main factors that hinder progress in this area are lack of training data and difficulty of automatic evaluation. Thus, to reliably evaluate the quality of such models, one needs to resort to time-consuming and expensive human evaluation. We tackle these problems by organizing the Conversational Intelligence Challenge (ConvAI) — open competition of dialogue systems. Our goals are threefold: to work out a good design for human evaluation of open-domain dialogue, to grow open-source code base for conversational systems, and to harvest and publish new datasets. Over the course of ConvAI1 and ConvAI2 competitions, we developed a framework for evaluation of chatbots in messaging platforms and used it to evaluate over 30 dialogue systems in two conversational tasks — discussion of short text snippets from Wikipedia and personalized small talk. These large-scale evaluation experiments were performed by recruiting volunteers as well as paid workers. As a result, we succeeded in collecting a dataset of around 5,000 long meaningful human-to-bot dialogues and got many insights into the organization of human evaluation. This dataset can be used to train an automatic evaluation model or to improve the quality of dialogue systems. Our analysis of ConvAI1 and ConvAI2 competitions shows that the future work in this area should be centered around the more active participation of volunteers in the assessment of dialogue systems. To achieve that, we plan to make the evaluation setup more engaging.*

Dialogue systems are traditionally divided into several types according to their purposes, and include question-answering systems that can give information on a particular topic of a range of topics (see for example Zhang, Yang, and Zhao 2020), goal-oriented dialogue systems that can accomplish a task such as looking for flights or ordering a pizza (see for example Madotto et al. 2020), and, finally, chatbots that do not have any particular task in a dialogue (that is, open domain; see for example Serban et al. 2017); their goal is to make a conversation interesting and enjoyable for a user, and to encourage the user to keep chatting for a longer time. For that, they can use a range of instruments, such as giving user quizzes (for example, the kAIb chatbot,[1] which participated in the first Conversational Intelligence Challenge [ConvAI1]), expressing empathy (Lin et al. 2019), trying to find a topic that is of interest to a user, and so on.

 ISSN 0738-4602

Similarly to other fields of natural language processing (NLP), dialogue systems benefitted from the recent deep learning revolution. The introduction of encoder–decoder neural architectures with attention (Bahdanau, Cho, and Bengio 2014) and, later, self-attentive transformers (Vaswani et al. 2017) took the quality of machine translation models to a new level (Wu et al. 2016). These advances also had a significant impact on dialogue systems. Architectures based on a transformer network were successfully applied to tasks of question answering (Lan et al. 2019) and goal-oriented dialogue (Lee, Lee, and Kim 2019). However, such an approach is moderately successful for nongoal-oriented chatbots.

One-turn-factoid question answering is probably the simplest task for deep learning techniques. The answer is usually unique, which allows building training and validation datasets easily. The quality of a system can be evaluated on a test set automatically via standard precision, recall, and F_1 metrics.

On the other hand, goal-directed dialogue, such as a restaurant booking, presents a more challenging setting. A task-oriented dialogue system needs to classify user intents and recognize entities mentioned by the user. Extracted entities should be retrieved from a knowledge base, sometimes with the help of reasoning. Another challenging problem is state tracking; that is, updating the information and beliefs about user intents after getting new input. Such dialogue systems are often modular, and the performance of natural language understanding, state tracking, and other modules is evaluated by comparing with ground truth answers. In addition to that, the accomplishment of the user's task can be easily tracked. Thus, goal-oriented systems can be evaluated with the percentage of successful dialogues.

Finally, building open-domain conversational systems poses a huge challenge for machine learning approaches. Open-domain dialogue data have high variability as there might be a large number of eligible responses given the same dialogue context. This creates problems with training and evaluation of neural chatbots.

The automatic metrics that are often used to assess the quality of an open-domain dialogue are weakly correlated with human scores (Liu et al. 2016). The goal of a chatbot is to generate answers that are natural and suit the context. Naturalness is commonly measured with the perplexity of a generated string (Serban et al. 2015), but perplexity cannot grasp the adequacy of an answer. To measure the appropriateness of a response, a generated string can be aligned with some oracle answer. The most popular metrics used for such oracle-based evaluation are BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002) and Metric for Evaluation of Translation with Explicit ORdering (Lavie and Agarwal 2007), originally used for evaluating machine translation models. Unfortunately, this is not an optimal way of evaluating chatbots either, because many different responses might be valid for the given context, so the oracle should not be unique (Liu et al. 2016).

On the other hand, there are recent works on the automatic evaluation of open-domain dialogue. The Automatic Dialogue Evaluation Model metric by Lowe et al. (2017) is trained on real human data to predict the scores of utterances given these utterances and their context. It yields only moderate correlation with human judgements. A work by Ghandeharioun et al. (2019) suggests using a linear combination of automatically computed statistics (sentiment scores of utterances, the similarity of human and bot utterances, the number of questions among bot utterances, and so forth) and shows that such a combined metric strongly correlates with the dialogue quality. Hashimoto, Zhang, and Liang (2019) describe a discriminator of bot and human responses whose error rate can be used as a quality score for a chatbot. While these metrics are promising and can be used for testing new features of dialogue systems and comparing model variants, the human assessment still is the most reliable method of evaluation. It is available at large scale in industry, as in, for example, a popular social bot XiaoIce (Zhou et al. 2018). It has a few hundred-million active users — which allows testing any new approaches on end-users directly. However, for the majority of academic or small industrial research groups, such a scenario is not feasible.

