Published @ CLIC-It 2018, Turin

Automatically Predicting User Ratings for Conversational Systems

A. Cervone[1], **E. Gambi**[1], **G. Tortoreto**[2], **E.A. Stepanov**[2], **G. Riccardi**[1]
[1]Signals and Interactive Systems Lab, University of Trento, Trento, Italy
[2] VUI, Inc., Trento, Italy
{alessandra.cervone, enrico.gambi, giuseppe.riccardi}@unitn.it, {eas,gtr}@vui.com

Abstract

English. Automatic evaluation models for open-domain conversational agents either correlate poorly with human judgment or require expensive annotations on top of conversation scores. In this work we investigate the feasibility of learning evaluation models without relying on any further annotations besides conversation-level human ratings. We use a dataset of rated (1-5) open domain spoken conversations between the conversational agent Roving Mind (competing in the Amazon Alexa Prize Challenge 2017) and Amazon Alexa users. First, we assess the complexity of the task by asking two experts to re-annotate a sample of the dataset and observe that the subjectivity of user ratings yields a low upper-bound. Second, through an analysis of the entire dataset we show that automatically extracted features such as user sentiment, Dialogue Acts and conversation length have significant, but low correlation with user ratings. Finally, we report the results of our experiments exploring different combinations of these features to train automatic dialogue evaluation models. Our work suggests that predicting subjective user ratings in open domain conversations is a challenging task.

Italiano. *I modelli stato dell'arte per la valutazione automatica di agenti conversazionali open-domain hanno una scarsa correlazione con il giudizio umano oppure richiedono costose annotazioni oltre al punteggio dato alla conversazione. In questo lavoro investighiamo la possibilità di apprendere modelli di valutazione attraverso il solo utilizzo di punteggi umani dati all'intera conversazione. Il corpus utilizzato è composto da conversazioni parlate open-domain tra l'agente conversazionale Roving Mind (parte della competizione Amazon Alexa Prize 2017) e utenti di Amazon Alexa valutate con punteggi da 1 a 5. In primo luogo, valutiamo la complessità del task assegnando a due esperti il compito di riannotare una parte del corpus e osserviamo come esso risulti complesso perfino per annotatori umani data la sua soggettività. In secondo luogo, tramite un'analisi condotta sull'intero corpus mostriamo come features estratte automaticamente (sentimento dell'utente, Dialogue Acts e lunghezza della conversazione) hanno bassa, ma significativa correlazione con il giudizio degli utenti. Infine, riportiamo i risultati di esperimenti volti a esplorare diverse combinazioni di queste features per addestrare modelli di valutazione automatica del dialogo. Questo lavoro mostra la difficoltà del predire i giudizi soggettivi degli utenti in conversazioni senza un task specifico.*

1 Introduction

We are currently witnessing a proliferation of conversational agents in both industry and academia. Nevertheless, core questions regarding this technology remain to be addressed or analysed in greater depth. This work focuses on one such question: *can we automatically predict user ratings of a dialogue with a conversational agent?*

Metrics for task-based systems are generally related to the successful completion of the task. Among these, contextual appropriateness (Danieli and Gerbino, 1995) evaluates, for example, the degree of contextual coherence of machine turns with respect to user queries which are classified with ternary values for slots (appropriate, inappro-

priate, and ambiguous). The approach is somewhat similar to the attribute-value matrix of the popular PARADISE dialog evaluation framework (Walker et al., 1997), where there are matrices representing the information exchange requirements between the machine and users towards solving the dialog task, as a measure of task success rate.

Unlike task-based systems, non-task-based conversational agents (also known as chitchat models) do not have a specific task to accomplish (e.g. booking a restaurant). The goal of these can arguably be defined as the conversation itself, i.e. the entertainment of the human it is conversing with. Thus, human judgment is still the most reliable evaluation tool we have for such conversational agents. Collecting user ratings for a system, however, is expensive and time-consuming.

In order to deal with these issues, researchers have been investigating automatic metrics for non-task based dialogue evaluation. The most popular of these metrics (e.g. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005)) rely on surface text similarity (word overlaps) between machine and reference responses to the same utterances. Notwithstanding their popularity, such metrics are hardly compatible with the nature of human dialogue, since there could be multiple appropriate responses to the same utterance with no word overlap. Moreover, these metrics correlate weakly with human judgments (Liu et al., 2016).

