Towards a Metric for Automated Conversational Dialogue System Evaluation and Improvement

Jan Deriu
Zurich University of Applied Sciences
deri@zhaw.ch

Mark Cieliebak
Zurich University of Applied Sciences
ciel@zhaw.ch

Abstract

We present ”AutoJudge”, an automated evaluation method for conversational dialogue systems. The method works by first generating dialogues based on self-talk, i.e. dialogue systems talking to itself. Then, it uses human ratings on these dialogues to train an automated judgement model. Our experiments show that AutoJudge correlates well with the human ratings and can be used to automatically evaluate dialogue systems, even in deployed systems. In a second part, we attempt to apply AutoJudge to improve existing systems. This works well for re-ranking a set of candidate utterances. However, our experiments show that AutoJudge cannot be applied as reward for reinforcement learning, although the metric can distinguish good from bad dialogues. We discuss potential reasons, but state here already that this is still an open question for further research.

1 Introduction

Conversational dialogue systems (also referred to as chatbots, social bots, or non-task-oriented dialogue systems) allow for a natural conversation between computer and humans. Research on these dialogue systems has recently reemerged due to the availability of large dialogue corpora, (Serban et al., 2018) as well as the popularization of deep learning (Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016b).

One major challenge in developing high-quality dialogue systems is the evaluation process. Ideally, an evaluation method should be automated, have a high correlation to human judgements and be able to discriminate between different dialogue strategies. Most common techniques to evaluate conversational dialogue systems rely on crowdsourcing, where human judges are asked to rate the *appropriateness* (or *quality*) of a generated response given a context. Although this procedure allows to discriminate between different strategies, it has several drawbacks: it is time and cost intensive, it has to be redone for every change in dialogue strategy, and the results cannot be used to improve the system.

On the other hand, the automated evaluation is usually performed by applying word-overlap metrics borrowed from the machine translation or text summarization community, which have been shown to correlate poorly to human judgements on the utterance level (Liu et al., 2016).

Trained Metrics. Recently, the notion of *trained metrics* was introduced for conversational dialogue systems (Lowe et al., 2017). The main idea is that humans rate the generated response of a dialogue system in relation to a given context (i.e. the dialogue history). Based on these ratings, a regression model is trained which models the human judges. For this, the context, the candidate response, and the gold-standard response are used as input and the judgement is predicted. This approach correlates well with human judgements on the turn level as well as on the system level.

However, these metrics rely on a gold-standard and work on static contexts, which is problematic for two reasons. First, as the context is written by humans it does not reflect the behaviour of the dialogue system. Second, it cannot be used in deployed systems where no gold-standard is available. Dynamic context evaluation (Gandhe and Traum, 2016), on the other hand, usually requires human-computer interaction, which is costly, and puts an additional cognitive strain on the users if they are to rate live during the conversation (Schmitt and Ultes, 2015).

Contribution. In this work we propose to automatically generate the dialogues relying on *self-talk*, which is derived from *AlphaGo* self-play (Silver et al., 2016). Dialogues are generated by two instances of the same system conversing with each other. Then the automatically generated dialogues are rated by human judges. That is, the judges read the dialogues and rate it on the turn level. Based on these ratings, we train a regression model which learns to predict the ratings of the human judges. Our results show that this method, which we refer to as *AutoJudge*, achieves high correlation to human judgements. Thus, it can be applied to fully automatically assess the quality of a dialogue system without being dependent on gold standard responses.

Applications. Since our approach is fully automatic and requires no humans in the loop, we want to go one step further and apply it to *improve* the dialogue system at hand. More precisely we attempt to apply the metric in two different ways: (i) response ranking similar to (Shalyminov et al., 2018; Hancock et al., 2019), and (ii) reward for reinforcement learning. It turns out that only the re-ranking shows promising results, whereas the metric is not useful as a reward function. This is very surprising, since the trained metric correlates well to human judgements, and it can discriminate between good and bad utterances. Why this happens, and how it can be resolved, is an open research question, which we discuss towards the end of this paper.