A possible solution to this problem is *citizen science* (Hand 2010). Volunteers can chat with bots in messengers and evaluate them. To attract volunteers, evaluation should be framed as an engaging activity such as playing a game, participating in a competition, or solving a challenge. Here, machine learning competitions provide good incentive to participate for volunteers. Participating as an assessor gives a possibility to interact with cutting-edge artificial intelligence (AI) technology and make a contribution to its development.

Large-scale public evaluation of open-domain dialogue systems not only allows us to compare submitted solutions, but also helps us to solve the problem of not having annotated data. Data annotated for quality of individual responses and overall dialogue is a valuable source for error analysis and training of the next generation of systems. The appropriate design of data-collection setup during chatbot competitions enables the organizers to release this data after the competition.

A number of conversational AI challenges and competitions with live human evaluation emerged in recent years. The most well-known industrial competition is the Alexa Prize[2] (Ram et al. 2018; Khatri et al. 2018). This is a competition targeted at building a socialbot that can converse coherently and engagingly with humans on popular topics for 20 minutes. Another small-scale analog is the Loebner Prize[3]. It is usually a one-day event where a small number of human judges converse in an open-ended manner via a textual interface with a chatterbot or another human for a few minutes and decide whether the peer is a machine or not. The Build It Break It[4] competition proposes a new type of shared task for AI problems that pits AI system builders against human breakers in an attempt to learn more about the generalizability of current NLP technology.

In 2016, we initiated and co-organized a series of conversational AI challenges as a part of the Conference on Neural Information Processing Systems[5]. The primary goal of this ConvAI series is to provide a framework for human evaluation of open-domain dialogue and explore the main challenges of state-of-the-art conversational systems.

ConvAI is designed to address the issues outlined earlier by pursuing the following goals. The first one is to work out a standard evaluation setup for open-domain dialogue systems and evaluate state-of-the-art conversational AI in order to define best practices of development. The second goal is to create a pool of open-source solutions that can be used as baselines for further research. The third goal is to collect and release, for public use, a dataset of evaluated human-to-human and human-to-bot conversations.

Description of Competitions

To this date, ConvAI competitions were organized twice — ConvAI1 in 2017 (Burtsev et al. 2018) and ConvAI2 2018 (Dinan et al. 2019). Both competitions were built around a focused open-domain conversational task. This means that the goal of dialogue was explicitly defined for peers, but the topic varied from one interaction to another. Such an approach makes it easier for a user to evaluate the quality of a conversation because its goal is clearly defined. Thus variance of scores is reduced, and the evaluation is more reliable. On the other hand, the creation of one particular conversational skill for many domains is much easier for participants than the creation of many skills for many domains.

ConvAI 1

The conversational task of ConvAI1 was to discuss a short text given to both peers at the beginning of a conversation. The text was a paragraph from a Wikipedia article randomly selected from the Stanford Question Answering Dataset (Rajpurkar et al. 2016). This makes possible narrowing down the topic of a conversation to the topic of a text. This setup is also targeted at making chatbot responses more informative and less uniform because the repetitive short answers have been identified as one of the problems in human-to-bot conversations (Yu et al. 2016). Importantly, this task poses a challenge for chatbots to retain a context in short-term memory, which is also one of the crucial problems in the development of dialogue systems (Li et al. 2016). The availability of open-source solutions for question answering over the Stanford Question Answering Dataset provided teams with a solid baseline and possibly some starting core code for the solution.

During a dialogue, the user was able to evaluate every bot answer by giving it a thumbs-up or thumbs-down interpreted as a positive or a negative score, respectively. This evaluation was not compulsory. Users initiated the end of dialogue themselves; this could be done any time. After that, they were asked to evaluate their experience according to three parameters: quality, breadth, and engagement. Quality was defined as the overall impression of dialogue, whereas breadth meant the extent to which the topic was covered, and engagement evaluated the peer's enthusiasm about the conversation. Each parameter was evaluated in terms of a 1-to-5 scale; this evaluation was compulsory.

ConvAI 2

The task of ConvAI1 turned out to be very challenging for chatbots as it was hard for them to maintain a coherent conversation about a provided text snippet.

Likewise, humans found it boring and difficult to adhere to the suggested topic, often because the texts were full of names and numbers, and hard to understand. To make a conversation more engaging, we chose Smalltalk as a conversational goal for the second ConvAI. The task was to tell about oneself and learn something about a peer. Every chat user and a chatbot had assigned profiles that consisted of four to five short sentences containing facts about a persona to represent in a conversation.

Each participant received a new persona for every dialogue. This scenario was also targeted at checking chatbot reading comprehension skills and short-term memory. Similarly to the previous year, chatbots were evaluated at the end of a conversation, but only with the quality score. An example of a typical ConvAI2 dialogue is shown in figure 1.

