*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 508–512
Marseille, 11–16 May 2020


# Treating Dialogue Quality Evaluation as an Anomaly Detection Problem

**Rostislav Nedelchev[1], Jens Lehmann[1,2], Ricardo Usbeck[1,2]**
[1]Smart Data Analytics Group, University of Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin and Dresden, Germany
rostislav.nedelchev@uni-bonn.de, jens.lehmann@cs.uni-bonn.de
ricardo.usbeck@iais.fraunhofer.de

**Abstract**

Dialogue systems for interaction with humans have been enjoying increased popularity in the research and industry fields. To this day, the best way to estimate their success is through means of human evaluation and not automated approaches, despite the abundance of work done in the field. In this paper, we investigate the effectiveness of perceiving dialogue evaluation as an anomaly detection task. The paper looks into four dialogue modeling approaches and how their objective functions correlate with human annotation scores. A high-level perspective exhibits negative results. However, a more in-depth look shows limited potential for using anomaly detection for evaluating dialogues.

**Keywords:** Dialogue, Evaluation Methodologies, Discourse Annotation, Representation, and Processing

## 1. Introduction

Recently, machine-learning powered dialogue systems have been gathering much attention from industry and academia alike (Chen et al., 2017). These systems have applications in various contexts, starting from personal speech assistants like Amazon Alexa or Apple Siri, through the "chatbots" on instant messaging platforms like Skype or Slack, and finally, conversational services like Wit.ai and Dialogflow that allow themselves deployed in various situations. While the majority of these systems have the purpose of completing a specific task like purchasing a product, booking service (e.g., hotel, flight), they can still benefit from conversational skills that are open-domain, for example, the ability to chit-chat to allow natural dialogues.

Nowadays, researchers and developers who work on dialogue systems rely mostly on human annotators to evaluate the quality of a conversation (Dinan et al., 2019; Logacheva et al., 2018; Yoshino et al., 2019). This can be very costly in terms of resources. Thus, the research and development of these systems could benefit significantly from an automated approach that can evaluate conversations.

Earlier works in machine translation and text summarization have developed automated measures for evaluation - for the former, BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) and, for the latter, ROUGE (Lin, 2004). These are also adopted by works researching dialogue systems (Ritter et al., 2011; Serban et al., 2016; Yoshino et al., 2019). However, Liu et al. show that word overlap metrics are not reliable for evaluating the quality of dialogues (2016). Thus, more advanced approaches are needed that consider the context and semantics of a dialogue.

Human annotators distinguish low from high-quality dialogues similarly to anomaly detection. Conversations generated from computer systems can appear to human annotators as very unusual, i.e., an outlier or an anomaly. Their perception is based on extensive conversational experience with real people, rather than using an explicit reference that helps to determine what is correct or wrong.

Thus, the main contribution of this effort is to investigate whether dialogue modeling approaches used for dialogue systems can detect anomalous conversations in contrast to normal ones. To the best of our knowledge, this is the first paper that attempts solving dialogue evaluation by treating it as an anomaly detection problem.

## 2. Related Work

### 2.1. Dialogue Evaluation

Lowe et al. (2017) propose a work that approximates human judgment using scored dialogues together with the context, reference response, and an utterance generated by a dialogue system. Reference responses and human annotation scores are hard to obtain and renders the approach difficult to use on a scale. Tao et al. (2018) propose a method consisting of two components: 1) a score capturing the resemblance between a generated as well as reference response using word vector pooling and 2) a neural network that evaluates relatedness of a reply given the context using only negative sampling from real dialogues. The first component also uses reference responses, which could also be hard to acquire. Both approaches lack the interpretability of the scores that they output regarding different aspects of dialogue like coherency or fluency.

The Dialogue Breakdown Detection Challenges (DBDC) (Higashinaka et al., 2017; Higashinaka et al., 2019) aim at detecting whether during a conversation one of the utterances causes a breakdown, i.e., a scenario where the participant is not able to continue with the dialogue.

Larson et al. (2019) propose outlier detection to detect erroneous utterances within a dialogue for clean data annotation in an NLP dataset. The approach averages word embedding of a reply's content to obtain an utterance level representation. After that, the second stage clusters the vectors, and

the top-$N$[1] are considered anomalous. The approach provides no dialogue-level information about the coherency of the conversation and does not offer a replacement for human annotators.

### 2.2. Anomaly detection

Anomaly detection, very commonly also outlier or novelty detection, deals with the problem of finding instances of data that do not belong to the regular pattern like most of the others (Chandola et al., 2009).

