

NIPS 2017 Live Competition Proposal: The Conversational Intelligence Challenge

Mikhail Burtsev* Valentin Malykh Ryan Lowe Iulian Serban
Alexander Rudnicky Alan W. Black Yoshua Bengio
burtcev.ms@mipt.ru

March 15, 2017

1 Overview of the competition

Today, evaluation of dialogue agents is severely limited by the absence of accurate formal metrics. Existing statistical measures such as perplexity, BLEU, recall and others are not sufficiently correlated with human evaluation [1]. Blind assessment of communication quality by humans is a straightforward solution famously proposed by Alan Turing as a test for machine intelligence [2]. Unfortunately, human assessment is time and resource consuming. Here we propose to crowdsource evaluation of dialogue systems in the form of a live competition. Participants of the competition, as well as volunteers, will be asked to perform a blind evaluation of a discussion about a news/wikipedia article with either a bot or a human peer. As a result we expect to have two outcomes: (1) a measure of quality of state of the art dialogue systems compared to human level, and (2) an open source dataset collected from evaluated dialogs.

1.1 Keywords

Dialogue systems, human evaluation, reading comprehension, question answering

1.2 Novelty

The closest analog of the proposed competition is Alexa Prize [3]. This is a competition to build a socialbot that can converse coherently and engagingly with humans on popular topics for 20 minutes. Two socialbots selected by Amazon Alexa customers and another one or few selected by the Amazon panel will compete head-to-head in front of three judges in November 2017. Another small scale analog is the Loebner Prize [4]. It usually takes a

*The lead organizer.

form of a one day event, where a small number of human judges converse in an open-ended manner via textual interface with a chatterbot or another human for a few minutes, and decide whether the peer is a machine or not. There is also Build It Break It [5] competition proposing 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. The key differences of our proposal are:

- a large number of evaluators (up to a few thousands);
- a deliverable: an open source dataset created as a result of the competition and potentially highly valuable for research community (up to 5K-15K dialogs with 5-15 turns per dialog);
- judges produce graded evaluations of dialogs (or additionally of single answers);
- every conversation has a specific goal - to discuss a news/wikipedia (to be decided) article with maximum (1) engagement (e.g. conversation length with humans, ratings by humans) and (2) breadth (e.g. did the conversation cover all events/facts in the article). This partially combines the tasks of reading comprehension and question answering, both of which have baselines and datasets for model pretraining;
- building user simulators for a selected subset of the articles. These will enable people to test their systems easily and automatically before the competition, without having to resort to crowd-sourcing experiments which has been a bottleneck for much work in this domain. The user simulators will also serve as valuable tools for future research.

2 Competition description

2.1 Background and impact

In recent years with the advent of mobile devices, textual channels of communication have become a major medium for information exchange. This creates a pressing demand for automation of conversational interfaces. Last year all leading messaging platforms have opened APIs preparing for the rapid adoption of chat bots. In spite of high expectations, the widespread use of conversational interfaces has not yet happened. Traditional rule-based dialogue systems just cannot handle all the richness of human communication even in constrained domains.

Availability of large volumes of conversational data and growing computational power open possibilities for development of end-to-end neural conversational systems. So, this area attracts more and more interest from the NIPS community and demonstrates promising results. But the problem is still far from being solved. Ongoing research in the field identifies two primary factors hampering the progress in development of neural dialogue systems. These are: (1) a lack of good methods for training, and (2) the related issue

of measuring the quality of dialogue systems [1]. To solve these issues, active research is unfolding to create learning tasks for training end-to-end question answering and dialogue systems including the dialog state tracking challenge [6], live dialogs transcripts [7], the movie dialog dataset [8], Q&A datasets [9] [10][11][12][13], bAbI tasks [14][15], progressive learning of language [16].

With the proposed Live competition, we want to make valuable contributions to these ongoing research efforts by human evaluation of existing dialogue systems and collection of rated dialogs for future training of end-to-end systems.

We believe that the dialogue community would benefit significantly from having a shared environment to measure progress for competing algorithms. Right now, it seems that since the automatic evaluation metrics in dialogue are so poor [1], researchers need to conduct human experiments – but since researchers often work on slightly different datasets or problems, proper comparison to existing methods requires either re-implementing or re-training networks from other groups, which can be very time-consuming. Thus, many researchers usually compare to algorithms that they’ve already implemented, which is most often their own previous work. To remove this bottleneck for progress, having a shared platform to evaluate dialogue agents would be very beneficial.