ConvAI2 competition provided to participants the PersonaChat dataset (Zhang et al. 2018), which could be used for training and evaluation of models. PersonaChat consists of conversations of eight to ten turns between two humans. The participants of these conversations discuss their work, hobbies, families, or pets. The information about themselves should be consistent with persona profiles received at the beginning of the conversations. Users were not required to use all the information from their profiles, but were recommended to divert a conversation toward topics covered in them. They were not allowed to use phrases from their profiles without changes, but rather elaborate on facts given there. Dialogues, as well as persona profiles, were collected via Amazon Mechanical Turk[6] service.

For this competition we expected a higher number of participants, so we could not arrange human assessment for all of the submitted models. Therefore, we organized a two-stage competition where, at the first stage, we selected a set of best-performing models using automatic evaluation, and then launched human assessment for these models. Thus, during the first stage, we had to use automatic metrics, although they are not an optimal evaluation strategy. We used three metrics: perplexity, F_1-score, and hits@1. We selected metrics that can evaluate generated text (perplexity, F_1-score) as well as retrieval models (hits@1, F_1-score).

Finally, the second edition of ConvAI had quite strong baselines. Their source code was made available in ParlAI (Miller et al. 2017)[7] so that participants could build upon these models. The models include a Key-Value Memory Network, which is a retrieval model, and two generative solutions: a

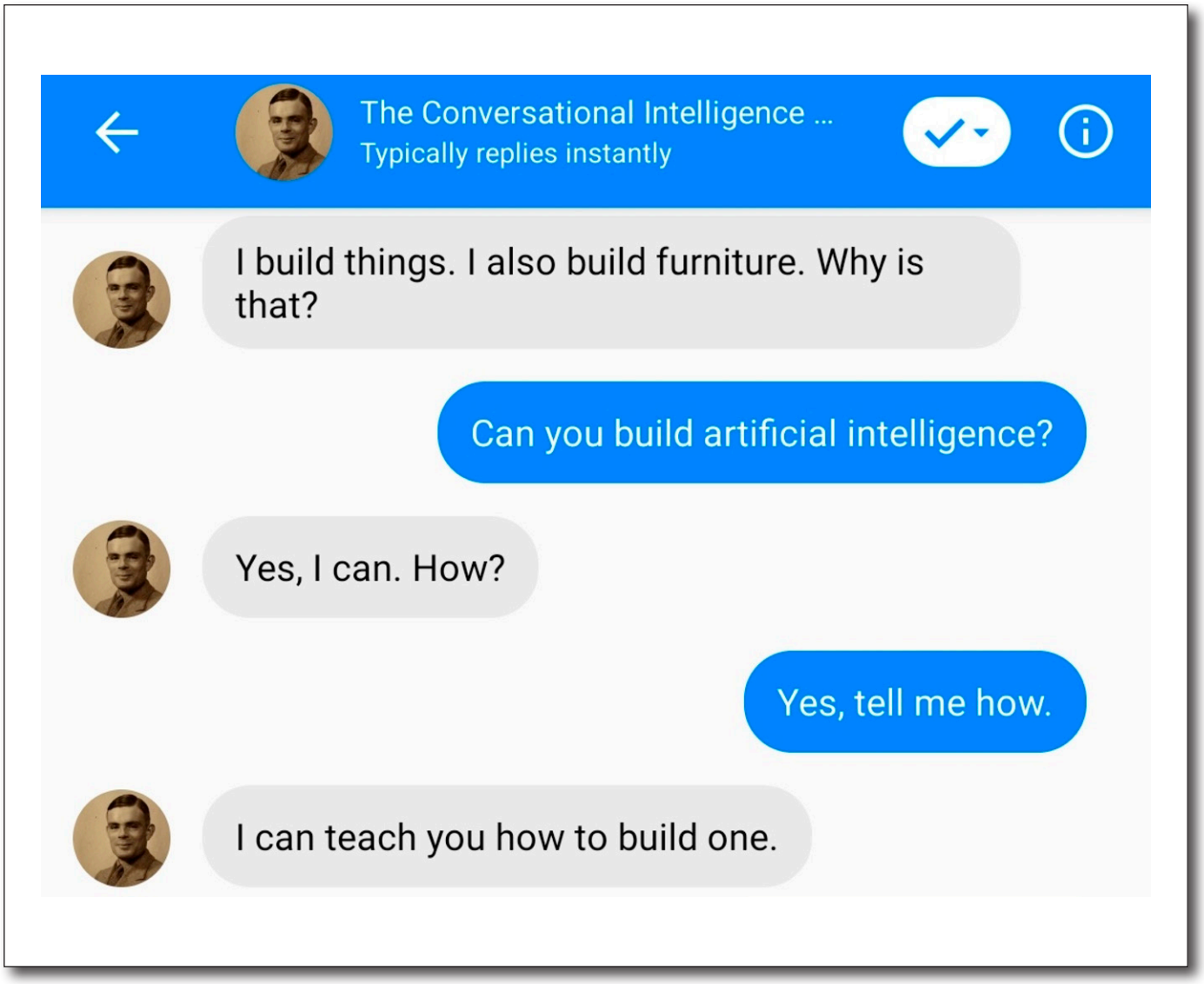

Figure 1. Example of ConvAI2 Dialogue with Chatbot Utterances on the Left and Human Volunteer Answers on the Right.

sequence-to-sequence model and a language model. All models are described in more details in the paper by Zhang et al. (2018).