There is a long list of works in NLP that have considered using anomaly detection for discovering incorrect annotations (Hollenstein et al., 2016; Guthrie et al., 2008; Larson et al., 2019). Most of them use handcrafted features to solve the problem.

In the field of deep learning, autoencoders found usage in significant amounts of research to solve problems from various domains. According to Chalapathy et al. (2019), they are at the core of all unsupervised neural-network-based anomaly detection methods. They have found application in a wide variety of domains like intrusion or malware detection, bank, or insurance fraud. Autoencoders learn to create another representation of data (usually, one of lower dimension) and then reconstruct from it the original input. Their effectiveness is measured using a reconstruction error. Thus, on examples that an autoencoder has observed and trained on, it has a lower reconstruction rate. At the same time, on rare or not-previously seen samples, it will exhibit a consistently higher error.

## 3. Methodology

### 3.1. Dialogue Modelling

To investigate the usability of anomaly detection for dialogue evaluation, we consider four neural network models for dialogue modeling. These approaches tackle conversations by first encoding the input context and using that representation by decoding it into the response. *While this is not the same as autoencoders, we can use the loss measuring the correctness of mapping the context to the reply in the same manner as a reconstruction loss.* In this subsection, we concisely present the models used for this study. For more detail on each of the approaches, we would forward the reader to the appropriate reference, during each of their presentations.

The first model we consider is a recurrent sequence-to-sequence approach, as described by Vinyals and Le (2015). It models a dialogue as a sequence of pairs of query and response, i.e., it considers a response as related only to the last utterance before it. The context is encoded using a recurrent neural network (RNN), and another RNN decodes the representation into the response. Cross-entropy acts as a reconstruction loss measuring how well the utterance maps to the context.

Next is Hierarchical Recurrent Encoder-Decoder (HRED) by Serban et al. (2016), which builds upon the sequence-to-sequence (Seq2Seq) approach by considering multiple utterances from the context. It does so by using a third RNN. The context utterances are each encoded using an RNN, and then encoded together one vector representation by the additional RNN. The rest is as in the sequence-to-sequence approach described earlier.

Thirdly, Serban et al. (2017) propose an extended version of HRED, a Hierarchical Latent Variable Encoder-Decoder (VHRED), by adding a latent variable at the decoder that parametrizes the context. Kullback-Leibler (KL) divergence provides measures of the reconstruction between the original context representation and its latent variable version. This way, the approach can model hierarchically-structured sequences in a two-step generation process-first sampling the latent variable, and then generating the output sequence-while maintaining long-term context.

Finally, Park et al. (2018) report that VHRED sufferers from a degeneration of the latent variable, which renders the model to behave almost like an HRED. They introduce a global conversation latent variable such that it is responsible for generating each of the utterances of the dialogue rather than capturing the whole context post-factum.

To train all the models, we use the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011). It has 220,579 conversations and a total of 304,713 utterances. The training is done by iterating over each dialogue turn and considering the full query context. The first sequence-to-sequence approach is using only the last dialogue turn as a context.

### 3.2. Dialogue Datasets

| Feature | ConvAI1 | ConvAI2 |
|---|---|---|
| # Dialogues | 2154 | 2237 |
| Avg # Utterances | 13.9 | 18.1 |
| Avg # Words per Utterance | 7.3 | 8.2 |
| Task | Topic discussion | Person impersonation |

Table 1: Key features of the dialogue datasets. Only dialogues with three or more utterances were considered as part of this work.

For evaluating the usefulness of anomaly detection to indicate the quality of dialogues, we use the results of the ConvAI1[2] (Burtsev et al., 2018; Logacheva et al., 2018) and ConvAI2[3] (Zhang et al., 2018; Dinan et al., 2019) challenges. Participants had to create dialogue systems that had to fulfill specific criteria. For the former, the systems had to be able to discuss a topic conversationally. For the latter, the dialogue systems had to engage in a chit-chat dialogue by impersonating a personality profile ("persona"). In both challenges, the dialogues of the participating systems have had their quality assessed by human evaluators.

