Scoring Interactional Aspects of Human–Machine Dialog for Language Learning and Assessment using Text Features

Vikram Ramanarayanan[†], **Matt Mulholland**[‡] **and Yao Qian**[†]
Educational Testing Service R&D
[†]90 New Montgomery Street, Suite 1500, San Francisco, CA
[‡]660 Rosedale Rd., Princeton, NJ
<vramanarayanan, mmulholland, yqian>@ets.org

Abstract

While there has been much work in the language learning and assessment literature on human and automated scoring of essays and short constructed responses, there is little to no work examining text features for scoring of dialog data, particularly interactional aspects thereof, to assess conversational proficiency over and above constructed response skills. Our work bridges this gap by investigating both human and automated approaches towards scoring human–machine text dialog in the context of a real-world language learning application. We collected conversational data of human learners interacting with a cloud-based standards-compliant dialog system, triple-scored these data along multiple dimensions of conversational proficiency, and then analyzed the performance trends. We further examined two different approaches to automated scoring of such data and show that these approaches are able to perform at or above par with human agreement for a majority of dimensions of the scoring rubric.

Index Terms: dialog systems, computer assisted language learning, conversational assessment, dialog scoring, intelligent tutoring systems.

1 Introduction

Learning and assessment solutions in today's educational marketplace are placing increasing importance and resources on developing technologies that are dialogic (as opposed to monologic) in nature. Conversational proficiency is a crucial skill for success in today's workplace (Weldy and Icenogle, 1997; Oliveri and Tannenbaum, 2019), which makes R&D on technologies that help develop and assess this skill important to complement our understanding from sociolinguistics (see for example Young, 2011; Doehler and Pochon-Berger, 2015). Dialog system technologies are one solution capable of addressing and automating this need by allowing learners to practice and improve their interactional compentence at scale (Suendermann-Oeft et al., 2017; Yu et al., 2019). However, such conversational technologies need to be able to provide targeted and actionable feedback to users in order for them to be useful to learners and widely adopted. Automated scoring of multiple aspects of conversational proficiency is one way to address this need.

While the automated scoring of text and speech data has been a well-explored topic for several years, particularly for essays and short constructed responses in the case of the former (Shermis and Burstein, 2013; Burrows et al., 2015; Madnani et al., 2017) and monolog speech for the latter (Neumeyer et al., 2000; Witt and Young, 2000; Xi et al., 2012; Bhat and Yoon, 2015), there has been a relative dearth of work on the *interpretable* automated scoring of dialog. Evanini et al. (2015) examined the automatic scoring of pseudo-dialogues, i.e., there were no branching dialog states; the system's response was fixed and did not vary based on the learner's response. Litman et al. (2016) developed a system to predict expert human rater scores based on audio signal and fluency features. Ramanarayanan et al. (2017a) analyzed this scoring problem at the level of each response in the dialog (i.e., each turn) instead of the entire conversation and across multiple dimensions of speaking proficiency. However, no study has performed a comprehensive examination of the automated scoring of *content* of whole dialog responses (with branching) based primarily on text features, based on a comprehensive multidimensional rubric and scoring paradigm designed specifically for dialog data, and interaction aspects in particular.

This study describes our contributions toward (i) developing a comprehensive rubric design

Table 1: *Human scoring rubric for interaction aspects of conversational proficiency. Scores are assigned on a Likert scale from 1-4 ranging from low to high proficiency. A score of 0 is assigned when there were issues with audio quality or system malfunction or off-topic or empty responses.*

Construct	Sub-construct	Description
Interaction	Engagement	Examines the extent to which the user engages with the dialog agent and responds in a thoughtful manner.
	Turn Taking	Examines the extent to which the user takes the floor at appropriate points in the conversation without noticeable interruptions or gaps.
	Repair	Examines the extent to which the user successfully initiates and completes a repair in case of a misunderstanding or error by the dialog agent.
	Appropriateness	Examines the extent to which the user reacts to the dialog agent in a pragmatically appropriate manner.
Overall Holistic Performance		Measures the overall performance.

specifically tailored to conversational dialog along multiple dimensions, particularly those focused on interaction, (ii) triple-scoring a selection of dialog data based on this rubric, and finally (iii) examining the performance of two methods for automated scoring of such data – the first a state-of-the-art feature engineering method that passes word and character *n*-grams, length and syntax features into multiple state-of-the-art classifiers, and the second a model engineering method that leverages end-to-end memory networks to model dependencies between turn and prompt histories using memory components – and analyzing this performance vis-a-vis human raters. Note that for the purposes of this paper, while our data is spoken dialog, we will focus on text features derived from transcriptions, and therefore will focus on how they can be used to score various aspects of interaction in an interpretable manner. A subsequent future analysis will comprehensively examine how these can be combined with speech features.

