AUTOMATIC USER EXPERIENCE EVALUATION OF GOAL-ORIENTED DIALOGS USING PRE-TRAINED LANGUAGE MODELS

Mika Rebensburg[1], *Stefan Hillmann*[1], *Nils Feldhus*[2]

[1]*Technische Universität Berlin,* [2]*DFKI Berlin*
m.rebensburg@campus.tu-berlin.de

Abstract: Dialog evaluation methods based on Pre-trained Language Models (Pr-LMs) have been primarily used for open-domain dialogs with the goal of comparing systems in terms of dialog skills relevant in casual chats, such as naturalness, engagement, and relevance. Automatic evaluation metrics for goal-oriented and closed-domain dialogs often measure few and objective metrics like task success rate and ignore subjective aspects of the User Experience (UX). Important subjective usability aspects like satisfaction go beyond simple objective metrics and have traditionally been assessed using questionnaires in an experimental setup. Information about subjective UX is often implicitly contained in the dialog text which could therefore be used to estimate the true UX in an automated fashion using Pr-LMs. This works aims to explore automatic text-based and multifaceted UX evaluation of goal-oriented chatbot interactions using Pr-LMs. We examine both a supervised learning approach and an approach based on an automatic, reference-free and unsupervised dialog evaluation metric. With supervised learning, we train a Pr-LM that predicts several relevant UX aspects with moderate correlation values. SimCSE embeddings perform best and even outperform the UX ratings of human observers collected in a previous study. While the reference-free approach manages to achieve low to moderate correlations, we suspect that this method mainly exploits the correlation between dialog length and user satisfaction and could hence fail in scenarios where these are not correlated.

1 Introduction

Evaluating goal-oriented dialog systems based on UX can be challenging, because the UX during a chatbot interaction is dependent on many diverse factors. Users interacting with customer support systems often find it important to achieve their goal efficiently. However, other factors such as the politeness of the chatbot and the naturalness of replies might also affect the UX. Surveying real users of the chatbot is the most obvious solution, but has many drawbacks. Firstly, users are often not motivated to answer long feedback surveys without any reward [1]. Additionally, feedback surveys require careful preparation and evaluation to gain interpretable insights. Finally, evaluating new chatbot systems on real users can have irreversible consequences to the users' support experience if evaluation results are unsatisfactory, although this can be mitigated by inviting users to a voluntary testing program. Another approach is to carry out evaluation studies with hired participants that are not real users of the chat system, e.g. using crowd-sourcing tools. This is, nevertheless, very costly and such an experiment is complex to conduct. Automatic evaluation metrics on the other hand can be very cheap and have few of these drawbacks while also being consistent. They can therefore complement or even replace human evaluations if they show high agreement with human ground truths.

Objective evaluation metrics such as task success rate and dialog length are easy to automate and have been commonly used for goal-oriented dialog systems [2, 3, 4], but they might not always be sufficient for the analysis of UX. Subjective metrics are harder to automate but capture a broader picture and give a better evaluation of the system when customer satisfaction is important. They measure subjective impressions of the user that can depend on the user's personality, mood and more. Subjective metrics are often reflected in the dialog text through the way a user responds to system messages, e.g. by expressing anger. They are needed to understand the impact of new chatbot strategies such as using customer profiles to personalize chatbot responses [5].

Pre-trained Language Models are already in use to assess subjective qualities of open-domain chit-chat dialogs [6]. Based on the importance of diverse subjective metrics for dialog evaluation and the promising results of Pr-LMs for open-domain chit-chat dialogs, we want to evaluate the use of Pr-LMs for assessing the UX in goal-oriented dialogs. The metrics should be able to capture multiple aspects of the UX and should be practical to use and implement for different systems.

More specifically, we want to investigate if sentence embeddings generated by Pr-LMs contain enough knowledge for predicting the UX of customer support dialogs. We further modify the dialog evaluation metric FED [7] to assess the UX of closed-domain goal-oriented chatbot conversations in practice.

2 Related Work

Automatic dialog evaluation metrics should be repeatable and explainable in the sense that they show qualities that the system lacks. If they correlate well with human judgments, they represent a cheap, reliable and fast substitute for human dialog evaluation [8].