Results of Competitions

The ConvAI competitions attracted a sizeable number of teams from academia and industry. In total, 10 teams participated in ConvAI1, and six of them were admitted for the final human assessment. At ConvAI2, solutions from 23 teams were submitted to the first round, and seven of them were selected for the human evaluation.

We required that the winners of ConvAI made their code open-source for future competitions and research in the area. Other participants were also encouraged to do so. Four participants of ConvAI1, including the winning teams, made their code available,[8] as well as the winners of ConvAI2.[9] Moreover, baseline models for ConvAI2 task are available online.

Overview of Solutions

In both editions of ConvAI, the participants used state-of-the-art methods existing at that time. However, they were held a year apart; therefore, ConvAI2 used more advanced network architectures. Besides that, the tasks of ConvAI1 and ConvAI2 were substantially different, and this difference is bound to be reflected in the design of the participating systems.

ConvAI1

At ConvAI1, chatbots had to solve very diverse tasks, such as reading of a text and talking to a user about this text. At the same time, during a dialogue, a user might want to switch to another arbitrary topic. Therefore, many of the solutions presented at ConvAI1 had a modular structure. They consisted of a set of modules, each targeted at solving a particular conversational task, and a dialogue manager. The

dialogue manager selected the most suitable module to generate an answer at each point in a conversation.

The design of the task suggested that users can ask questions to chatbots. Therefore, the majority of systems included a question-answering (QA) module. They used established QA architectures, such as DrQA (Chen et al. 2017) or FastQA (Weissenborn, Wiese, and Seiffe 2017), or designed their own approaches. QA modules were mostly trained on the Stanford Question Answering Dataset (Rajpurkar et al. 2016). Alternatively, they could retrieve answers from external sources — for example, Wikipedia or DBpedia. In addition to QA, some systems contained question-generation modules, anticipating that the user would enjoy a quiz-shaped conversation.

Most solutions also included a sequence-to-sequence model to generate free dialogue answers. The majority of teams used encoder–decoder architectures based on recurrent neural networks — for example, a bidirectional recurrent neural network with Long Short-Term Memory cells, or its more advanced versions, such as a variational hierarchical recurrent encoder–decoder (Serban et al. 2016). Instead of generating utterances, some teams used retrieval models that selected suitable phrases from a set of candidate responses. Many teams used an open-source chit-chat bot, Alicebot (Wallace 2009), as a separate module.

At ConvAI1, we did not provide any official training dataset and did not prohibit the use of any data, as long as it is freely available. In addition to the Stanford Question Answering Dataset and Wikipedia, models were trained on data fetched from Twitter, Reddit, and other web resources. Dialogue managers were trained on dialogues manually labeled for quality at the utterance level to classify generated utterances as good or bad. Several teams also used reinforcement learning. They used human actions during dialogue as rewards.

ConvAI2

ConvAI2 was different from the previous competition in many respects. First, the task did not imply switching between conversation styles. Therefore, the majority of participants designed end-to-end conversational models without explicit dialogue management. The architectures of these models were also different from the previous year. While most of the generation models at ConvAI1 were recurrent-neural-network–based, many of participants of ConvAI2 preferred Transformer architecture (Vaswani et al. 2017), which was introduced shortly after ConvAI1.

They trained Transformers themselves or used publicly available pretrained models, such as the Generative Pretrained Transformer model (Radford et al. 2018). On the other hand, some teams that participated in the final round still adhered to recurrent neural networks, and one of them preferred a retrieval model adapted from Wu et al. (2016).

Another novelty of ConvAI2 was the official training dataset, namely PersonaChat. However, similarly to ConvAI1, PersonaChat was not the only allowable training resource, so all teams used additional data for training. These were chit-chat dialogue datasets, such as DailyDialog (Li et al. 2017), Reddit comments, OpenSubtitles (Lison and Tiedemann 2016), and Switchboard.[10] In addition to that, models were pretrained on plaintext datasets — for example, the One Billion Word dataset (Chelba et al. 2013).

Comparison of Dialogues

Throughout the overall ConvAI competitions, we collected four datasets[11] the ConvAI1 dataset (Logacheva et al. 2018), and three ConvAI2 datasets (Logacheva et al. 2019), which include an official human evaluation dataset collected via Amazon Mechanical Turk,[12] a human evaluation by paid workers collected via Yandex. Toloka[13] whom we further denote as *tolokers*, and wild evaluation by volunteers. The evaluation setups for tolokers and volunteers were identical. The only difference was that the former were paid for chatting with systems, and the latter were doing that just for the fun. The instructions to users and evaluation scheme were identical.

On the other hand, the official human evaluation at ConvAI2 collected via Amazon Mechanical Turk was not fully consistent with the other three datasets, because it used the 1-to-4 scale for dialogue-level evaluation instead of the 1-to-5 scale. In addition to that, the dialogues were automatically finished by the evaluation framework when reaching the length of four to six turns. The length was selected randomly from $L = \{4, 5, 6\}$. Therefore, evaluation of these data is incomparable with the evaluation for the other three datasets, so we do not report its statistics here. For the thorough analysis of these dialogues, we refer readers to the official description of ConvAI2 (Dinan et al. 2019).