2 Data

2.1 Collection

We crowdsourced, using Amazon Mechanical Turk, the collection of 2288 conversations of non-native speakers interacting with a dialog application designed to test general English speaking competence in workplace scenarios, and pragmatic skills in particular. The application, dubbed "Request Boss" requires participants to interact with their boss and request a meeting with her to review presentation slides using pragmatically appropriate language. To develop and deploy this application, we leveraged HALEF[1], an open-source modular cloud-based dialog system that is compatible with multiple W3C and open industry standards (Ramanarayanan et al., 2017b). The HALEF dialog system logs speech data collected from participants to a data warehouse, which are then transcribed and scored.

2.2 Human Scoring

In order to understand how well participants performed in our conversational task, we had each of the 2288 dialog responses triple scored by human expert raters on a custom-designed rubric. This rubric was iteratively modified and refined to score constructs specific to dialog data[2]. The final conversational scoring rubric defined 12 subconstructs under the 3 broad constructs of linguistic control, task fulfillment and interaction, apart from an overall holistic score. However, for purposes of this first study, we will focus on the relatively understudied interaction construct, in particular aspects of engagement, turn-taking, repair and (pragmatic) appropriateness. See Table 1 for more details. We asked expert raters to score each dialog for each rubric dimension on a scale from 1 to 4, and to assign dialogs that contained no or corrupted or significantly off-topic audio responses a score of 0. The expert raters were scoring leaders with significant experience in scoring various spoken and written assessments of English language proficiency. We used an automatic randomized design to assign three (out of eight possible) raters to every dialog such that (i) all raters had a commensurate number of responses to rate, and (ii) the same group of raters did not rate the same set of files (achieved by randomization; this prevents unwitting biases due to individual raters affecting the overall score analysis).

[1] `http://halef.org`

[2] Three scoring leaders first collaboratively adapted a rubric originally developed to score spoken interaction based on selected benchmark dialog responses. Based on this modified rubric and accompanying scoring notes specific to the task, 8 scoring leaders performed the final round of scoring.

Table 2: *c-rater ML features used for machine scoring.*

Feature	Description
Word *n*-grams	Word *n*-grams are collected for n = 1 to 2. This feature captures patterns about vocabulary usage (key words) in responses.
Character *n*-grams	Character *n*-grams (including whitespace) are collected for *n* = 2 to 5. This feature captures patterns that abstract away from grammatical and other language use errors.
Response length	Defined as *log(chars)*, where *chars* represents the total number of characters in a response.
Syntactic dependencies	A feature that captures grammatical relationships between individual words in a sentence. This feature captures linguistic information about "who did what to whom" and abstracts away from a simple unordered set of key words.

3 Machine Scoring

This section first lays out our setup for interpretable machine scoring including details of the feature extraction and machine learning methods. We then analyze human performance (by examining inter-rater statistics) and use this to benchmark the performance of machine scoring methods. Following standardized convention in automated scoring, we only consider dialogs with a non-zero score to train scoring models (because a separate filtering mode is typically trained to eliminate "unscorable" responses, which include responses with no, garbled or out-of-topic audio data, see Higgins et al., 2011, for a more detailed motivation and rationale for this approach).

3.1 Feature Engineered Content Scoring

We used a set of features that have been employed in many previously published approaches to building content scoring models (see Madnani et al., 2017, 2018, for instance). We refer to this system as *c-rater ML*; see Table 2 for more details. All of the features are binary (indicating presence or absence) and try to capture how well responses contain (a) the right concepts (approximately captured by words and bigrams), (b) the right syntactic relationships between those concepts (approximately captured by dependency triples), (c) spelling and morphological relations (character *n*-grams) and (d) length of the response (captured by length features).