### 3.3. Scoring

*As presented in 3.1., the cross-entropy loss function will act as a reconstruction loss to detect anomalies*. For obtaining the scores, the dialogues presented in subsection 3.2. go

[1] $N$ is a hyperparameter

[2] http://convai.io/2017/data/

[3] http://convai.io/data/

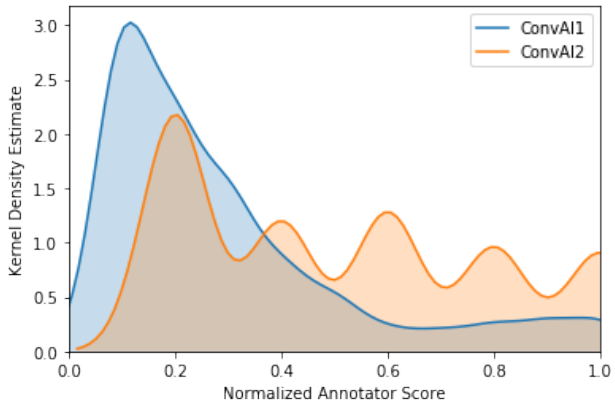


Figure 1: Kernel density estimation of the distribution of annotator scores of the dialogues in ConvAI1 and ConvAI2. We see that the majority of dialogues are evaluated as low quality.

through the same iterative manner described in subsection 3.1.. After that, the scores are averaged on the dialogue level to obtain a single value that summarizes the whole conversation.
Cross-entropy is defined as:

$$L = \frac{1}{T}\sum_{t=1}^{T} l_t$$
$$l_t = -w_r(\sum_{v=1}^{V} y'_v \log(y_v)) \quad (1)$$

where $t$ stands for the $t$-th token in the response, $y'_v$, and $y_v$ are the true and the predicted words from the vocabulary ($V$), respectively, $w_r$ are weights used for ignoring padding tokens in a sequence. All of the scores obtained from a single model applied to a dataset undergo a rescaling such that the maximum will have a value of 1.0.

## 4. Evaluation

In this section, we will analyze the dialogue datasets, ConvAI1 separately, and ConvAI2, for possible correlations between the cross-entropy values exhibited from each of the models and the respective annotator score. The results are summarized in Table 2.

| Dataset | ConvAI1 | | ConvAI2 | |
|---|---|---|---|---|
| | $r$ | $\rho$ | $r$ | $\rho$ |
| **Seq2Seq** | 0.2150 | 0.3006 | 0.3444 | 0.4892 |
| **HRED** | 0.1869 | 0.2832 | 0.3469 | 0.4876 |
| **VHRED** | 0.2210 | 0.3009 | 0.3384 | 0.4885 |
| **VHCR** | 0.2249 | 0.3037 | 0.3408 | 0.4888 |

Table 2: Pearson's correlation coefficients, $r$, and Spearman's correlation coefficients, $\rho$, on the two dialogue datasets' human scores and cross-entropy scores. All of the scores are with a confidence of $p <= 0.0001$

The first immediate observation is that all of the models across the two datasets demonstrate a significant positive correlation with the scores from the human annotators. The result is contrary to the initial expectation for the following reason. Cross-entropy measures the models' ability to reconstruct a response from the given query context. Thus, the higher the loss function's value is, the more difficult it is for the model to relate the input to the response. The positive correlation states that as the annotator's score increases, so does the cross-entropy. Ideally, the correlation between the two variables should be negative, since the models used training data with proper examples and, thus have difficulties to process anomalous conversations from dialogue systems. Then, the outlier exchanges will be lowly evaluated by the human annotators, and the models should have a comparatively higher loss score.
Furthermore, all of the approaches appear to have a shared understanding and perspective of the conversations because they are demonstrating a very similar correlation with the annotators' scoring. The sequence-to-sequence approach is also on par with the others, which is noteworthy because unlike the others, it cannot capture long-term dependencies in dialogues. Thus, long-term context appears to be not necessary for the scoring of these dialogues by the annotators.
We see that in Figure 1 that the dialogue scores by the annotators have a non-uniform distribution. Thus, we set to investigate if there are any patterns within the various quality subgroups. For that purpose, we split the dialogues into five equal-width bins based on the minimum (0.0) and maximum (1.0) values for the human annotator scores. All of the sub-groups that exhibit somewhat negative correlation coefficients are in Table 3.

