A Semi-automated Evaluation Metric for Dialogue Model Coherence

Sudeep Gandhe and David Traum

Abstract We propose a new metric, *Voted Appropriateness*, which can be used to automatically evaluate dialogue policy decisions, once some wizard data has been collected. We show that this metric outperforms a previously proposed metric *Weak agreement*. We also present a taxonomy for dialogue model evaluation schemas, and orient our new metric within this taxonomy.

1 Introduction

There has been a lot of work in end-to-end evaluation of dialogue systems, but much less so on the dialogue modelling component itself. The key task here is: given a context of prior utterances in the dialogue, choose the next system utterance. There are many possible ways of evaluating this decision, including whether it replicates an original dialogue move, how close it is to that move (e.g., [4]), and human evaluations of quality or coherence. In Sect. 2 we provide a taxonomy that organizes types of evaluation along a series of dimensions regarding evaluation metric, evaluator and evaluation context.

For the purposes of using machine learning for improving dialogue policies, it is critical to have a high-quality automatic evaluation method. MDP [8] and POMPD [17] dialogue models are generally evaluated with respect to a reward function, however these reward functions typically function at the level of whole dialogues and not specific choices (even though reinforcement learning models estimate the contribution of individual moves). There is still much work needed in picking good reward functions, and this task is much harder, when the metric of importance concerns dialogue coherence rather than task success.

S. Gandhe (✉) · D. Traum
Institute for Creative Technologies, University of Southern California,
Los Angeles, CA 90094, USA
e-mail: srgandhe@gmail.com

D. Traum
e-mail: traum@ict.usc.edu

A. Rudnicky et al. (eds.), *Situated Dialog in Speech-Based Human-Computer Interaction*, Signals and Communication Technology,
DOI 10.1007/978-3-319-21834-2_19

We propose a semi-automated evaluation paradigm, similar to BLEU used in machine-translation [10] or ROUGE, used in summarization [9], and improving on the previously proposed metric *weak-agreement* [2]. In this paradigm, a set of human "wizards" make the same decisions that the system will have to make, and this data is used to evaluate a broader set of system decisions. This approach is particularly appropriate in a *selection* paradigm for producing system utterances, where the system (or wizard) selects from a corpus of previous dialogue utterances rather than generating a novel utterance.

The work described in this paper is done within the scope of *Virtual Human Dialogue Systems*. Virtual Humans are autonomous agents who can play the role of humans in simulations [14]. Virtual Human characters have proved useful in many fields; some have been used in simulations for training negotiation skills [13] or tactical questioning skills [12]; some virtual humans are used in settings where a face-to-face conversation can have a stronger impact in presenting some information (e.g., a Virtual Nurse used for counseling hospital patients who have inadequate health literacy at the time of discharge [1], Museum Docents promoting science and technology interests in middle school students [11]); some virtual humans are used as non-playing characters in interactive games (e.g., [7]). Although different virtual humans may have different sets of goals, one common requirement for all of them is the ability to take part in natural language conversations.

2 Evaluation Schema for Conversational Dialogue Models

Evaluating a dialogue model requires making a series of decisions. Figure 1 shows a schematic representation of such decisions for evaluation of dialogue models.

The first decision is which evaluation metric to use. This is dependent on the goals of the dialogue system. In case of a task-oriented dialogue system, some suitable

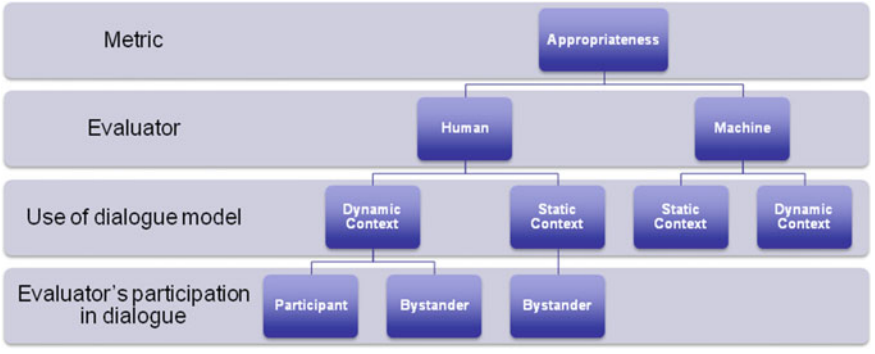

Fig. 1 A schematic representation of various decision factors in evaluating dialogue models for virtual humans