We used SKLL,[3] an open-source Python package that wraps around the *scikit-learn* package (Pedregosa et al., 2011) to perform machine learning experiments. We experimented with rescaled linear support vector machine (SVM) and multilayer perceptron (MLP) regressors. The former allows us to interpret how the algorithm performs, while the latter is used for comparison purposes to understand how deep neural networks might perform on this task given the data we have. In our case, we found that the SVM classifier beat the MLP across the board, possibly because our feature space is sparse and high-dimensional, consisting of binary presence/absence features. We ran 10 fold cross-validation experiments and report the best overall results for the SVM system. We used cross entropy (log-loss) as an objective function for optimizing learner performance. We further tuned and optimized the free parameters of each learner using a grid-search method. We computed both accuracy and quadratic weighted kappa (which takes into account the ordered nature of the categorical labels) as metrics, reported in Table 3.

3.2 End to End Memory Network (MemN2N) architecture

We also investigated the efficacy of the End to End Memory Network (MemN2N) architecture (Sukhbaatar et al., 2015; Chen et al., 2016) adapted to the dialog scoring task. The end to end MemN2N architecture models dependencies in text sequences using a recurrent attention model coupled with a memory component, and is therefore suited to modeling how response and prompt histories contribute to a dialog score. In our case, the MemN2N architecture learns a mapping between an output score and an input tuple consisting of the current response, the response history and the prompt history. See Figure 1. We modified the original MemN2N architecture in Sukhbaatar et al. (2015) in the following ways: (i) instead of the original *(query, fact history, answer)* tuple that is used to train the network in the original paper, we have an *(current response, response his-*

[3] https://github.com/EducationalTestingService/skll

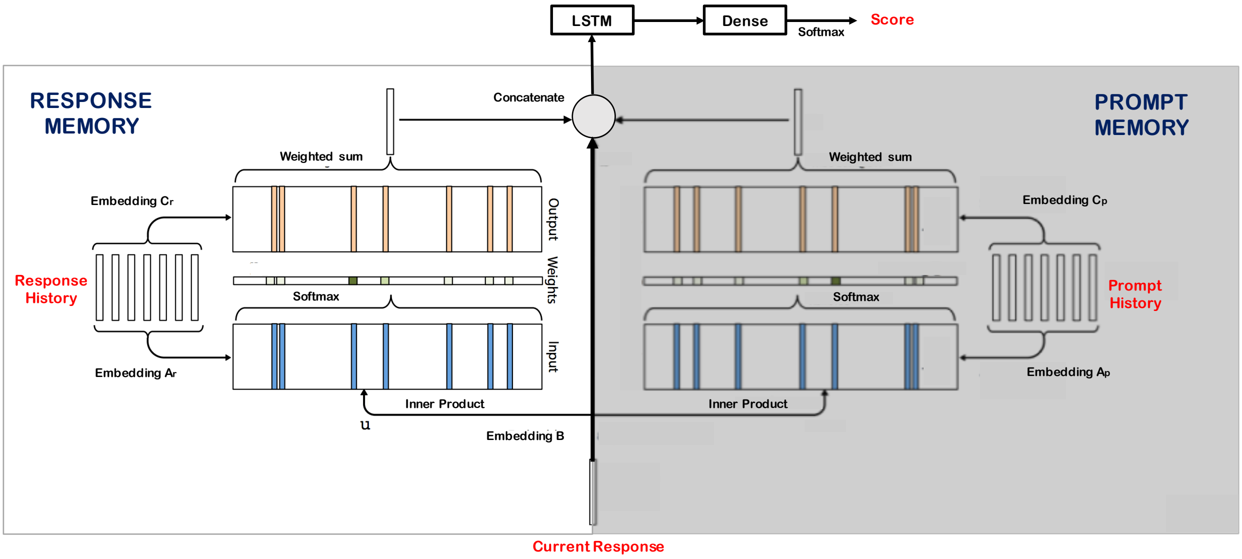

Figure 1: Schematic of a *single hop* module of our modified end-to-end memory network (MemN2N) adapted from Sukhbaatar et al. (2015) for our dialog scoring experiments. Stacking modules on top of each other allows us to model multiple hops.