2 Experimental Setup

Our experimental pipeline follows three phases. First, the data generation phase, where we let the dialogue systems generate dialogues automatically. Second, the data annotation phase, where we rely on crowdsouring to rate the dialogues on the turn level. Third, the improvement phase, where we train an automated judgement model on the annotated data and apply this model to improve the dialogue system.

2.1 Dialogue Systems

For our experiments we relied on the following state-of-the-art dialogue systems (the training details are in Appendix A):

Seq2Seq. The Sequence-to-Sequence model as proposed by (Vinyals and Le, 2015) consists of an encoder and a decoder. Both modules are based on Long Short-Term Memory cells (LSTM) (Hochreiter and Schmidhuber, 1997), where the encoder consumes the last utterance and produces a hidden representation, which is passed as initial state to the decoder to condition the generation process.

HRED. The Hierarchical Recurrent Encoder-Decoder (HRED) model proposed by (Serban et al., 2016a) enhances the Seq2Seq model by a hierarchical encoding procedure. Here, the context-turns are encoded by first encoding each turn separately and then by applying a recurrent encoder over the hidden states of the turns. The decoding procedure is conditioned on the hidden state produced by the context encoder.

VHRED. The Hierarchical Latent Variable Encoder-Dcoder model (VHRED) (Serban et al., 2017a) enhances the aforementioned HRED model by introducing a stochastic latent variable at the utterance level. This stochastic variable aims to inject variability at the utterance level, which in turn increases the variety of responses a model generates.

MrRNN. The Multi-resolution Recurrent Neural Ntwork (MrRNN) (Serban et al., 2017b) enhances the HRED model by introducing an abstraction layer. More precisely, the dialogue is modelled by processing the inputs and outputs at various level of abstractions (e.g. at the level of meaning bearing words and the usual word-level).

DE. The Dual Encoder (DE) (Lowe et al., 2015) is a selection based model, which differs from the generation based approaches of the aforementioned models. The DE encodes both the context and a candidate response (using the same encoder as the VHRED model) and then classifies if the candidate is a valid response to the given context.

2.2 Turn-Level Annotation

We apply *self-talk* to automatically generate dialogues. For this, we sample 100 different contexts randomly from a set of unseen contexts and let the dialogue system generate a dialogue starting from this context, which consist of 10 turns each. For the annotation process, we use Amazon Mechanical Turk (AMT) [1] and follow the procedure outlined by (Lowe et al., 2017), i.e. the judges rated the *overall quality* of each turn on a scale from 1 (low quality) to 5 (high quality). Each turn is annotated by three different judges. We required the AMT workers to be from an english speaking country (USA, UK, Ireland or Australia)

[1] https://www.mturk.com/

in order to ensure that they are native speakers, since the generated messages are highly colloquial and make heavy usage of slang. For each annotation, we paid 15 cents, where we assumed that each annotation takes between 60 to 90 seconds. For the selection of the final turn-label, we apply the MACE procedure (Hovy et al., 2013), which learns confidence scores for the annotators. Our final dataset consists of a total of 500 annotated dialogues, which amounts to 5000 annotated pairs of contexts and responses.

2.3 AutoJudge

Similarly to the *ADEM* procedure proposed by (Lowe et al., 2017), we train a regression model on the annotated data. For this, we use the pre-trained context and response encoder from the VHRED model. Unlike *ADEM*, our dialogues are generated automatically, thus, we do not have access to a gold-standard response. For this reason, we use the following scoring function: $score(c, r) = (c^T M r - \alpha)/\beta$ where $M \in \mathbb{R}^{d \times d}$ is a learned similarity matrix, α, β are scalar constants, and c, r are the context and response embeddings respectively. The model is optimized to minimize the mean squared error between the predicted ratings and the human judgements.

2.4 Improving Dialogue Systems

Since *AutoJudge* is fully automated, we apply it to improve the existing dialogue systems. For this, we implemented the following two applications: as reward for reinforcement learning (RL), and as re-ranking candidate utterances.