Open-domain dialogs often rely on human assessments or automatic reference-based metrics. Reference-based metrics work by comparing the system response to a reference response, e.g. that of a human. In dialogs, valid responses are however semantically diverse as there is usually not a single correct way to respond to a user statement. Therefore, dialog evaluation metrics should ideally be reference-free. With USR [6] and FED [7], Mehri and Eskenazi utilized Pr-LMs to create reference-free metrics that also offer very fine-grained evaluation scores. The FED metric exploits implicit knowledge about dialog quality by comparing the likelihoods of different manually constructed positive and negative follow-up utterances. For each of the 18 dialog qualities measured by the metric, positive and negative follow-up utterances were handwritten by the authors. The likelihoods of these follow-up utterances are computed using a Pr-LM. In order to measure the quality *interesting* for example, the FED metric will compute the likelihood of the user hypothetical stating the follow-up utterance "Wow, that is really interesting.", among others.

Openly accessible data of goal-oriented dialogs with annotated UX is limited. For investigating the effect of self-efficacy in goal-oriented dialog systems, Cao et al. [9, 10] conducted a study in which various subjective aspects of UX were measured for chat interaction with a bot that gives technical support for mobile phones. The participants of the study were given one of three scenarios that they had to resolve by interacting with the chatbot. They then rated their individual UX on a Likert Scale. The chatbot of the study uses click-options with coherent texts (e.g., "Do you see the battery icon while charging? [Yes, I can] [No, I can't]"), but also allows for free text input, e.g., required for formulating the user's technical problem. In a second stage, other study participants rated the first stage dialogs from an observer point of view. This gives us an estimation of non-expert human performances for UX prediction that we can compare our two methods against.

3 Methods

In the following we present two approaches to automatically predict UX ratings for task-driven and goal-oriented dialogs. The first approach is based on supervised learning and we compare the performance of three Pr-LMs. Furthermore, we evaluate the reference-free FED metric without additional training. The data for both approaches contain 218 UX-labeled dialogs from the dataset of Cao et al. (cp. Sec. 2 and [9, 10]).

3.1 Supervised Learning of UX-Evaluation Using Pr-LMs

We tested three different Pr-LMs to use for predicting the UX-score of customer support dialogs, i.e. SBERT [11], SimCSE [12] and TOD-BERT [13]. SBERT and SimCSE are Pr-LMs intended for general sentence-level representations, while TOD-BERT was trained on a token-level and specifically created for the use in task-oriented dialog tasks. We used the supervised SimCSE variant that is based on the base, uncased BERT model.

Instead of directly training a Pr-LM by adding a fully connected output layer to the model, the process was split: First, embeddings of the dialog texts were generated using a Pr-LM. A feed-forward network was then trained on the task of dialog UX-evaluation with the dialog embeddings as inputs. The feed-forward network predicts the score for each UX item using ordinal regression [14]. Each of the 15 UX items has five output nodes where the five predicted binary values form one prediction value on the Likert scale based on an ordinal regression threshold value of 0.5. The model architecture is equivalent to a Pr-LM that has all layers of the original model frozen with added layers acting as a head for the task of ordinal regression. The fine-tuning of all layers of the Pr-LM was neglected, because early tests showed that the results were worse than with the chosen approach, splitting the process as described. Some of the dialog lengths used for training and testing exceeded the maximum input length of the models, but not by more than a factor of two. To ensure that all words could be encoded, some dialogs were converted to two sentence embeddings.

The results of using the three Pr-LMs in the model architecture were compared to the results of a baseline model using static GloVe embeddings [15] and to the trivial solution of always guessing the most frequent class for a given UX item. Additionally, the correlations between the produced scores and the real ratings were analyzed in order to better examine the potential benefits of using Pr-LMs for the task of dialog UX prediction.

For the evaluation of the model, one has to consider the ordinality and imbalance of the data. Gaudette and Japkowicz [16] argue that for ordinal regression, recognizing small errors (e.g. a rating of five stars as four stars) is less important than recognizing large errors (e.g. a rating of five stars as one star), as even humans find it hard to differentiate a one-star rating distance. The argument can be applied to ratings on the Likert scale in subjective tests and our study as well. We are, for example, not as interested in whether the user perceived the chatbot as *very helpful* or just as *helpful* as we are in whether the chatbot was helpful or not helpful. The Mean Squared Error (MSE) has the advantage of penalizing large errors stronger than small errors, providing a quadratic loss function. Motivated by this, the evaluations of the experiments focus on the Mean Squared Error and not the Mean Absolute Error. The UX item scores are highly imbalanced as many users gave the chatbot similar ratings. This needs to be considered, because a model could always predict the trivial solution, i.e. the most common class of a label, and achieve a very small MSE value. A model that is less conservative and is further away from the trivial solution should be compensated for the higher MSE that results from the class imbalance if it generally models the UX well. To address this problem, we choose the Macro-Averaged Mean Squared Error (MSE^M), a macro-averaged version of the MSE as