Table 3: *Human and machine score statistics*

Construct	Sub-construct	c-rater ML		MemN2N		c-rater ML + MemN2N		Human Metrics	
		Accuracy	QWκ	Accuracy	QWκ	Accuracy	QWκ	Conger κ	Krippendorff α
Interaction	Engagement	0.70	0.70	0.65	0.65	0.71	0.72	0.69	0.72
	Turn Taking	0.69	0.67	0.68	0.40	0.71	0.70	0.71	0.74
	Repair	0.66	0.60	0.64	0.58	0.67	0.64	0.73	0.72
	Appropriateness	0.67	0.67	0.62	0.58	0.67	0.67	0.70	0.72
Overall Holistic Performance		0.69	0.72	0.66	0.65	0.70	0.72	0.75	0.75

tory, prompt history, score) tuple in our case. In other words, we not only embed and learn memory representations between the current response and the history of previous responses, but the history of prior system prompts that have been encountered thus far; (ii) we used an LSTM instead of a matrix multiplication at the final step of the network before prediction; and (iii) we experimented with *Google word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014) initializations for word embeddings in addition to experimenting with multiple memory hops. We train the network at the turn level; in other words, for each turn, the training data would consist of an input of *(response for current turn, response history, prompt history)* and an output of the *dialog*-level score (in other words, each turn is assumed to have the same score as that of the full dialog). During testing, we compute the score for each dialog in the test set as the median of scores predicted by the trained network for each turn in that dialog.

We used a similar crossvalidation setup as described in §3.1 with the exact same 10 folds with experiments optimizing a cross-entropy-based objective function as in the earlier case to enable a fair comparison. We tuned hyperparameters of the network using the *hyperas* toolkit[4]. This included the number of neurons in the *Dense* and *LSTM* layers as well as the addition of *Dropout* layers after each memory component. We experimented with 1, 2 and 3 memory hops and found 2 to be optimal. Interestingly, we also found that initializing the memory embedding matrics with pretrained *Google word2vec* or *GloVe* embeddings worked better than randomly-initialized ones for prompt history encoding as compared to response history encoding.

4 Observations and Results

The final two columns of Table 3 display two inter-rater agreement statistics – Conger κ and Krippendorff α – for the human expert scores assigned to the data. Recall that each dialog was scored by 3 out of 8 possible raters. We observe a moderate to high agreement between raters for all dimensions

[4] http://maxpumperla.com/hyperas/

of the scoring rubric, which is not too surprising given that all our raters had significant experience in rating monologic speech data.

Table 3 also shows the performance of our two different systems in scoring various aspects of interaction at the level of the entire dialog. Observe that fusing the MemN2N with the c-rater ML system leads to a small but significant improvement over either of the systems alone. Additionally, it is interesting to note that the quadratic weighted kappa ($QW\kappa$) of the fusion system is in a similar ballpark as the κ and α metrics for human inter-rater agreement, particularly for engagement and turn-taking subscores. While these measures are *not directly comparable*, this trend is encouraging nonetheless, suggesting that a combination of n-gram, length, syntactic dependency and memory-based attention over embedding representations of words over the entire dialog are useful in capturing at least some aspects of these sub-constructs of interaction. On the other hand, the fusion system performance for repair and appropriateness subscores is still below par, suggesting that more feature engineering and modeling research is required to model these aspects of interaction. These dimensions of interaction are also harder to predict, given that repair and pragmatic appropriateness are more high-level and abstract in nature.

5 Discussion

This paper has examined approaches to both human and machine scoring of text dialogs collected as part of a language learning application, particularly looking at interactional aspects. We observed, through careful design of the human scoring paradigm, a moderate-to-high agreement between the raters. We further examined two methods for automated scoring of such data – the first a feature engineering method that passes word and character n-grams, length and syntax features into an SVM based classifier, and the second a model engineering method that leverages end-to-end memory network (MemN2N) to model dependencies between turn and prompt histories using memory components – and found that a fusion of both methods performs close to or at par with human inter-rater agreement statistics.

While our results are encouraging, there is still much work ahead in understanding and scoring interactional competence. One of the key reasons for this has to do with the fact that the features were considered were text-based, and it is unclear how some features that don't directly consider information from audio or visual channels are useful in predicting properties related to interaction (engagement, for instance). Repair and appropriateness, and even turn taking to a lesser extent are related to proficiency in language use, and hence it makes sense that features such as n-grams and syntax use might be somewhat useful in predicting these aspects of interaction. However, some of the results might also be explained by some of our examined features being highly correlated with more interpretable/relevant features. For instance, length might be an indication of a more proficient and verbose speaker, which might in turn correlate with a high level of engagement. Nonetheless, an understanding of how meaningful our text-based results are will be incomplete without examining features derived from audio (and visual streams, if available), including non-verbal and prosodic cues.