Table 1 shows the statistics of the three datasets collected in 2017 and 2018. There, we see that only 50 percent of dialogues were long enough, that is, they lasted more than three turns. This suggests that short dialogues are inevitable in crowdsourcing experiments, and when planning such experiments one needs to take into account that a large proportion (up to 50 percent) of the collected data will not be usable.

Another common feature of all datasets is the difference in the length of utterances generated by human participants and chatbots. Chatbots produce longer answers on average. As was already suggested in the description of the 2017 data (Logacheva et al. 2018), humans do not strive to output grammatically correct and self-contained sentences. They tend to use elliptic constructions, and can produce extremely short answers. This trend also holds in the 2018 datasets, and suggests that it is an inherent characteristic of human-to-bot conversations.

Task Difficulty for Humans and Bots

Table 1 shows that the length of dialogues increased in 2018 compared with 2017. In general, this means that chatbots managed to keep user attention for a longer time. This does not mean that the dialogues became better; as was shown in the analysis of

Dataset	2017 Data Volunteers	2018 Data Tolokers	2018 Data Volunteers	Total
Dialogues	4,750	3,197	1,209	4,406
Long Dialogues (>3 turns)	2,640 (56%)	1,696 (53%)	537 (44%)	2,233 (51%)
Average Utterances per Dialogue	17.16	21.41	23.97	22.03
Average Words per Utterance				
Human	5.73	6.26	5.17	5.96
Bot	8.88	10.19	10.41	10.26
Total	7.38	8.37	7.77	8.21
Average Score				
Human	3.69	—	—	—
Bot	1.07	2.70	2.76	—

Table 1. Statistics of Dialogue Datasets Collected during ConvAI Competitions.

ConvAI1 data (Logacheva et al. 2018), the length of dialogues is only weakly correlated with their quality. Nevertheless, being able to capture user attention for a long time is an essential (although not sufficient) step on the way to a good conversation. Systems participating in ConvAI2 not only yielded longer dialogues but also received higher scores from users. Table 2 shows the comparison of scores given by users to chats with bots during ConvAI1 and ConvAI2. Analysis of ConvAI1 data (Logacheva et al. 2018) suggests that scores for quality are strongly correlated with scores for the other two parameters, breadth and engagement. So we use quality as the only proxy for overall user satisfaction for ConvAI1. In ConvAI2, we did not use breadth and engagement anymore. Utterance-level scores are optional thumbs-up and thumbs-down given by users to individual utterances during a conversation. We interpreted positive scores as "1" and negative scores as "0" and averaged them for a particular chatbot (or for all chatbots in the "Average" row). Note that only around 15 percent utterances per bot were rated. Unrated utterances did not participate in the utterance-level evaluation.

Category	2017	2018
Dialogue-Level Scores		
Best Bot	2.4	3.3
Worst Bot	1.4	2.1
Average	2.2	2.9
Utterance-Level Scores		
Best Bot	0.52	0.72
Worst Bot	0.19	0.29
Average	0.44	0.61

Table 2. User Scores in 2017 and 2018 Experiments for the Best-Performing and the Worst-Performing Systems.

The overall level of user satisfaction is higher for ConvAI2 because the scores of both worst-performing and best-performing chatbots grew in 2018 compared with the previous year. This can be explained by the fact that the quality of models submitted to ConvAI2 is higher. However, it might be that the ConvAI2 task is easier or more engaging for users than the task from the first ConvAI.

Analysis of Evaluation Schemes

For human evaluation of ConvAI1 dialogues, we used three dialogue-level evaluation metrics: overall quality of dialogue, dialogue breadth, and peer engagement.

At ConvAI2, users explicitly evaluated dialogues with overall quality only. In addition to that, we included a new metric for evaluation of chatbot consistency. After a dialogue, we asked users to choose a persona that belonged to their peer from two persona descriptions. Our intuition was that being consistent with a profile allows a chatbot to have a better conversation because lack of consistency is a persistent complaint about current state-of-the-art dialogue systems. However, the correlation between dialogue quality and persona detection scores was only moderate (0.45). This shows that a model can be consistent without producing a good dialogue — such as, for example, by randomly throwing in random facts from its profile. Such a strategy yields a high persona detection score, but does not improve the system's quality.

In addition to human assessment, at the first stage of ConvAI2 we evaluated the chatbots with automatic metrics (perplexity, F_1-score, and hits@1). However, they turned out to disagree with subsequent evaluation by users. Table 3 shows the automatic and manual scores of the six chatbots that showed

Category	Human Evaluation	Automatic Evaluation		
		Perplexity	Hits@1	F_1-score
Lost in Conversation	3.11	—	17.1	17.77
Hugging Face	2.68	16.28	80.7	19.5
Little Baby	2.44	—	64.8	—
ParlAI Team (Baseline)	2.44	—	55.2	11.9
Mohd Shadab Alam	2.33	29.94	13.8	16.91
Happy Minions	1.92	29.01	—	16.01
ADAPT Centre	1.6	31.4	—	18.39

Table 3. The Official Results of ConvAI2 in Terms of Human and Automatic Metrics.

the best results in the first stage and were selected for the human assessment. The ranking of these models produced by the automatic scores shows a weak negative correlation with ranking based on human-dialogue–level quality scores. For example, the model that outperformed other competitors by a large margin in all three automatic metrics was considered only second-best by humans. This means that these scores cannot replace human evaluation.