introduced by Baccianella et al. [17]. The macro-averaged version of the MSE is analogous to the macro-averaged versions of common multi-class classification metrics such as precision and the F1-score. These however ignore the order of classes which is why the MSE is preferred for ordinal regression.

3.2 Reference-Free UX Evaluation Using FED

In order to make the FED metric suitable for goal-oriented dialogs and the UX qualities of our data, we create custom follow-up utterances that match the UX qualities that we aim to measure. Since we want to predict the UX of an entire conversation, we use follow-up utterances on a dialog level. If a user perceived a conversation with the system as helpful they are more likely to end the dialog with "Thanks, that was helpful!". Other parts of the FED implementation remain identical, including the use of Dialogue Generative Pre-trained Transformer (DialoGPT) [18] to calculate the likelihoods of follow-up utterances.

4 Results

In the following, we present our results on the performance evaluation when using Pr-LMs for supervised learning (cp. Sec. 3.1) as well as FED as a reference-free and unsupervised approach (cp. Sec. 3.2).

4.1 Comparison of Pr-LMs for Supervised Learning

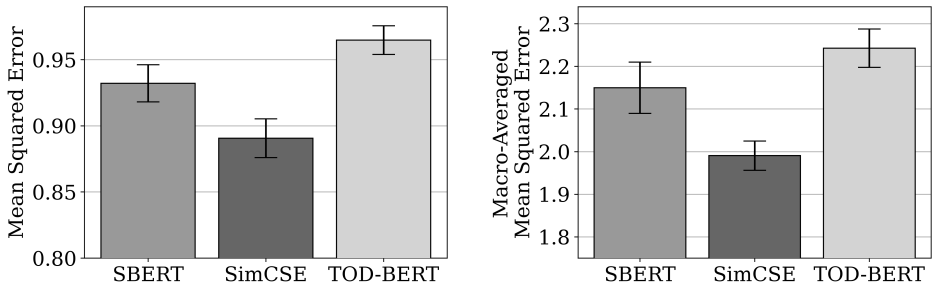
The results presented in Table 1 and Fig. 1 show the performances of the three Pr-LMs (i.e., SBERT, SimCSE, and TOD-BERT) and the GloVe baseline on predicting the UX of the dataset's dialogs. The table also shows the errors of the trivial solution and of the human observer-assessed ratings from the crowdsourcing experiment conducted by Cao et al. [9, 10].

Using SBERT for generating dialog embeddings achieves similar performance to using the baseline model GloVe instead, while TOD-BERT performs worse than the baseline. SimCSE performs best, having the smallest MSE and MSE^M of all Pr-LMs. The macro-averaged MSE of the predictions with SimCSE is similar to the macro-averaged MSE of the observer-assessed scores. The MSE of SimCSE is smaller than that of human observers. To further assess how SimCSE performs in comparison to the human observers, we analyze the Spearman's correlation.

	SBERT	SimCSE	TOD-BERT	GloVe	Trivial Sol.	*Human Observers*
MSE	0.93	**0.89**	0.96	0.93	1.08	*1.41*
MSE^M	2.15	**1.99**	2.24	2.11	3.13	*1.98*

Table 1 – Mean results for the three different Pr-LMs together with GloVe and the trivial solution as baselines. The best result among the Pr-LMs is highlighted for each metric. The column *Human Observers* shows the performance of the observer-assessed ratings.

Fig. 2 shows the correlations of the real UX ratings with the scores predicted by the model, using SimCSE as the Pr-LM. The figure also compares this to the correlations with the observer-assessed ratings collected in Cao et al. [9, 10]. The supervised model's predictions for A5v ("The dialogue was not too long") have no significant correlations with the real UX scores. The score predictions for most other categories however show higher correlation values than those of the human observers from the study.



