# Competition on Automatic Evaluation of Dialogue Quality

**Mikhail Burtsev** **Varvara Logacheva** **Valentin Malykh**
**Vadim Poluliakh** **Aleksandr Seliverstov**
Neural Networks and Deep Learning Lab
Moscow Institute of Physics and Technology, Russia
{burtcev.ms, logacheva.vk, poluliakh.vv}@mipt.ru,
valentin.malykh@phystech.edu, seliverstov.a@gmail.com

## Abstract

We present a competition on automatic evaluation of chatbots. This is a challenging task which has not been solved with acceptable quality. In order to stimulate research in this field we organised a hackathon where multiple teams submitted their solutions. Unlike the majority of previous works, we perform evaluation of quality at the level of dialogues.

## 1 Introduction

Evaluation is an essential part of development of machine learning (ML) methods and any systems that use ML. Evaluation is particularly challenging when dealing with tasks whose output is a text in natural language (e.g. machine translation, summarisation, image captioning). There exist metrics that evaluate generated text by comparing it with gold standard text: e.g. BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) for machine translation, ROUGE (Lin, 2004) for summarisation. However, they do not suit for the evaluation of dialogue systems, especially chatbots (Liu et al., 2016). The reason is that the space of possible answers at any point in a dialogue is so wide that one cannot list all of them, so gold standard is inherently incomplete.

It was suggested to evaluate dialogue systems using a system that is trained to predict the quality of bot's answer by analysing the answer itself and previous utterances of a dialogue (Lowe et al., 2017; Dušek et al., 2017). It is expected to be more reliable than conventional metrics. Unfortunately, at present such systems still have too low correlation with human judgements. In order to find new ideas for this problem we organised a hackathon where the participants' task was to create a system for the evaluation of human-to-bot dialogues.

## 2 Hackathon "Deephack.Turing"

The hackathon[1] is a spinoff of ConvAI challenge[2] — a competition of chatbots which is conducted via human evaluation. Volunteers have conversations with bots and evaluate their performance. As a byproduct of this evaluation we acquire a dataset of human-to-bot conversations labelled for quality. This dataset better suits for training of evaluation system because it contains real human-to-bot dialogues of different quality, whereas previous solutions had to use human-to-human dialogues and mimic bad answers by selecting them randomly (Lowe et al., 2016).

The hackathon lasted for five days. Participants were developing their systems and attending lectures on topics related to dialogue systems. Besides that, they were collecting training data for dialogue evaluation (having dialogues with bots). The dialogues were rated at both utterance and dialogue levels, but the task was to predict only dialogue-level score (in terms of 1-to-5 scale). This is different from previous systems (Lowe et al., 2016; Lowe et al., 2017) which predicted utterance-level quality. We decided to predict scores of entire dialogues because we supposed that this task is easier: larger context is available and human evaluation is less subjective.

Note that the majority of dialogues were human-to-bot, but we also collected a small number of human-to-human dialogues. When classifying dialogues teams were not told which of dialogue participants is a bot. Therefore, they predicted dialogue quality from the point of view of each participant, which gives two scores per dialogue (since bots do not evaluate their human peers, they are considered to give a score of 0).

[1] http://turing.tilda.ws/
[2] http://convai.io/

## 3 Results

The participants of the hackathon collected 2,778 dialogues (2,029 used for training and 749 for testing) and submitted 13 dialogue evaluation systems. We prepared three baselines: **random** selection of labels, **SVM** regression which uses dialogue statistics (e.g. length of dialogues, number of utterances from each user, etc.), and hierarchical RNN encoder (**HRE**) (Sordoni et al., 2015).