choices for an evaluation metric are user satisfaction, task success rate, task efficiency, etc. [16]. For tutoring dialogue systems, some suitable evaluation metrics can be user satisfaction or learning gain as measured by differences between post-test and pre-test scores [3]. Since the goal for virtual humans is to be as human-like as possible, a suitable evaluation metric for virtual human dialogue systems is how appropriate or human-like the responses are for a given dialogue context. These evaluation metrics can be subjective or objective and can measured at different levels of granularity such as utterance-level, dialogue-level, user-level, etc.

The next decision is who evaluates the dialogue models. The dialogue models we need to evaluate are designed to be part of a virtual human who will engage human users in natural language conversations. Judging appropriateness of a response utterance given a dialogue context in such conversations is not an easy process and may require human-level intelligence. This is why human judges are a natural choice for such a subjective evaluation metric.

Although humans are best suited to evaluate appropriateness of responses, using humans as judges is costly and time-consuming. For these and other reasons, automatic evaluation becomes an attractive alternative.

The next decision criterion is how the dialogue model to be evaluated is used in the process of generating response utterances and the corresponding dialogue contexts. There are two possible settings *Dynamic Context* and *Static Context*. Figure 2 shows a schematic representations for these different settings.

Dynamic Context

In *dynamic context* evaluation, the dialogue model is used for generating the response utterances as well as the dialogue contexts with respect to which the subsequent responses are evaluated. In this case, we build a dialogue system using the dialogue model that needs to be evaluated. A human user interacts with this dialogue system. The system's response is the top-ranked response utterance for the given dialogue context as ranked by the dialogue model.

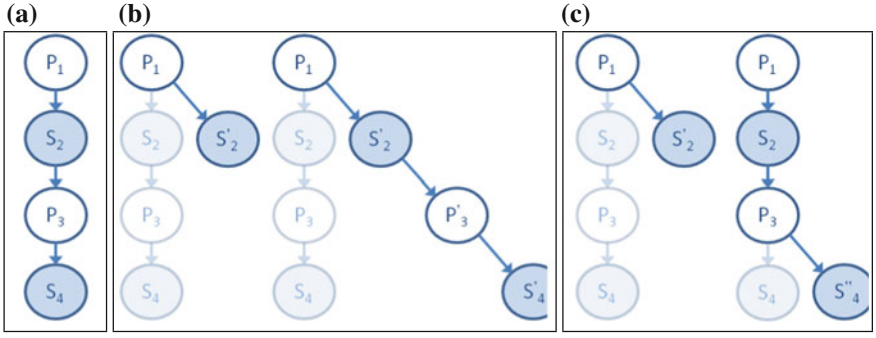

Fig. 2 Schematic representation of *dynamic context* and *static context* evaluation settings. **a** Original human-human dialogue, **b** dynamic context setting, **c** static context setting

Figure 2b shows first two stages of the *dynamic context* evaluation process. At first, the user produces an utterance P_1. Based on the context $\langle P_1 \rangle$, the dialogue model being evaluated produces the response utterance $S_2^{'}$. This response may be different from utterance S_2, which was the response in original human-human dialogue (Fig. 2a). The user continues the dialogue and responds to the system's response with utterance $P_3^{'}$. The next response from the system produced by the dialogue model being evaluated is based on the context $\langle P_1, S_2^{'}, P_3^{'} \rangle$. This context is dependent on the dialogue model being evaluated. Thus during dynamic context evaluation the resulting dialogue (and the intermediate dialogue contexts) are generated through an interactive process between a human user and a dialogue model. If an inappropriate response is chosen by the dialogue model then it becomes part of the context used to select the next response. Thus the dialogue model has the potential to recover from its errors or to build on them. System's responses are evaluated for appropriateness with respect to the same contexts that were used to generate them.

Static Context

In *static context* evaluation the dialogue model is used for generating only the response utterances. The dialogue contexts are not affected by the specific dialogue model being evaluated. These dialogue contexts are extracted from actual in-domain human-human dialogues. For every turn whose role is to be played by the system, we predict the most appropriate response in place of that turn given the dialogue context.