(a) MSEs **(b)** Macro-Averaged MSEs

Figure 1 – Comparison of the three Pr-LMs after a repeated 10-fold cross validation with $n = 20$ repetitions. Error bars indicate one standard deviation.

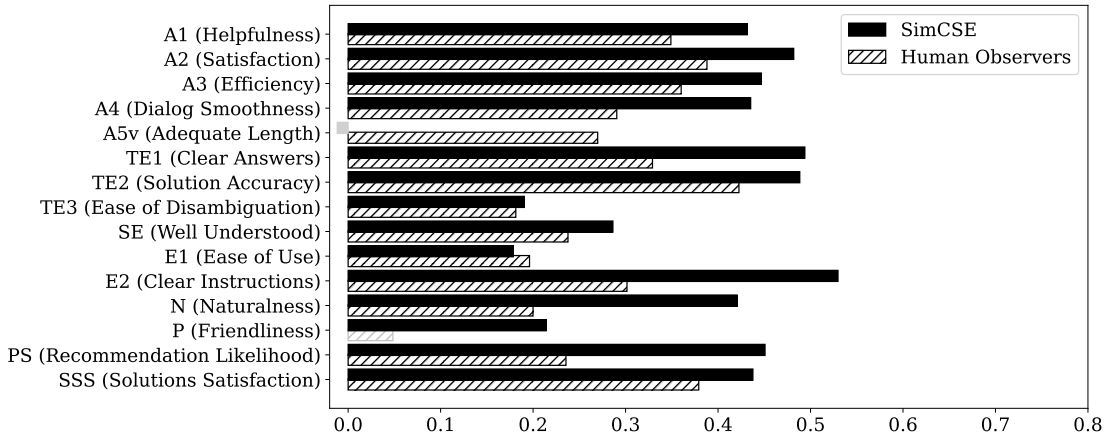


Figure 2 – Spearman's correlations between the real UX, with the scores generated by the model that uses SimCSE, and with the observers-assessed scores annotated by humans. Grey bars indicate insignificant correlations ($p > 0.05$).

4.2 Correlation Evaluation for FED

The custom FED metric has low to moderate *negative* correlation values for most UX items (Fig. 3). The highest absolute correlation values are achieved for the categories E2 ("I knew at each point of the dialogue what the system expected from me"), SSS ("I was satisfied with the answer or solution offered to the given problem") and A3 ("I was able to interact efficiently with the chatbot"). We refer to Cao et al. for details about the used categories [9, 10].

5 Discussion

SimCSE performed better than SBERT, which is in line with results found by Zhang et al. [19] for open-domain dialogs. This is plausible, as SimCSE improves on SBERT in many ways and is seen as the state-of-the-art in both supervised and unsupervised sentence representation learning [19]. The results of the supervised approach are potentially limited by the proposed model architecture which could be modified to improve performance. During pre-training, the employed sentence-level language models get fed very short texts to adjust the model's weights. The representations of sentence embeddings might therefore be non-optimal for longer texts like dialogs. The Pr-LMs could also benefit from domain-adaptive pre-training on the domain of (mobile) customer support chats. The available data of only around 200 dialogs poses another limitation. The quality of crowd-sourced annotations might not be perfect and could be improved by surveying real users of a system with actual goals that they want to achieve through the interaction with the chatbot.

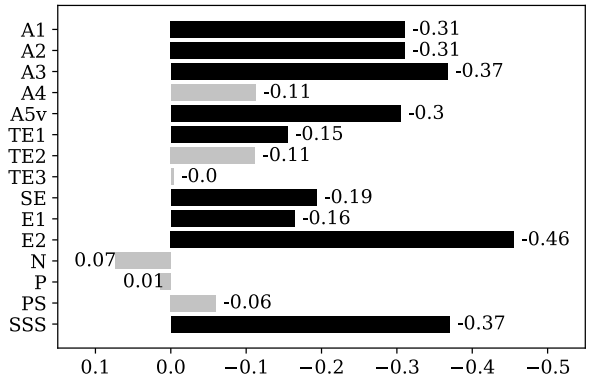


Figure 3 – Correlation between self-assessed UX and scores generated by the custom FED metric. Grey bars indicate statistical insignificance ($p > 0.05$).