Recently, a few studies proposed metrics having a better correlation with human judgment. ADEM (Lowe et al., 2017) is a model trained on appropriateness scores manually annotated at the response-level. Venkatesh et al. (2017) and Guo et al. (2017) combine multiple metrics, each capturing a different aspect of the interaction, and predict conversation-level ratings. In particular, Venkatesh et al. (2017) shows the importance of metrics such as coherence, conversational depth and topic diversity, while Guo et al. (2017) proposes topic-based metrics. However, these studies require extensive manual annotation on top of conversation-level ratings.

In this work, we investigate non-task based dialogue evaluation models trained without relying on any further annotations besides conversation-level user ratings. Our goal is twofold: investigating conversation features which characterize good interactions with a conversational agent and exploring the feasibility of training a model able to predict user ratings in such context.

In order to do so, we utilize a dataset of non-task based spoken conversations between Amazon Alexa users and Roving Mind (Cervone et al., 2017), our open-domain system for the Amazon Alexa Prize Challenge 2017 (Ram et al., 2017). As an upper bound for the rating prediction task, we re-annotate a sample of the corpus using experts and analyse the correlation between expert and user ratings. Afterwards, we analyse the entire corpus using well-known automatically extractable features (user sentiment, Dialogue Acts (both user and machine), conversation length and average user turn length), which show a low, but still significant correlation with user ratings. We show how different combinations of these features together with a LSA representation of the user turns can be used to train a regression model whose predictions also yield a low, but significant correlation with user ratings. Our results indicate the difficulty of predicting how users might rate interactions with a conversational agent.

2 Data Collection

The dataset analysed in this paper was collected over a period of 27 days during the Alexa Prize 2017 semifinals and consists of conversations between our system Roving Mind and Amazon Alexa users of the United States. The users could end the conversation whenever they wanted, using a command. At the end of the interaction users were asked to rate a conversation on a 1 (not satisfied at all) to 5 (very satisfied) Likert scale. Out of all the rated conversations, we selected the ones longer than 3 turns to yield 4,967 conversations. Figure 1 shows the distribution (in percentages) of the ratings in our dataset. The large majority of conversations are between a system and a "first-time" users, as only 5.25% of users had more than one conversation.

3 Methodology

In this section we describe conversation representation features, experimentation, and evaluation methodologies used in the paper.

3.1 Conversation Representation Features

Since in the competition the objective of the system was to entertain users, we expect the ratings to reflect how much they have enjoyed the interaction. User "enjoyment" can be approximated

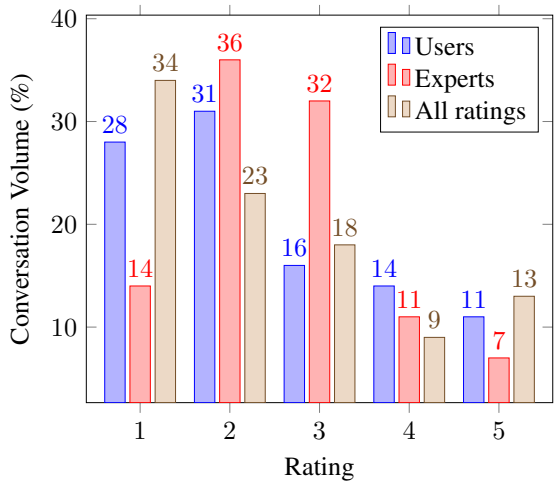

Figure 1: Distribution of user and expert ratings on the annotated random sample of 100 conversations (test set) compared to the distribution of ratings in the entire dataset ("All ratings"). For clarity of presentation, from the latter we excluded the small portion of non integer ratings (2.3% of the dataset).

using different metrics that do not require manual annotation, such as conversation length (in turns), mean turn length (in words), assuming that the more users enjoy the conversation the longer they talk; sentiment polarity – hypothesizing that enjoyable conversations should carry a more positive sentiment. While length metrics are straightforward to compute, the sentiment score is computed using a lexicon-based approach (Kennedy and Inkpen, 2006).

Another representation that could shed a light on enjoyable conversations is Dialogue Acts (DA) of user and machine utterances. DAs are frequently used as a generic representation of intents and the considered labels often include *thanking*, *apologies*, *opinions*, *statements* and alike. Relative frequencies of these tags potentially can be useful to distinguish good and bad conversations. The DA tagger we use is the one described in Mezza et al. (2018) trained on the Switchboard Dialogue Acts corpus (Stolcke et al., 2000), a subset of Switchboard (Godfrey et al., 1992) annotated with DAs (42 categories), using Support Vector Machines. The user and machine DAs are considered as separate vectors and assessed both individually and jointly.

Additional to Dialogue Acts, sentiment and length features, we experiment with word-based text representation. Latent Semantic Analysis (LSA) is used to convert a conversation to a vector. First, we construct a word-document co-occurrence matrix and normalize it. Then, we reduce the dimensionality to 100 by applying Singular Value Decomposition (SVD).