Figure 2c shows first two stages of the *static context* evaluation process. The first system response is generated based on the context $\langle P_1 \rangle$ and is $S_2^{'}$, the same as in the case of *dynamic context*. But for the second response from the system, the context is reset to $\langle P_1, S_2, P_3 \rangle$ the same as the original human-human dialogue and does not depend on the dialogue model being evaluated. The system's response then is $S_4^{''}$, which can be different from both S_4 (human-human) and $S_4^{'}$ (dynamic context). Again, the system's responses are evaluated for appropriateness with respect to the same contexts that were used to generate them.

The next decision criterion in evaluating dialogue models is whether the evaluator takes part in the conversation. If we require that the evaluator participates in the dialogue then each dialogue can be evaluated by only one evaluator—the participant himself. This evaluation scheme assumes that the conversational participant is in the best position to judge the appropriateness of the response. The Turing test [15] calls for such a dynamic context evaluation by the participant where instead of appropriateness, the evaluation metric is whether the conversational participant is human or machine.

Although evaluation by a dialogue participant is the most faithful evaluation possible, it is costly. As only one evaluator can judge a dialogue, we need to create a large enough test corpus by conducting conversations with the system. Moreover, volunteers may find playing two roles (dialogue participant and evaluator) difficult. In such cases, evaluation by a bystander (overhearer) can be a suitable alternative. In this type of evaluation the evaluator does not actively participate in the conversation and more than one evaluator can judge a dialogue for appropriateness of responses.

In case of multiple judges, the average of their judgments is used as a final rating for appropriateness. For static context evaluation, the evaluator is always a bystander if s/he doesn't take part in creating the original human-human dialogue.

3 Automatic Static Context Evaluation

Recently we evaluated 7 dialogue models for a *Virtual Human Dialogue System*. We used the negotiation scenario where a human trainee tries to convince a virtual doctor to move his clinic [13]. We conducted a *Static Context* evaluation of response appropriateness using human judges [5]. We evaluated 5 computer dialogue models and 2 wizard dialogue models as upper human-level baselines. For wizard dialogue models, we collected data from four wizards as to which utterances are appropriate responses for given dialogue contexts using the tool described in [6]. The data collected from wizards is used to build two models: *Wizard Max Voted* model, which returns the response with the maximum number of votes from the four wizards; and *Wizard Random* model, which returns a random utterance from the list of all utterances marked as appropriate by any of the wizards. We also collected ratings for appropriateness of responses from different dialogue models on a scale of 1–5 (1 being very inappropriate response and 5 perfectly appropriate). The ratings were provided by four human judges for the same dialogues as used in wizard data collection.[1] This results in a collection of appropriateness ratings for a total of 397 unique pairs of $\langle u_t, context_t \rangle$, where u_t is a response utterance for a dialogue context $context_t$. We use this data for proposing and evaluating automatic evaluation measures in static context setting.

3.1 Weak Agreement

DeVault et al. [2] used an automatic evaluation measure based on wizard data collection for evaluating various dialogue models in a static context setting. The dialogue models evaluated in that study operate at the dialogue act level and consequently the wizard data collection is also done at the dialogue act level. Their proposed automatic evaluation, *weak agreement*, judges the response dialogue act for a given context as appropriate if any one of the wizards has chosen that dialogue act as an appropriate response. In their study DeVault et al. do not correlate this automatic measure with human judgments of appropriateness.

Let $R(u_t, context_t)$ denote the average appropriateness of the response utterance u_t for the dialogue context $context_t$ as judged by the four human judges. Also let $W(context_t)$ be the union of set of responses judged appropriate for the dialogue context $context_t$ by the four wizards. Then following [2], an automatic evaluation for response appropriateness along the lines of *weak agreement* can be defined as,

[1] Two of the judges also performed the role of the wizards, but the wizard data collection and the evaluation tasks were separated by a period of over 3 months.

$$R_{weak}(u_t, context_t) = \begin{cases} 5 \text{ if } u_t \in W(context_t) & \text{Appropriate response} \\ 1 \text{ if } u_t \notin W(context_t) & \text{Inappropriate response} \end{cases} \quad (1)$$

In order to test the validity of this automatic evaluation metric (R_{weak}), We correlate it with human judgments (R). This correlation can be computed either at the level of an individual response (i.e., for every unique value of $\langle u_t, context_t \rangle$) or at the system level (i.e., by aggregating the ratings over each dialogue model). The Pearson's correlation between R_{weak} and R is 0.485 ($p < 0.001, n = 397$) at individual response level and 0.803 ($p < 0.05, n = 7$) at the system level. Although we report both correlation values, we're primarily interested in comparing dialogue models with each other. So we focus on the system level correlation. *Weak Agreement*, R_{weak} turns out to be a good evaluation understudy for judging appropriateness of responses given a dialogue context especially at the system level.