The use of supervised learning introduces a general limitation. To get a supervised learning model that can predict the UX, annotated data of user interactions with the chatbot is needed. The annotated data itself could already be a good proxy of the chatbot qualities if this is what we are looking for. This problem could be addressed by training a single model that generalizes well enough to successfully predict the UX in other domains and with different chatbots. Nevertheless, if such a general model is unavailable, it might still be beneficial to collect annotated UX data and train a supervised model for the prediction of unseen dialogs to understand system failures and problems. One might, for example, use such a model to detect the need for handing over the chatbot conversation to a human support agent. A well-performing supervised learning model could also be used to rate the quality of support agents in a company, although this might raise ethical concerns

The FED metric achieved low to moderate, but negative correlations on the dataset. We speculate that the metric fails to capture more than just dialog length for the given data, or that is at least affected by the correlations of dialog length. The FED metric uses log-likelihoods of dialog texts, and because longer dialogs are often less likely for the DialoGPT model than short dialogs, the metric could implicitly capture the dialog length. In the case of chit-chat dialogs longer dialogs usually get a higher score, so in that domain, longer dialogs tend to get a better rating which is in line with the FED scores. In the case of our goal-oriented dialogs dataset, however, the UX tends to be worse for longer dialogs which would explain that the correlations between the FED score and the self-assessed UX ratings are negative instead of positive. Additional correlation analysis on the dataset showed that dialog length would indeed better model the UX than the adjusted FED metric in the given dataset.

Further improvements to the adjusted FED metric for goal-oriented dialog systems could be made. The original FED metric uses DialoGPT as a Pr-LM which has been trained on open-domain dialogs between humans. A more suitable approach for a goal-oriented dialog evaluation metric would use an autoregressive language model trained on the domain of goal-oriented dialog sets instead. More openly available goal-oriented dialog datasets for training these language models are favorable. Furthermore, the FED metric we implemented only considers follow-up utterance likelihoods on the level of the entire dialog. Evaluating follow-up utterances on the turn-level and combining these scores might improve the results. The follow-up utterances also create room for improvements, as they are handwritten and could be further fine-tuned and extended. One could also try to focus only on negative follow-up utterances, because users often do not explicitly express positive feelings such as “You cleared any misunderstandings" which might add noise to the values if DialoGPT’s predicted likelihoods of positive statements do not correlate with the UX.

6 Conclusion

We compared the use of three different Pr-LMs for the supervised learning of UX scores in goal-oriented dialogs and found that SimCSE performs best in combination with our model architecture and even outperforms the UX ratings of human observers (cp. Sec. 4.1). An unsupervised and reference-free approach using the FED metric was tested, which showed worse performance than the supervised approach (cp. Sec. 4.2). Utilizing the implicit knowledge of Pr-LMs in this way still seems to be promising due to the discussed advantages of an unsupervised approach over a supervised one. The FED metric already achieved good results for the UX evaluation of open-domain dialogs, and with the rapid advances of large language models that improve their knowledge and abilities, it seems reasonable to expect that the approach is generally also suitable for goal-oriented dialogs.

7 Acknowledgment

Parts of the presented work and this paper have been funded by the Federal Ministry of Education and Research (Germany) and the Federal State of Berlin under grant no. 16DHBKI088 for the project USOS at Technische Universität Berlin.

References

[1] GUO, Y., J. A. KOPEC, J. CIBERE, L. C. LI, and C. H. GOLDSMITH: *Population survey features and response rates: a randomized experiment*. *American Journal of Public Health*, 106(8), pp. 1422–1426, 2016. doi:10.2105/AJPH.2016.303198.

[2] LI, X., Y. WANG, S. SUN, S. PANDA, J. LIU, and J. GAO: *Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems*. *ArXiv*, abs/1807.11125, 2018.

[3] WEN, T.-H., D. VANDYKE, N. MRKŠIĆ, M. GAŠIĆ, L. M. ROJAS-BARAHONA, P.-H. SU, S. ULTES, and S. YOUNG: *A network-based end-to-end trainable task-oriented dialogue system*. In *Proc. EACL 2017, Volume 1*, pp. 438–449. ACL, 2017.

[4] BUDZIANOWSKI, P., T.-H. WEN, B.-H. TSENG, I. CASANUEVA, S. ULTES, O. RAMADAN, and M. GAŠIĆ: *MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling*. In *Proc. EMNLP 2018*, pp. 5016–5026. ACL, Brussels, Belgium, 2018. doi:10.18653/v1/D18-1547.