3.2 Correlation Analysis Methodology

The two widely used correlation metrics are Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC). While the former evaluates the linear relationship between variables, the latter evaluates the monotonic one.

The metrics are used to assess correlations of different conversation features, such as sentiment score or conversation length, with the provided human ratings for those conversations; as well as to assess the correlation of the predicted scores of the regression models to those ratings. For the assessment of the correlation of both features and regression models raw rating predictions are used.

3.3 Prediction Methodology

Using the conversation features described above, we train regression models to predict human ratings. We experiment with both Linear Regression and Support Vector Regression (SVR) with radial basis function (RBF) kernel using scikit-learn (Pedregosa et al., 2011). Since the latter consistently outperforms the former, we report only the results for the SVR. The performance of the regression models is evaluated using the standard metrics of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Additionally, we compute Pearson and Spearman's Rank Correlation Coefficients for the predictions with respect to the reference human ratings.

We experiment with the 10-fold cross-validation setting. The performance of the regression models is compared to two baselines: (1) mean baseline, where all instances in the testing fold are assigned as a score the mean of the training set ratings, and (2) chance baseline, where an instance is randomly assigned a rating from 1 to 5 with respect to their distribution in the training set. The models are compared for statistical significance to these baselines using paired two-tail T-test with $p < 0.05$. In Section 6 we report average RMSE and MAE as well as average correlation coefficients.

	RMSE	MAE	PCC	SRCC
Exp 1 vs. Exp 2	0.875	0.660	0.705	0.694
Exp 1 vs. Users	1.225	0.966	0.538	0.526
Exp 2 vs. Users	1.286	1.016	0.401	0.370

Table 1: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson (PCC) and Spearman's rank (SRCC) correlation coefficients among user and expert ratings.

4 Upper bound

Since human ratings are inherently subjective, and different users can rate the same conversation differently, it is difficult to expect the models to yield perfect correlations or very low RMSE and MAE. In order to test this hypothesis two human experts (members of our Alexa Prize team) were asked to rate a random subset of the corpus (100 conversations). The rating distributions for both experts and users on the sample is reported in Figure 1. We observe that expert ratings tend to be closer to the middle of the Likert scale (i.e. from 2 to 4), while users had more conversations with ratings at both extremes of the scale (i.e. 1 and 5).

The RMSE, MAE and Pearson and Spearman's rank correlation coefficients of expert and user ratings are reported in Table 1. We observe that the experts tend to agree with each other more than they agree individually with users, since compared to each other the experts have the highest Pearson and Spearman correlation scores (0.705 and 0.694, respectively) and the lowest RMSE and MAE (0.875 and 0.660, respectively). The fact that expert ratings do not correlate with user ratings as well as they correlate among themselves, confirms the difficulty of the task of predicting subjective user ratings even for humans.

5 Correlation Analysis Results

The results of the correlation analysis are reported in Table 2. From the table, we can observe that conversation length has a positive correlation with human judgment, while the average user turn length has a negative correlation. The positive correlation with conversation length confirms the expectation that users tend to have longer conversations with the system when they enjoy it. The negative correlation with average user turn length, on the other hand, is unexpected. As expected, sentiment score has a significant positive correlation with human judgments.

Feature	PCC	SRCC
Conversation Length	0.133**	0.111**
Av. User Turn Length	-0.068**	-0.079**
User Sentiment	0.071**	0.088**
User Dialogue Acts		
yes-answer	0.081**	0.088**
appreciation	0.070**	0.115**
thanking	0.062**	0.089**
action-directive	-0.069**	-0.052**
statement-non-opinion	0.050**	0.037*
...		
Machine Dialogue Acts		
yes-no-question	0.042**	0.038**
statement-opinion	-0.027*	-0.032*
...		

Table 2: Pearson (PCC) and Spearman's rank (SRCC) correlation coefficients for conversation lengths, sentiment score, and user and machine Dialogue Acts. Correlations significant with $p < 0.05$ are marked with * and $p < 0.01$ with **.

Due to the space considerations, we report only a portion of the DAs that have significant correlations with human ratings. The analysis confirms our expectations that user DAs, such as *thanking* and *appreciation*, have significant positive correlations. We also observe that the *action-directive* DA has a negative correlation. Since this DA label covers the turns where a user issues control commands to the system, we hypothesize this correlation could be due to the fact that in such cases users were using a task-based approach with our system which was instead designed for chitchat and might therefore feel disappointed (e.g. requesting the Roving Mind system to perform actions it was not designed to perform, such as playing music).