The results are given in table 1. Baseline systems performed well, despite their simple architecture. The training data was scarce and insufficient for neural network architectures preferred by the majority of teams. In such cases the careful feature engineering (as in SVM baseline) can be more effective. Likewise, the winning system used a big number of dataset-specific features and was trained with an out-of-the-box ML method. Other participants used neural networks: RNNs or CNNs, also combined with hand-crafted features. In the majority of cases the combinations of different models worked better than individual models.

| Team | Spearman's $r$ |
|---|---|
| • Conundrum | 0.772 |
| Turing Quest | 0.738 |
| newbies+ | 0.723 |
| Plastic world | 0.721 |
| I have no mouth and I must chat | 0.702 |
| **HRE baseline** | 0.677 |
| XL-shell | 0.649 |
| DATA Siegt | 0.644 |
| **SVM baseline** | 0.592 |
| Warp Drive | 0.570 |
| Agent Smith | 0.564 |
| TEAM | 0.483 |
| StackingOverflow | 0.358 |
| **Random baseline** | 0.01 |
| fattakhov | 0.004 |
| chatme | -0.06 |

Table 1: Quality of bot evaluation systems. The winning system is marked with •. Systems in gray area are not significantly different from baselines.

Many teams also explicitly classified each user as a human or a bot and then predicted their quality. Those not performing classification as a separate stage still used features that implicitly divided users into bots and humans (e.g. whitespace in front of a punctuation mark occurs mainly in bots' answers, whereas typos are a human feature).

| Rating | Team | Spearman's $r$ |
|---|---|---|
| 1-2 | DATA Siegt | 0.498 |
| 1-3 | Agent Smith | 0.458 |
| 2-4 | I have no mouth | 0.432 |
| 3-4 | Turing Quest | 0.388 |
| | **HRE baseline** | 0.364 |
| | **SVM baseline** | 0.311 |
| 5-6 | newbies+ | 0.276 |
| 5-9 | Conundrum | 0.249 |
| 6-10 | Warp drive | 0.191 |
| 6-10 | chatme | 0.189 |
| 6-10 | Stacking Overflow | 0.184 |
| 6-10 | XLshell | 0.160 |
| 11-12 | fattakhov | 0.004 |
| 11-12 | TEAM | -0.009 |
| | **Random baseline** | -0.04 |
| 13 | Plastic world | -0.201 |

Table 2: Performance of evaluation systems on user scores (zero scores by bots omitted).

However, bot detection task is less important in real-world scenarios: when assessing bot quality and user experience we usually already know who is a bot. Therefore, we also compared the performance of systems only on human scores (see table 2). This task turned out to be more challenging: the highest achieved correlation with human judgements is 0.5. Moreover, the winning systems from table 1 did not perform well in it — this suggests that features used by them (in particular, features that helped detecting bots) are not useful for dialogue evaluation. On the other hand, the majority of teams that explicitly solved bot detection task reported good results in it.

## 4 Conclusion

We conducted a competition of systems that evaluate quality of chatbots at the dialogue level. We are the first to introduce this task, whereas previous work considered evaluation at the utterance level. We suggested that dialogue-level evaluation should be easier.

The winning teams showed high correlation with human judgements, but these are results of joint bot detection and quality prediction tasks. When regarded on its own, quality prediction turned out to be more challenging. Bot detection itself proved to be quite easy. However, teams used many dataset-specific features which will probably not be effective on different data.

## Acknowledgments

The hackathon organisers are grateful to lecturers who gave talks at the hackathon: Danica Damljanovic, Pyry Takala, Jason Weston, Zhou Yu, Pei-Hao Su, Rodrigo Nogueira, Konstantin Vorontsov, Peter Schlecht, Maksim Kretov, Tomas Mikolov. The official partners of the hackathon are MTS, HostKey, 1C, Flint Capital. The work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F630002.

## References

Ondej Dušek, Jekaterina Novikova, and Verena Rieser. 2017. Referenceless Quality Estimation for Natural Language Generation. In *ICML-2017: 34th International Conference on Machine Learning, 1st Workshop on Learning to Generate Natural Language (LGNL 2017)*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation.

Ryan Lowe, Iulian V. Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the Evaluation of Dialogue Systems with Next Utterance Classification. *Sigdial*, page 5.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. *Acl*, pages 1–19.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob G. Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder For Generative Context-Aware Query Suggestion.