Collected data should make it possible to analyze how human raters make their decisions. This can help to develop a system that specifically minimizes the clues humans use to detect non-human conversations. The dataset should open new opportunities for training a dialogue evaluation model, for example the ADEM evaluation model proposed in [17]. One could also use this data to train the discriminator of a GAN to distinguish between human responses and machine responses. This has been proposed, for example in [18] [19], but has not yet been implemented at a large scale for dialogue evaluation. Such models should foster further progress in the field by allowing a rapid automatic approximation to human evaluation.

2.2 Tasks and application scenarios

The goal of the competition is to test the ability of a dialogue agent (chat bot) to discuss the content of a news/wikipedia (to be decided) article with a user. The evaluation is performed through a blind cross testing of bots and other users in a series of dialogs. Members of participating teams, as well as volunteers, will log into an anonymous chat system and communicate either with bot or another human user. We expect every human evaluator to complete 10-30 dialogs with the following scenario.

1. Connect randomly with a peer. The peer might be a chat bot or other human user. No information about identity of the peer is provided.
2. Peers receive the text of a news/wikipedia article.

3. Discuss the content of the article with the peer as long as you wish. Evaluate the quality of every response and the dialog as a whole.
4. Choose another news/wikipedia article and/or anonymous peer.

We plan to split competition into two stages.

The goals of the first stage are (1) select the best teams for live competition at NIPS, (2) give teams feedback on the quality of submitted dialogue systems, (3) collect dataset that can be used by teams to finetune bots for the competition at NIPS. At the beginning of the stage in the qualification round teams demonstrate quality of their agents on some fixed dataset. Teams with highest scores are passed to the human evaluation round. During week long NLP summer school-hackathon we run human evaluation of agents submitted by selected teams. This evaluation is performed by team members, school participants and remote volunteers recruited via Competition web page. School participants also organize in teams and compete in building discriminator models to differentiate between human and chat bot answers. The team winning school's hackathon will be awarded with the travel grant to visit NIPS conference. Members of teams from the main track of the Competition are invited for a full participation in the school or giving a remote talk on their research.

The second stage starts with the tuning round when teams can modify models to improve performance on the dataset collected at the human evaluation round. Three weeks prior the NIPS conference submission of new models is closed. Two weeks before the NIPS human evaluation of the final bots begins remotely by team members and volunteers, and then continues through the NIPS as a live competition. During the conference attendees, team members and remote volunteers evaluate competing dialogue systems. At the end of the NIPS results of the competition are presented on the dedicated session. We also propose to extend presentation of results with a satellite workshop where teams discuss implementation of their dialogue systems. The stage is closed with publication of collected dataset under an open source license and submission of the book chapter to the Springer Series on Challenges in Machine Learning: NIPS 2017 Competitions volume.

The competition task can be easily matched to the applications in industry that require discussion of some content with the user:

- presentation of a new product or service;
- support related to some product or service;
- engagement in desired activity;
- entertainment;
- education.

To maximize outcomes of the proposed Live competition the following strategy of presentation on NIPS is suggested:

1. A short presentation of the Live competition at the very first day of NIPS to engage a maximal number of evaluators from NIPS participants for the maximal time. Evaluation will be performed via messaging platform through the whole conference duration.
2. A dedicated landing page with Live Leaderboards for competing dialogue systems and human evaluators. These Leaderboards can be presented on screen in some public space through all the conference.
3. Posters with info on the competition and invitations to volunteer as an evaluator with a link to dedicated channel in a messaging platform.
4. On the last day of the conference a session consisting of (1) a presentation of overall competition and the final results, (2) talks from top 3 best performing teams, and (3) 1-2 famous researchers talk to the bot on a particular topic (e.g. their own biography).

2.3 Metrics and Judging

The quality of dialogs generated by the contest entries should be human-evaluated with values from 0 (bad) to 10 (excellent). The final rating of an entry is calculated as an average over all evaluated dialogs for that entry. Human evaluators will be asked to rate individual responses as well as provide subjective judgement of engagement, breadth and an overall quality of the dialog. Evaluators will be also required to indicate their fluency in English.

We plan to employ three groups of evaluators.