Re-Ranking. Given a list of responses from the five aforementioned dialogue systems for a given context, *AutoJudge* re-ranks them by their predicted score. In our experiments, we use the dialogue systems, which we trained for the self-talk experiment, i.e. we re-rank the outputs of the five aforementioned dialogue systems. Thus, the re-ranker serves as a meta-selection module.

Reinforcement Learning Reward. We apply the predicted ratings as reward in the RL framework. For this, we apply the Policy Gradient formulation, as done in (Li et al., 2016), which is defined as follows: $\bigtriangledown J_{RL}(\theta) = \sum_i \bigtriangledown \log p(r_i|c_i) \times \sum_i R(r_i, c_i)$, where r_i and c_i are the response and context in the i^{th} turn, $R(r_i, c_i)$ is the predicted reward by *AutoJudge*, and $\sum_i \log p(r_i|c_i)$ is the reconstruction error.

3 Results and Discussion

In our experiments we use the Twitter Dialogue Corpus (Ritter et al., 2011)[2]. The Twitter Dialogue Corpus provides social interactions, which we believe to be a good basis for being annotated via crowdsouring.

Data Aggregation. The turn-level ratings provide us with 5000 annotated pairs of context and responses. The distribution over the labels is balanced (i.e. each class is represented between 19% and 21% of the cases). However, the agreement scores among the human judges is rather low: the median pairwise Spearman correlation between two judges is only at 0.403. Furthermore, the MACE procedure reports on the confidence score (between 0 and 1) of single judges, which is used as basis for selecting the final label. The average confidence is at only 0.15. We assume that these problems stem from the high degree of subjectivity of the problem.

	Pearson Corr	Spearman's Rho	MAE
CONVO SPLIT	0.573	0.577	0.928
SYSTEM SPLIT	0.544	0.53	0.984

Table 1: Average correlations between the judgements predicted by AutoJudge and the human judgement. CONVO SPLIT denotes the cross-validation split according to the contexts and SYSTEM SPLIT denotes the cross-validation split according to the dialogue system.

AutoJudge. We train *AutoJuge* using k-fold cross validation. There are two ways of splitting the data into folds, in order to ensure that all turns of the same dialogue are in the same fold. First, we group the 100 contexts into 10 folds, thus, each fold consists of 50 dialogues (i.e. 10 contexts times the number of dialogue systems), this is denoted as CONVO SPLIT. The second option is to split the data according to the system which created the conversation, which evaluates the performance of *AutoJuge* in rating dialogues of unseen dialogue systems. We denote this as SYSTEM SPLIT. In Table 1, we report the average Pearson correlation, Spearman's rho and mean absolute error (MAE) over all folds for the *conversation split* and the *system split*. With moderate correlations of 0.573 on the dialogue level, we get results which are comparable to (Lowe et al., 2017),

[2] We use the IDs provided by (Serban et al., 2017a), which can be found here: www.iulianserban.com/Files/TweetIDs.zip

where ADEM achieves a Pearson correlation of 0.436. Note that we cannot directly compare our results to BLEU score and ADEM, since these base their predictions on gold standards, which we do not have in our setting. An interesting result is the *System Split*, i.e. that our approach is able to maintain a high correlation (0.544) with the ratings of a dialogue system when removing the data of that system from the training, which is not the case in (Lowe et al., 2017) where the correlation for a different system dropped significantly.

Answer Selection. In order to evaluate the improvements achieved by the re-ranking method, we sample a disjoint set of 100 new contexts and apply *self-talk* to generate conversations. Then, we use AMT to let humans judge the automatically generated conversations on the dialogue level (i.e. a rating for the entire dialogue as opposed to turn-based ratings). We compare the performance of the five base dialogue systems to the performance of the re-ranking strategy. Table 2 shows the average scores for each dialogue system. Our results show that the *re-ranking* approach works very well. It raises the score to 3.47, which is 0.16 points higher than the best base-system (i.e. SEQ2SEQ).