In addition to explicit evaluation, we tried to find features of dialogues that could serve as a proxy for a dialogue quality. We observed only a weak correlation of dialogue level quality and the number of utterances in dialogue, the number of unique words, and unique trigrams. The length of dialogue was particularly interesting for us because it is often used as a quality metric in commercial chatbots (Zhou et al. 2018; Kumar et al. 2017). However, we found out that a dialogue of any length is equally likely to have any score. The only exception was for the score of "1." It was more often given to short dialogues. Therefore, if a dialogue is short, it is likely to be considered bad. Otherwise, dialogue length does not tell us much about user experience.

Discussion

Application of machine learning methods in the field of dialogue systems opens highly promising possibilities, but poses problems with training and evaluation of open-domain chatbots.

What makes a good conversation for a chatbot? There is no clear answer to this question yet. Lack of understanding prohibits implementation of automatic metrics of open-domain dialogue quality. Attempts to apply metrics from machine translation were not very successful. BLEU and other metrics based on the precision of words or n-grams do not suffice for this purpose, even if they are extended with word or sentence vector representations. They assume lexical or at least semantic similarity of an answer to an oracle, which is often wrong for the dialogue. Here, we have a fundamental problem with the high variability of meaningful responses for the same dialogue context. Today, a reliable evaluation of a dialogue system can only be performed by humans.

Human evaluation can be done via crowd-sourcing platforms like Amazon Mechanical Turk or by volunteers. In our competitions, we experimented with both types. The motivations and goals of volunteers and paid workers during conversation are different. Paid workers earn money and, in general, are not excited by the task itself. On the other hand, volunteers have a natural interest in exploring different aspects of conversation with an intelligent agent. When we compared the performance of volunteers from wild evaluation and paid users from Yandex. Toloka (tolokers), we found that there is a larger number of recurring users among tolokers. There were around 100 tolokers who conducted at least 10 conversations, whereas only around 10 volunteers had more than two dialogues. Thus, because experimental setups were identical for volunteers and paid workers (except for the payment), we can see that the task was not exciting enough to provide the motivation for conducting a large number of dialogues.

A large number of recurring users might be a better scenario for research purposes. Having many dialogues from the same user allows for detection of, and removal of, user-specific biases. Moreover, users who conduct only one or two dialogues tend to give lower scores to bots. This might indicate their overly high expectations at the beginning of a testing session. After one dialogue, a user might be dissatisfied and rate a chatbot very low. However, if she or he has multiple dialogues, then expectations and scores of subsequent dialogues would be adjusted and become more stable. Thus, human evaluation of chatbots should probably take into account only scores from recurring users, and would allow a training period for new users.

An important aspect of a competition is that all systems are evaluated on the same population of users. This is vital for the comparison of different solutions and reproducibility of results. Thus, even in the case when two studies are using the same crowd-sourcing platform, factors such as slight

differences in the description of the task for workers, parameters of worker selection, and the fact that a pool of workers changes over time, inevitably bias results, making them incompatible. Competitions provide a much better setup by having the same testing conditions for all systems.

The development of advanced open-domain conversational agent requires not only advanced machine learning techniques but also a large amount of engineering effort spiced with inevitable tips and tricks on how to choose hyperparameters and train a system. This is why we asked winners to provide open-source code of their solutions. As a result, the teams uploaded their code, technical papers, and presentations on GitHub and even promoted them with blog posts. This is how ConvAI competitions can lower barriers to enter the field and accelerate research and development of more intelligent conversational AI systems.

As a part of ConvAI, we collected about 5,000 human-to-human and human-to-bot dialogues where utterances, as well as whole dialogues, are labeled for quality. The potential of this data are waiting to be explored by the community. In contrast to a huge number of raw dialogues on the web, ConvAI data gives a unique opportunity to compare the performance of humans and bots for the same task. The use of dialogue scores to train a response-selection model for the dialogue system is a practical application of the dataset.

Future Directions

Each of the two editions of ConvAI provided many interesting and useful insights into the testing of dialogue systems. Still, a reasonable and robust evaluation and comparison framework for open-domain conversational systems does not exist. Thus, we are looking forward to exploring untapped possibilities with new public competitions.

Another reason we need a new conversational competition is the fact that NLP is developing very fast. The new powerful architectures and, in particular, new ways of pretraining language models (InferSent, Conneau et al. 2017; Bidirectional Encoder Representations from Transformers, Devlin et al. 2018; Universal Language Model Fine-Tuning, Howard and Ruder 2018; and Generative Pretrained Transfomer, Radford et al. 2018) dramatically improved the performance of neural networks on many NLP tasks. Thus, a new competition on dialogue modeling can measure to what extent these new developments influenced chatbots. In addition to that, many pretrained models of good quality are freely available now, so to develop a well-performing dialogue system, one does not need access to powerful machines and other resources. This fact increases the number of potential participants of such competitions.