- Members of participating teams. About ten teams to be selected with an average size of 5. Each team will be required to evaluate 150 dialogs per round for 2 human evaluation rounds.
- About 50 students of the Summer school. Each student will be required to evaluate 30 dialogs.
- Volunteers. We expect to recruit about 300-500 volunteers via Competition web page for the first human evaluation round and about 1000 volunteers for the NIPS round. Estimated number of evaluated dialogs per volunteer is 3.

Given these estimates a total number of evaluated dialogs is expected to be about 10 thousand.

2.4 Materials provided

We plan to provide baseline solution pretrained on selected datasets as well as reference to the datasets that can be used for the development of the competition entry such as [7] [8] [10] [11] [12] [13] [14] [15].

2.5 Facilities at NIPS

Requirements to run live competition on site:

- internet access for evaluators;
- info screen in a public space for live leaderboard broadcast (optional);
- space for presentation of final competition results and solutions.

2.6 Data

We expect teams to use open datasets such as [7] [8] [10] [11] [12] [13] [14] [15] for the pretraining of their dialogue systems.

3 Organizational aspects

3.1 Rules

3.1.1 Competition rounds

The competition consists of four rounds.

1. **Qualification round.** Starting from the 1st of May registered participants submit their results on the task associated with dataset X (to be selected). Submission of results is closed on the 1st of July. Top N teams are selected for the Human evaluation round.
2. **Human evaluation round.** Members of selected teams are invited to participate in a week long NLP summer school by giving a talk on their research. Participation can be on site or remote. During the school week members of teams, school participants, and volunteers recruited via the competition web page evaluate the submitted dialogue systems on the competition task. At the end of Human Evaluation Round, up to 10 teams are selected for the NIPS session. Every team is required to evaluate at least 150 dialogs during the Round.
3. **Tuning round.** Dataset of rated dialogs collected during Human Evaluation Round (1K-5K dialogs) is open sourced and can be used by participating teams to tune their solutions.

4. **NIPS round.** Starting two weeks before the NIPS conference teams and volunteers perform evaluation of submitted dialog systems. At the beginning of NIPS the conference participants are invited to volunteer in evaluation of teams entries adjusted over the Tuning Round. Final rating of submissions is presented on the Competition session at NIPS.

3.1.2 Task

Both human evaluators and dialogue agents complete the same task.

1. Connect randomly with a peer. The peer might be a chat bot or other human user. No information about identity of the peer is provided.
2. Both parties are given a text of a recent news/wikipedia article.
3. Discuss content of the article with the peer as long as you wish.
4. Choose another news/wikipedia article and/or anonymous peer.

3.1.3 Evaluation

1. Evaluator will not be given any information about identity of the peer.
2. Members of the team will be automatically excluded from evaluation of their own submission and each other.
3. The quality of every response is subjectively evaluated on the 0 to 10 range.
4. The quality of the dialog as a whole as well as its breadth and engagement are evaluated on the 0 to 10 range.
5. Final rating is calculated as an average of evaluation values accumulated by submission during the NIPS Round of Competition.

3.1.4 Technical infrastructure

1. Competitors will provide their solutions in the form of executable source code supporting a common interface (API).
2. These solutions will be run in isolated virtual environments (containers).
3. The solutions will not be able to access any external services or the Internet, and will only be able to communicate with the supervisor bot to guard against cheating.

4. The master bot will facilitate communication between human evaluators and the competitors solutions. It will be available in popular messenger services (Facebook/Telegram). Its main function will be to connect a participant to a (randomly selected) solution or peer and log the evaluation process.
5. The master bot will provide the instructions and a context necessary for human evaluation of presented solutions.

3.1.5 Dataset

Dataset collected during competition will be distributed under open source license.

3.1.6 Discussion of the rules

Two major goals of proposed competition are (1) the evaluation of current state of the art in the area of dialogue systems and (2) the collection of dataset of rated dialogs. To maximize progress towards these goals we split our competition into four rounds. During Qualification Round teams that will demonstrate high scores on the related task will be selected to test their bots in real settings during Human Evaluation Round. As a result of this stage teams will be rated for the final competition at the NIPS and preliminary dataset of evaluated dialogs will be collected along the way. To improve quality of submissions for NIPS live competition teams will have an access to the dataset collected. Two weeks prior to the NIPS conference teams and volunteers will start to evaluate final submissions. This stage will guarantee statistically significant rating of teams for the final NIPS session. Over the course of the conference we will try to engage as many people as possible in the evaluation.