Systems	Dialogue Level Rating
SEQ2SEQ	*3.31*
HRED	2.78
VHRED	3.20
MRRNN	2.37
DUAL ENCODER	2.02
RE-RANKING	**3.47**

Table 2: Human judgements on the dialogue level for each dialogue system. For this, a each dialogue system (the five base-systems and the re-ranking system) generate 100 dialogues using self-talk, which human judges rated on the dialogue level. Here we see the average ratings for each system.

Reinforcement Learning. When we apply *AutoJudge* as reward resulted in suboptimal dialogues. Although the return increases over time (from 21.74 to 37.41 over 80 episodes), the dialogues which the policy generates are often incoherent or completely useless. This seems counterintuitive when taking into account the aforementioned high correlation scores. We believe that the main reason for the suboptimal behaviour is that *AutoJudge* does not have enough coverage during training. Thus, very bad responses (e.g. empty responses, repeating responses, convergence to a single universal response) tend to receive high scores, since the training data for *AutoJudge* does not include these kinds of responses. However, it is not clear how to stabilize *AutoJudge* to handle these cases. For instance, by artificially enhancing the training data for *AutoJudge* with negative examples, the Pearson correlation score drops to 0.50 without any impact on the reinforcement learning.

4 Conclusion

Our results show that *AutoJudge* correlates well to human judgements and it is useful to measure the progress of a dialogue system, as it is able to discriminate among different strategies. Furthermore, it generalizes well to unseen strategies for the same domain. Since *AutoJudge* is independent of a gold-standard it can be applied to deployed systems where gold-standards are not available. Finally, it shows promising results when applied as answer selection module. As a next step, we intend to apply *AutoJudge* onto human-computer dialogues to measure the viability of *AutoJudge* in a real-world setting.

In this work we tried to use *AutoJudge* as a reward for reinforcement learning, which resulted in suboptimal dialogues. The main reason seems to be that *AutoJudge* cannot properly handle the bad utterance that are generated during the initial phase of reinforcement learning. This is surprising, since *AutoJudge* is able to distinguish good and bad utterances of fully-trained systems. This seems to indicate that there are different types of ”bad” utterances, and we need to adapt the training mechanism of *AutoJudge* if we want to apply it not only to evaluation, but also to improving dialogue systems. Our results indicate that trained metrics suffer from instabilities, which might be caused by the size of the dataset.

One major issue is that it is not clear which aspects *AutoJudge* captures. Although the correlation between the human judgements and the outputs of *AutoJudge* are high, we cannot make any statement about what aspects of the context or the response are relevant for the predicted rating. This is a fundamental problem with the evaluation of conversational dialogue systems, as there is no clear definition for ”adequate” responses. Thus, an important future work problem is the investigation into the definition of ”adequacy” for conversational dialogue systems.

We conjecture that this might apply also to other automated metrics, thus, this is an important research question that needs to be addressed if we want to understand how to better train and optimize dialogue systems.

5 Acknowledgements

This paper has been partially funded by the LIHLITH project, supported by ERA-NET CHIST-ERA and Swiss National Science Foundation (20CH21_174237).

References

Sudeep Gandhe and David Traum. 2016. *A Semi-automated Evaluation Metric for Dialogue Model Coherence*, pages 217–225. Springer International Publishing, Cham.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–1780.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 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. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven Response Generation in Social Media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 583–593, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Schmitt and Stefan Ultes. 2015. Interaction Quality: Assessing the quality of ongoing spoken dialog interaction by expertsAnd how it relates to user satisfaction. *Speech Communication*, 74:12 – 36.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building End-to-end Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press.

Iulian V. Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017a. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, page 1583.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C. Courville. 2017b. Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3288–3294.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016b. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A Survey of Available Corpora for Building Data-Driven Dialogue Systems: The Journal Version. *Dialogue & Discourse*, 1(9).

Igor Shalyminov, Ondřej Dušek, and Oliver Lemon. 2018. Neural response ranking for social conversation: A data-efficient approach. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 1–8, Brussels, Belgium. Association for Computational Linguistics.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.