A potential new round of ConvAI competition should have a task that combines strong points of the previous editions. Similarly to the first ConvAI, dialogues should be centered around some meaningful topic, but to make the conversation more engaging, we should let the user choose that topic. To make a goal of dialogue more grounded and clearer to the user, we should give a more concrete and detailed task — for example, to ask for a particular piece of information from a chatbot, as was done in ConvAI2. The task completion could be checked by a short quiz on the chosen topic. This will also provide additional gamification, which can help attract volunteers and encourage them to conduct more dialogues.

Acknowledgments

The authors thank all co-organizers of the ConvAI challenges, and members of the advisory board: Iulian Serban, Alexander Rudnicky, Emily Dinan, Jason Weston, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Ryan Lowe, Shrimai Prabhumoye, Jason Williams, Joelle Pineau, Yoshua Bengio, and Alan W. Black. The authors are grateful to the team of the Yandex.Toloka project for their help with the evaluation.

The work was supported by the National Technology Initiative (Russia) and PAO Sberbank project ID 0000000007417F630002.

Notes

1. github.com/lifelongeek/kaib_nips
2. developer.amazon.com/alexaprize
3. en.wikipedia.org/wiki/Loebner_Prize
4. bibinlp.umiacs.umd.edu/
5. neurips.cc
6. www.mturk.com
7. github.com/facebookresearch/ParlAI/tree/master/projects/convai2
8. github.com/DeepPavlov/convai/tree/master/2017/solutions
9. github.com/atselousov/transformer_chatbot
10. web.stanford.edu/?jurafsky/ws97
11. The datasets are available online at convai.io/data
12. www.mturk.com
13. toloka.yandex.ru
14. statmt.org/wmt18/quality-estimation-task.html

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. Paper presented at the International Conference on Learning Representations (ICLR) 2015, San Diego, CA, May 7–9.

Burtsev, M.; Logacheva, V.; Malykh, V.; Serban, I.; Lowe, R.; Prabhumoye, S.; Black, A. W.; Rudnicky, A.; and Bengio, Y. 2018. *The First Conversational Intelligence Challenge. The NIPS '17 Competition: Building Intelligent Systems*. Berlin, Germany: Springer.

Chelba, C.; Mikolov, T.; Schuster, M.; Ge, Q.; Brants, T.; and Koehn, P. 2013. *One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling*. arXiv:1312.3005. Ithaca, NY: Cornell University Library.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. *Reading Wikipedia To Answer Open-Domain Questions*. arXiv:1704.00051. Ithaca, NY: Cornell University Library.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. *Supervised Learning of Universal Sentence Representations From Natural Language Inference Data*. arXiv:/1705.02364. Ithaca, NY: Cornell University Library.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805. Ithaca, NY: Cornell University Library.

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R., et al. 2019. *The Second Conversational Intelligence Challenge (ConvAI2). NIPS 2018 Competition Track*. arXiv:1902.00098. Ithaca, NY: Cornell University Library.

Ghandeharioun, A.; Shen, J. H.; Jaques, N.; Ferguson, C.; Jones, N.; Lapedriza, A.; and Picard, R. 2019. Approximating Interactive Human Evaluation With Self-Play for Open-Domain Dialog Systems. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*. 13658–69. Redhook, NY: Curran Associates, Inc.

Hand, E. 2010. Citizen Science: People Power. *NATNews* 466(7307): 685–7.

Hashimoto, T.; Zhang, H.; and Liang, P. 2019. Unifying Human and Statistical Evaluation for Natural Language Generation. 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), 1689–1701. Stroudsburg, PA: Association for Computational Linguistics doi.org/10.18653/v1/N19-1169.

Howard, J., and Ruder, S. 2018. *Fine-Tuned Language Models for Text Classification*. arXiv:abs/1801.06146. Ithaca, NY: Cornell University Library.

Khatri, C.; Hedayatnia, B.; Venkatesh, A.; Nunn, J.; Pan, Y.; Liu, Q.; Song, H.; Gottardi, A.; Kwatra, S.; Pancholi, S., et al. 2018. *Advancing The State of the Art in Open Domain Dialog Systems Through the Alexa Prize*.arXiv:1812.10757. Ithaca, NY: Cornell University Library.

Kumar, A.; Gupta, A.; Chan, J.; Tucker, S.; Hoffmeister, B.; and Dreyer, M. 2017. *Just ASK: Building an Architecture for Extensible Self-Service Spoken Language Understanding*. arXiv:1711.00549. Ithaca, NY: Cornell University Library.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. *Albert: A Lite Bert for Self-Supervised Learning of Language Representations*. arXiv:1909.11942. Ithaca, NY: Cornell University Library.

Lavie, A., and Agarwal, A. 2007. Meteor: An Automatic Metric for MT Evaluation With High Levels of Correlation With Human Judgments. Paper presented at the Second Workshop on Statistical Machine Translation, StatMT 2007. Prague, Czech Republic, 23 June. doi.org/10.3115/1626355.1626389.