It is also worth mentioning tangentially related work on dialog interaction quality at this point (see for instance Schmitt and Ultes, 2015; Stoyanchev et al., 2019; See et al., 2019). While such work primarily focuses on investigating techniques to measure and improve the quality of the overall dialog interaction as opposed to providing targeted assessment and feedback on the quality of spoken language used during interactions, it might nonetheless be useful to take this body of work into account while developing techniques for automated proficiency scoring.

This lays out multiple avenues for future work. First, as mentioned earlier, would be examining both text and speech signals for a more complete examination of the scoring problem. Second, we would like to look at other broad aspects of conversational proficiency, such as delivery (for instance, fluency, intonation, vocabulary and grammar) and topic development (elaboration and task specificity, for example) in addition to building on the interaction aspects described here. Third, we will investigate combining feature-engineering and model-engineering approaches towards developing specific features and model architecture improvements that will help us push the automated scoring performance even higher. These will feed into our ultimate goal of being able to provide language learners with targeted, actionable feedback on different facets of conversational proficiency.

References

Suma Bhat and Su-Youn Yoon. 2015. Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, 67:42–57.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech*, pages 3245–3249.

S Pekarek Doehler and Evelyne Pochon-Berger. 2015. The development of l2 interactional competence: evidence from turn-taking organization, sequence organization, repair organization and preference organization. *Usage-based perspectives on second language learning*, 30:233.

Keelan Evanini, Sandeep Singh, Anastassia Loukina, Xinhao Wang, and Chong Min Lee. 2015. Content-based automated assessment of non-native spoken language proficiency in a simulated conversation. In *Proceedings of the Machine Learning for SLU & Interaction NIPS 2015 Workshop*.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2):282–306.

Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen, and David Vandyke. 2016. Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of english. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 270.

Nitin Madnani, Aoife Cahill, Daniel Blanchard, Slava Andreyev, Diane Napolitano, Binod Gyawali, Michael Heilman, Chong Min Lee, Chee Wee Leong, Matthew Mulholland, et al. 2018. A robust microservice architecture for scaling automated scoring applications. *ETS Research Report Series*, 2018(1):1–8.

Nitin Madnani, Anastassia Loukina, and Aoife Cahill. 2017. A large scale quantitative exploration of modeling strategies for content scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 457–467.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech communication*, 30(2):83–93.

Maria Elena Oliveri and Richard J Tannenbaum. 2019. Are we teaching and assessing the english skills needed to succeed in the global workplace? *The Wiley Handbook of Global Workplace Learning*, pages 343–354.

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*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vikram Ramanarayanan, Patrick L Lange, Keelan Evanini, Hillary R Molloy, and David Suendermann-Oeft. 2017a. Human and automated scoring of fluency, pronunciation and intonation during human-machine spoken dialog interactions. In *INTERSPEECH*, pages 1711–1715.

Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Robert Mundkowsky, Alexei V Ivanov, Zhou Yu, Yao Qian, and Keelan Evanini. 2017b. Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System. In *Multimodal Interaction with W3C Standards*, pages 295–310. Springer.

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.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Svetlana Stoyanchev, Soumi Maiti, and Srinivas Bangalore. 2019. Predicting interaction quality in customer service dialogs. In *Advanced Social Interaction with Agents*, pages 149–159. Springer.

David Suendermann-Oeft, Vikram Ramanarayanan, Zhou Yu, Yao Qian, Keelan Evanini, Patrick Lange, Xinhao Wang, and Klaus Zechner. 2017. A multimodal dialog system for language assessment: Current state and future directions. *ETS Research Report Series*, 2017(1):1–7.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Teresa G Weldy and Marjorie L Icenogle. 1997. A managerial perspective: Oral communi cation competency is most important for business students in the workplace jeanne d. maes. *The Journal of Business Communication*, 34(1):67–80.

Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2):95–108.

Xiaoming Xi, Derrick Higgins, Klaus Zechner, and David Williamson. 2012. A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, 29(3):371–394.

Richard F Young. 2011. Interactional competence in language learning, teaching, and testing. *Handbook of research in second language teaching and learning*, 2(426-443).

Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft. 2019. An open-source dialog system with real-time engagement tracking for job interview training applications. In *Advanced Social Interaction with Agents*, pages 199–207. Springer.