[5] LUO, L., W. HUANG, Q. ZENG, Z. NIE, and X. SUN: *Learning personalized end-to-end goal-oriented dialog*. In *Proc. 33rd AAAI Conf. on AI*. AAAI Press, 2019. doi:10.1609/aaai.v33i01.33016794.

[6] MEHRI, S. and M. ESKENAZI: *USR: An unsupervised and reference free evaluation metric for dialog generation*. In *Proc. 58th ACL*, pp. 681–707. ACL, Online, 2020. doi:10.18653/v1/2020.acl-main.64.

[7] MEHRI, S. and M. ESKENAZI: *Unsupervised evaluation of interactive dialog with DialoGPT*. In *Proc. SIGDIAL 2020*, pp. 225–235. Association for Computational Linguistics, 1st virtual meeting, 2020.

[8] DERIU, J., A. RODRIGO, A. OTEGI, G. ECHEGOYEN, S. ROSSET, E. AGIRRE, and M. CIELIEBAK: *Survey on evaluation methods for dialogue systems*. *Artificial Intelligence Review*, 54(1), pp. 755–810, 2021. doi:10.1007/s10462-020-09866-x. 1905.04071.

[9] CAO, Y., V. CARMONA, X. LIU, C. HU, N. ISKENDER, A. BEYER, S. MÖLLER, and T. POLZEHL: *On the impact of self-efficacy on assessment of user experience in customer service chatbot conversations*. In *Proc. IWSDS 2021*. 2021.

[10] CAO, Y.: *Crowdsourced and Automated User Experience Assessment in Customer Service Domain*. Master's thesis, Technische Universität Berlin, 2021.

[11] REIMERS, N. and I. GUREVYCH: *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. In *Proc. EMNLP-IJCNLP 2019*, pp. 3982–3992. ACL, Hong Kong, China, 2019. doi:10.18653/v1/D19-1410.

[12] GAO, T., X. YAO, and D. CHEN: *SimCSE: Simple contrastive learning of sentence embeddings*. In *Proc. EMNLP 2021*, pp. 6894–6910. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021. doi:10.18653/v1/2021.emnlp-main.552.

[13] WU, C.-S., S. C. HOI, R. SOCHER, and C. XIONG: *TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue*. In *Proc. EMNLP 2020*, pp. 917–929. ACL, Online, 2020. doi:10.18653/v1/2020.emnlp-main.66.

[14] CHENG, J., Z. WANG, and G. POLLASTRI: *A neural network approach to ordinal regression*. In *Proc. Int. Joint Conf. Neural Networks (IEEE World Cong. on Comp. Intel.)*, pp. 1279–1284. 2008. doi:10.1109/IJCNN.2008.4633963. ISSN: 2161-4407.

[15] PENNINGTON, J., R. SOCHER, and C. MANNING: *GloVe: Global vectors for word representation*. In *Proc. EMNLP 2014)*, pp. 1532–1543. ACL, Doha, Qatar, 2014. doi:10.3115/v1/D14-1162.

[16] GAUDETTE, L. and N. JAPKOWICZ: *Evaluation methods for ordinal classification*. In Y. GAO and N. JAPKOWICZ (eds.), *Proc. Canadian AI 2009*, pp. 207–210. Springer Berlin Heidelberg, 2009. doi:10.1007/978-3-642-01818-3_25.

[17] BACCIANELLA, S., A. ESULI, and F. SEBASTIANI: *Evaluation measures for ordinal regression*. In *Proc. 9th Int. Conf. on Intelligent Systems Design and Applications*, pp. 283–287. IEEE Computer Society, 2009. doi:10.1109/ISDA.2009.230.

[18] ZHANG, Y., S. SUN, M. GALLEY, Y.-C. CHEN, C. BROCKETT, X. GAO, J. GAO, J. LIU, and B. DOLAN: *DIALOGPT : Large-scale generative pre-training for conversational response generation*. In *ACL System Demonstrations*, pp. 270–278. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-demos.30.

[19] ZHANG, C., L. F. D'HARO, Y. CHEN, T. FRIEDRICHS, and H. LI: *Investigating the impact of pre-trained language models on dialog evaluation*. In S. STOYANCHEV, S. ULTES, and H. LI (eds.), *Proc. IWSDS 2022*, pp. 291–306. Springer Nature Singapore, Singapore, 2021. doi:10.1007/978-981-19-5538-9_21.