Evaluation is proposed to take a form of chat to discuss some text. Evaluator log into one of the popular messaging platforms (to be selected) and then asks the competition master bot to start evaluation session. Master bot returns text and connects the evaluator to anonymous human or machine peer to initiate discussion. When dialog is finished the user sends ratings to the master bot. This allows unbiased evaluation of the dialog quality. To safeguard against cheating every bot will be run in the virtual environment with the external connection to the master bot only.

3.2 Schedule

3.2.1 Prior to NIPS

10th of April **Competition promotion is started.** Call for participation is published on the dedicated web page and disseminated over social media. Volunteers for evaluation are invited to sign up. Competition is promoted at the upcoming conferences: ICLR, IJCNN, etc.

- 10th of April* **Registration is open.** Registered teams are allowed to submit dialogue agents for the Qualification round.
- end of June* **Qualification round is closed.** Qualification Leaderboard is published. Top 10 teams are selected for the Human evaluation round and NIPS session.
- 1st half of August* **Human evaluation round.** One week NLP summer school. Teams, school participants and volunteers evaluate entries of selected teams.
- 1st of September* **1st Dataset is published.** Data collected at Human evaluation round is published.
- 1st of September* **Tuning round.** Selected teams tune their solutions on the 1st Dataset and prepare papers for NIPS. Three weeks prior the NIPS conference submission of new models is closed.

3.2.2 At NIPS

- 1st Day* Short promotional talk presenting the goal of Competition, teams selected, results of summer Human evaluation round and requesting for participation in final NIPS round of Competition.
- during NIPS* Human evaluation of competition entries. Promotion of the Competition to recruit as many evaluators as possible.
- Last Day* Session with 3-5 talks and poster presentations. Announcement of winners. Winners present their dialogue systems. 1-2 famous researchers chat with the best bot.

3.3 Competition promotion

Outline of the competition promotion campaign.

- Creation of the dedicated landing web page with calls, competition rules and registration forms for teams and volunteers.
- Regular announcements via social media such as Facebook, Google+, twitter and mailing lists. Here we plan to actively involve key influencers from the organizing team and beyond.
- Posters and announcements on the conferences such as ICLR, IJCNN, ACL and others.

3.4 Prizes

We plan to have a number of prizes in different categories.

- Major nomination for the dialog agent with highest overall rating.
- Major nomination for the human evaluator with highest number of rated turns.
- Nomination for the best discriminator developed at the Summer School. Prize is a travel grant to NIPS conference.
- Nomination for the human evaluator with the highest rating. (optional)
- Nomination for the dialog agent with highest engagement rating. (optional)
- Nomination for the dialog agent with highest breadth rating. (optional)

We had no sufficient time to negotiate sponsors participation but started to reach out to potential partners. At the moment we have preliminary agreement with Sberbank of Russia for support of major prize nominations. Due to the large industrial interest in a dialogue automation and our previous experience we expect no problems to secure more funds.

3.5 Organizing team

Mikhail Burtsev - Head of Neural Networks & Deep Learning lab at MIPT, Moscow. His current research interests: application of neural nets and reinforcement learning in the NLP domain, neuroevolutionary and nurodevelopmental methods. He recently proposed and organized a series of DeepHack science schools-hackathons [20][21][22].

Valentin Malykh - Ph.D. student at the Neural Networks & Deep Learning lab at MIPT, Moscow under the supervision of Mikhail Burtsev. He has an extensive experience as NLP and IR developer in industry (Yandex, etc.). Research interests: dialog systems, reinforcement learning. He is one of winners of the first DeepHack hackathon [22] and co-organizer of the third DeepHack event [20].

Ryan Lowe - Ph.D. student in Computer Science in the Reasoning & Learning Lab at McGill University, Montreal under the supervision of Joelle Pineau. Research interests: dialog systems, generative models. Co-organizer of the Montreal AI Ethics Group to discuss technical issues related to AI safety with machine learning researchers from McGill and University of Montreal.

Iulian Serban - Ph.D. student at University of Montreal jointly supervised by Aaron Courville and Yoshua Bengio. His research interests: the generative deep learning models and reinforcement learning methods with the focus on dialogue modeling and question answering.

Alexander Rudnicky - Research Professor at the Computer Science Department in the School of Computer Science at Carnegie Mellon University. He is a part of the Carnegie Mellon Speech Group and serves as the Director of the Carnegie Mellon Speech Consortium. Also he is part of the faculty in the Language Technologies Institute. His research interests: language-based communication between humans and robots and aspects of core speech recognition, such as out-of-vocabulary (OOV) word processing.