| Model | Dataset | Quality Range | $r$ ($p \leq$) | $\rho$ ($p \leq$) |
|---|---|---|---|---|
| **Seq2Seq** | ConvAI1 | $[0.4, 0.6)$ | 0.0141 (0.8087) | -0.0513 (0.3791) |
| **Seq2Seq** | ConvAI1 | $[0.6, 0.8)$ | -0.0093 (0.9309) | 0.0941 (0.3776) |
| **Seq2Seq** | ConvAI2 | $[0.8, 1.0]$ | -0.0093 (0.9309) | -0.0093 (0.3791) |
| **HRED** | ConvAI1 | $[0.4, 0.6)$ | 0.0145 (0.8039) | -0.0514 (0.3783) |
| **HRED** | ConvAI2 | $[0.8, 1.0]$ | -0.2493 (0.0001) | -0.2778 (0.0001) |
| **VHRED** | ConvAI1 | $[0.4, 0.6)$ | 0.0093 (0.8737) | -0.0546 (0.349) |
| **VHRED** | ConvAI1 | $[0.6, 0.8)$ | -0.0097 (0.9279) | 0.0984 (0.3562) |
| **VHRED** | ConvAI2 | $[0.8, 1.0]$ | -0.2613 (0.0001) | -0.282 (0.0001) |
| **VHCR** | ConvAI1 | $[0.4, 0.6)$ | 0.0106 (0.8559) | -0.0507 (0.3843) |
| **VHCR** | ConvAI1 | $[0.6, 0.8)$ | -0.0196 (0.8546) | 0.0958 (0.3689) |
| **VHCR** | ConvAI2 | $[0.8, 1.0]$ | -0.2609 (0.0001) | -0.2841 (0.0001) |

Table 3: Selected sub-groups with negative correlation coefficients. The omitted groups have positive correlations aligned with the results from Table 2.

For the dialogues in ConvAI1, we discover that all of the models exhibit a very weak negative correlation in the quality scores between $0.4$ and $0.8$. The considerably lower amount of examples in the groups with higher quality contributes to low confidence estimates. Nevertheless, this discovery hints that there is limited potential in using anomaly detection for dialogue quality evaluation.

Meanwhile, for the conversations in ConvAI2, we identify stronger than in ConvAI1 negative correlations with the top-most in terms of quality samples. The dialogues in the quality range between $0.8$ and $1.0$ have negative Pearson's and Spearman's correlation coefficients. These samples provide further evidence to the potential of having an anomaly detection perspective on the issue.

## 5. Conclusion

On a high level, we saw that the method is unfit for replacing human annotators. However, when we consider only various quality sub-groups of the data, the models demonstrate an expected negative correlation and show some promise for using their loss function outputs for detecting anomalous conversations.

Overall, the limited ability to generalize or, otherwise, the insignificant amount of training data are obstacles for using outlier detection methods for evaluating dialogues. As future work, we would focus in this direction, so that models can better generalize and be able to demonstrate consistent behavior across various domains, thus, successfully assessing dialogue quality.

## 6. Bibliographical References


Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Burtsev, M., Logacheva, V., Malykh, V., Serban, I. V., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., and Bengio, Y. (2018). The first conversational intelligence challenge. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46. Springer.

Chalapathy, R. and Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15.

Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, pages 76–87. Association for Computational Linguistics.

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). *arXiv preprint arXiv:1902.00098*.

Guthrie, D., Guthrie, L., and Wilks, Y. (2008). An unsupervised approach for the detection of outliers in corpora. *Statistics*, pages 3409–3413.

Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., and Kaji, N. (2017). Overview of dialogue breakdown detection challenge 3. *Proceedings of dialog system technology challenge*, 6.

Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., and Kaji, N. (2019). Overview of the dialogue breakdown detection challenge 4. *Proceedings of Tenth International Workshop on Spoken Dialogue Systems Technology*.

Hollenstein, N., Schneider, N., and Webber, B. (2016). Inconsistency detection in semantic annotation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3986–3990.

Larson, S., Mahendran, A., Lee, A., Kummerfeld, J. K., Hill, P., Laurenzano, M. A., Hauswald, J., Tang, L., and Mars, J. (2019). Outlier detection for improved data quality and diversity in dialog systems. 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)*, pages 517–527.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, C.-W., Lowe, R., Serban, I., 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Logacheva, V., Burtsev, M., Malykh, V., Polulyakh, V., and Seliverstov, A. (2018). Convai dataset of topic-oriented human-to-chatbot dialogues. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 47–57. 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)*, pages 1116–1126.

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*, pages 311–318. Association for Computational Linguistics.

Park, Y., Cho, J., and Kim, G. (2018). A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natu-*

*ral language processing*, pages 583–593. Association for Computational Linguistics.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Tao, C., Mou, L., Zhao, D., and Yan, R. (2018). Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Yoshino, K., Hori, C., Perez, J., D'Haro, L. F., Polymenakos, L., Gunasekara, C., Lasecki, W. S., Kummerfeld, J. K., Galley, M., Brockett, C., et al. (2019). Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.