Lee, H.; Lee, J.; and Kim, T.-Y. 2019. SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5478–83. Stroudsburg, PA: Association for Computational Linguistics doi.org/10.18653/v1/P19-1546.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. *A Persona-Based Neural Conversation Model*.arXiv:1603.06155. Ithaca, NY: Cornell University Library.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A Manually Labelled Multi-Turn Dialogue Dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, 986–95. Stroudsburg, PA: Association for Computational Linguistics.

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 2019. MoEL: Mixture of Empathetic Listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 121–32. Stroudsburg, PA: Association for Computational Linguistics doi.org/10.18653/v1/D19-1012.

Lison, P., and Tiedemann, J. 2016. Opensubtitles2016: Extracting Large Parallel Corpora From Movie and Tv Subtitles. *In* Chair, N. C. C.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA). Luxembourg: European Language Resources Association.

Liu, C.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. *How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation*.arXiv:1603.08023. Ithaca, NY: Cornell University Library.

Logacheva, V.; Burtsev, M.; Malykh, V.; Polulyakh, V.; and Seliverstov, A. 2018. *ConvAI Dataset of Topic-Oriented Human-to-Chatbot Dialogues*. *The NIPS '17 Competition: Building Intelligent Systems*. Berlin, Germany: Springer.

Logacheva, V.; Malykh, V.; Litinsky, A.; and Burtsev, M. 2019. *ConvAI2 Dataset of Non-Goal-Oriented Human-to-Bot Dialogues*. *The NIPS '18 Competition*. Berlin, Germany: Springer.

Lowe, R.; Noseworthy, M.; Serban, I. V.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards An Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Volume 1 (Long Papers), 1116–26. Stroudsburg, PA: Association for Computational Linguistics doi.org/10.18653/v1/P17-1103.

Madotto, A.; Lin, Z.; Wu, C.-S.; Shin, J.; and Fung, P. 2020. *Attention Over Parameters for Dialogue Systems*. arXiv:2001.01871. Ithaca, NY: Cornell University Library.

Miller, A. H.; Feng, W.; Fisch, A.; Lu, J.; Batra, D.; Bordes, A.; Parikh, D.; and Weston, J. 2017. *Parlai: A Dialog Research Software Platform*. arXiv:1705.06476. Ithaca, NY: Cornell University Library.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, 311–8. Stroudsburg, PA: Association for Computational Linguistics.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. *Improving Language Understanding by Generative Pre-Training*. arXiv:1207.0580. Ithaca, NY: Cornell University Library.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–92. Stroudsburg, PA: Association for Computational Linguistics doi.org/10.18653/v1/D16-1264.

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A., et al. 2018. *Conversational AI: the Science Behind the Alexa Prize*. arXiv:1801.03604. Ithaca, NY: Cornell University Library.

Serban, I. V.; Sankar, C.; Germain, M.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Pieper, M.; Chandar, S.; Ke, N. R., et al. 2017. *A Deep Reinforcement Learning Chatbot*. arXiv: 1709.02349. Ithaca, NY: Cornell University Library.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2015. *Hierarchical Neural Network Generative Models for Movie Dialogues*. arXiv:1507.04808. Ithaca, NY: Cornell University Library.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2016. *A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues*. arXiv:1605.06069. Ithaca, NY: Cornell University Library.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. *Attention Is All You Need*. arXiv:1706.03762. Ithaca, NY: Cornell University Library.

Wallace, R. 2009. The Anatomy of A.L.I.C.E. In Epstein, R., Roberts, G., and Beber, G., eds., *Parsing the Turing Test*, 181–210. Berlin, Germany: Springer .

Weissenborn, D.; Wiese, G.; and Seiffe, L. 2017. *FastQA: A Simple and Efficient Neural Architecture for Question Answering*. arXiv:1703.04816. Ithaca, NY: Cornell University Library.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. *Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation*. arXiv:609.08144. Ithaca, NY: Cornell University Library.

Yu, Z.; Xu, Z.; Black, A. W.; and Rudnicky, A. I. 2016. Chatbot Evaluation and Database Expansion via Crowdsourcing. Paper presented at the Second Workshop on Chatbots and Conversational Agent Technologies (WOCHAT), Los Angeles, CA, 20 September. Corpus ID: 12117509.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. *Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?* arXiv:1801.07243. Ithaca, NY: Cornell University Library.

Zhang, Z.; Yang, J.; and Zhao, H. 2020. *Retrospective Reader for Machine Reading Comprehension*. arXiv:2001.09694. Ithaca, NY: Cornell University Library.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2018. *The Design and Implementation of XiaoIce, an Empathetic Social Chatbot*. arXiv:1812.08989. Ithaca, NY: Cornell University Library.

Mikhail Burtsev is head of the Neural Networks and Deep Learning laboratory at the Moscow Institute of Physics and Technology, Moscow, Russia. His research interests center around the application of neural networks and reinforcement learning in the NLP domain. He proposed and co-organized the ConvAI competitions and a series of DeepHack schools-hackathons.

Varvara Logacheva is a research scientist at the Center for Computational and Data-Intensive Science and Engineering at Skolkovo Institute of Science and Technology, Moscow, Russia. Her research interests are machine translation, dialogue systems, and evaluation of text processing systems. She co-organized the ConvAI competitions and shared tasks on Quality Estimation for Machine Translation.[14] Logacheva received her PhD in computer science at the University of Sheffield.