Alan W. Black - Full Professor in the Language Technology Institute at Carnegie Mellon University, Prof. Black has over 170 refereed published papers in many aspects of speech and language technologies and has served on the committees of many international conferences and workshops. His research interests are dialog systems and speech synthesis. With Prof. Tokuda (Nagoya Institute of Technology) initiated the annual (2005 - till present) Blizzard Challenge [23] evaluation for corpus based synthesis techniques, this is largest speech synthesis evaluation forum and involves academia and industry around the world. Prof. Black was the organizer for the Spoken Dialog Challenge 2010 [24] and co-organizer of the first Dialog State Tracking Challenge ([25], 2013).

Yoshua Bengio - Full Professor of the Department of Computer Science and Operations Research, head of the Montreal Institute for Learning Algorithms (MILA), CIFAR Program co-director of the CIFAR program on Learning in Machines and Brains, Canada Research Chair in Statistical Learning Algorithms. His main research ambition is to understand principles of learning that yield intelligence. Yoshua Bengio was Program Chair for NIPS' 2008 and General Chair for NIPS' 2009. Since 1999, he has been co-organizing the Learning Workshop with Yann Le Cun, with whom he has also created the International Conference on Representation Learning (ICLR). He has also organized or co-organized numerous other events, principally the deep learning workshops and symposia at NIPS and ICML since 2007.

References

- [1] Liu, Chia-Wei, et al. "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation." arXiv preprint arXiv:1603.08023 (2016).
- [2] Turing, Alan M. "Computing machinery and intelligence." *Mind* 59.236 (1950): 433-460.
- [3] <https://developer.amazon.com/alexaprize>
- [4] https://en.wikipedia.org/wiki/Loebner_Prize
- [5] <https://bibinlp.umi.acs.umd.edu/>
- [6] Williams, Jason, Antoine Raux, and Matthew Henderson. "The dialog state tracking challenge series: A review." *Dialogue & Discourse* 7.3 (2016): 4-33.

- [7] Lowe, Ryan Thomas, et al. "Training end-to-end dialogue systems with the ubuntu dialogue corpus." *Dialogue & Discourse* 8.1 (2017): 31-65.
- [8] Dodge, Jesse, et al. "Evaluating prerequisite qualities for learning end-to-end dialog systems." *arXiv preprint arXiv:1511.06931* (2015).
- [9] Serban, Iulian Vlad, et al. "A survey of available corpora for building data-driven dialogue systems." *arXiv preprint arXiv:1512.05742* (2015).
- [10] Yang, Yi, Wen-tau Yih, and Christopher Meek. "WikiQA: A Challenge Dataset for Open-Domain Question Answering." *EMNLP*. 2015.
- [11] Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint arXiv:1606.05250* (2016).
- [12] Nguyen, Tri, et al. "MS MARCO: A Human Generated MACHine Reading COMprehension Dataset." *arXiv preprint arXiv:1611.09268* (2016).
- [13] <https://data.quora.com/>
- [14] Weston, Jason, et al. "Towards ai-complete question answering: A set of prerequisite toy tasks." *arXiv preprint arXiv:1502.05698* (2015).
- [15] Bordes, Antoine, and Jason Weston. "Learning end-to-end goal-oriented dialog." *arXiv preprint arXiv:1605.07683* (2016).
- [16] Baroni, Marco, et al. "CommAI: Evaluating the first steps towards a useful general AI." *arXiv preprint arXiv:1701.08954* (2017).
- [17] Lowe, Ryan Thomas, et al. "Towards an automatic Turing test: Learning to evaluate dialogue responses." *arXiv preprint* (2017).
- [18] Li, Jiwei, et al. "Adversarial Learning for Neural Dialogue Generation." *arXiv preprint arXiv:1701.06547* (2017).
- [19] Kannan, Anjuli, and Oriol Vinyals. "Adversarial evaluation of dialogue models." *arXiv preprint arXiv:1701.08198* (2017).
- [20] <http://rl.deephack.me>
- [21] <http://qa.deephack.me>
- [22] <http://game.deephack.me>
- [23] <http://www.festvox.org/blizzard/>
- [24] <http://dialogrc.org/sdc>

- [25] Williams, Jason, et al. "The dialog state tracking challenge." Proceedings of the SIG-DIAL 2013 Conference. 2013.