Regarding machine DAs, we observe that even though some DAs exhibit significant correlations, overall they are lower than user DAs. In particular, *yes-no-question* has a significant positive correlation with human judgments, indicating that some users appreciate machine initiative in the conversation. The analysis confirms the utility of length and sentiment features, as well as the importance of some DAs (generic intents) for estimating user ratings.

6 Prediction Results

The results of the experiments using 10-fold cross-validation and Support Vector Regression are reported in Table 3. We report performances of each feature representation is isolation and their combi-

	RMSE	MAE	PCC	SRCC
BL: Chance	1.967	1.535	0.007	0.023
BL: Mean	1.382	1.189	N/A	N/A
Lengths	1.400	1.116*	0.153*	0.158**
Sentiment	1.423	1.128*	0.109*	0.122*
DA: user	1.378	1.106*	0.213**	0,207**
DA: machine	1.418	1.129*	0.104*	0.099*
DA: user+machine	1.375	1.106*	0.219**	0.211**
LSA	**1.350***	**1.075***	0.299**	0.288**
All - LSA	1.366*	1.100*	0.240**	0.230**
All	**1.350***	1.078*	**0.303****	**0.290****

Table 3: 10 fold cross-validation average Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson (PCC) and Spearman's rank (SRCC) correlation coefficients for regression models. RMSE and MAE significantly better than the baselines are marked with *. Correlations significant with $p < 0.05$ are marked with * and $p < 0.01$ with **.

nations. We consider two baselines – chance and mean. For the chance baseline an instance is randomly assigned a rating with respect to the training set distribution. For the mean baseline, on the other hand, all the instances are assigned the mean of the training set as a rating. The mean baseline yields better RMSE and MAE scores; consequently, we compare the regression models to it.

Sentiment and length features (conversation and average user turn) both yield RMSE higher than the mean baseline and MAE significantly lower than it. Nonetheless, their predictions have significant positive correlations with reference human ratings. The picture is similar for the models trained on user and machine DAs alone and their combination. The RMSE scores are higher or insignificantly lower and MAE scores are significantly lower than the mean baseline.

For the LSA representation of conversations we consider ngram sizes between 1 and 4. The representation that considers 4-grams and the SVD dimension of 100 yields better performances; thus, we report the performances of this models only, and use it for feature combination experiments. The LSA model yields significantly lower error both in terms of RMSE and MAE. Additionally, the correlation of the predictions is higher than for the other features (and combinations).

The regression model trained on all features but LSA, yields performances significantly better than the mean baseline. However, they are inferior to that of LSA alone. Combination of all the features retains the best RMSE of the LSA model, but achieves a little worse MAE score. While it yields the best Pearson and Spearman's rank correlation coefficients among all the models, the difference from LSA only model is not statistically relevant using Fisher r-to-z transformation.

7 Conclusions

In this work we experimented with a set of automatically extractable black-box features which correlate with the human perception of the quality of interactions with a conversational agent. Furthermore, we showed how these features can be combined to train automatic non-task-based dialogue evaluation models which correlate with human judgments without further expensive annotations.

The results of our experiments and analysis contribute to the body of observations that indicate that there still remains a lot of research to be done in order to understand characteristics of enjoyable conversations with open-domain non-task oriented agents. In particular, our analysis of expert vs. user ratings suggests that the task of estimating subjective user ratings is a difficult one, since the same conversation might be rated quite differently.

For the future work, we plan to extend our corpus to include interactions with multiple conversational agents and task-based systems, as well as to explore other features that might be relevant for assessing human judgment of interaction with a conversational agent (e.g. emotion recognition).

References

Satanjeev Banerjee and Alon Lavie. 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*, volume 29, pages 65–72.

Alessandra Cervone, Giuliano Tortoreto, Stefano Mezza, Enrico Gambi, and Giuseppe Riccardi. 2017. Roving mind: a balancing act between open–domain and engaging dialogue systems. In *Alexa Prize Proceedings*.

Morena Danieli and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*, volume 16, pages 34–39.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2017. Topic-based evaluation for conversational bots. In *NIPS 2017 Conversational AI workshop*.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

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.

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, pages 1116–1126.

Stefano Mezza, Alessandra Cervone, Giuliano Tortoreto, Evgeny A. Stepanov, and Giuseppe Riccardi. 2018. Iso-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics: Technical Papers*, pages 3539–3551.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2017. Conversational ai: The science behind the alexa prize. In *Alexa Prize Proceedings*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Benham Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2017. On evaluating and comparing conversational agents. In *NIPS 2017 Conversational AI workshop*.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.