3.2 Voted Appropriateness

We made an observation regarding R_{weak} which may lead to an improvement. According to weak agreement, we should expect *Wizard Max Voted* and *Wizard Random* models to have the same appropriateness rating of value 5 (by definition in 1). Instead, we observe that *Wizard Max Voted* model receives significantly higher appropriateness ratings than *Wizard Random*. This indicates that not all responses chosen by wizards are judged as highly appropriate by other judges. It also suggests that more votes from wizards for a response utterance are likely to result in higher appropriateness ratings.

Based on these observations, we propose an evaluation understudy *Voted Appropriateness*, R_{voted}. Let $V(u_t, context_t)$ be the number of wizards who chose the utterance u_t as an appropriate response to the dialogue context $context_t$. Following PARADISE [16], which models user satisfaction as a linear regression of observable dialogue features, we model R_{voted} as a linear regression based on V.

$$R_{voted}(u_t, context_t) = \alpha_0 + \alpha_1 \cdot V(u_t, context_t) \quad (2)$$

Figure 3 shows the appropriateness rating (R) as judged by human judges for response utterances as a function of number of wizard votes (V) received by those response utterances. For this analysis we use only distinct pairs of $\langle u_t, context_t \rangle$ ($n = 397$). We fit a linear regression model for this data. The number of votes received V is a significant factor in estimating R ($p < 0.001$). The final linear model estimated from all available data is, $R_{voted} = 3.549 + 0.449V$. The fraction of variance explained by the model is 0.238.

To verify whether a simple linear regression model can be used as an automatic evaluation for static context setting, we perform fivefold cross-validation analysis. During each fold, we hold out the data corresponding to one of the dialogues and train a linear model on the rest of the data. We use this trained model to compute *voted appropriateness* (R_{voted}) for the held-out data and then correlate it with the actual

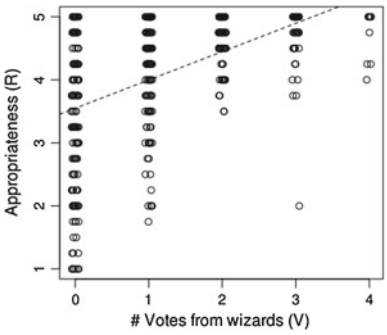

Fig. 3 Appropriateness of responses (R) as judges by 4 human judges plotted against the number of wizard votes (V) received by those responses. The *dashed line* indicates a fitted linear model. A small amount of jitter is added to V for visualization

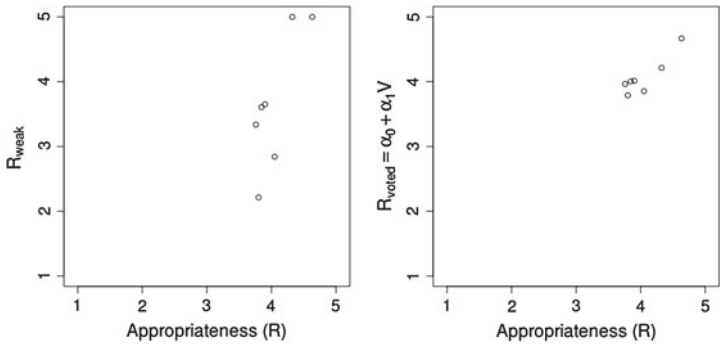

Fig. 4 Comparison between two automatic evaluation understudy measures at system level in static context setting

observed value of appropriateness rating (R) as judged by humans. The Pearson's correlation between R_{voted} and R is 0.479 ($p < 0.001, n = 397$) at the individual response level. At the system level the Pearson's correlation between R_{voted} and R is 0.893 ($p < 0.01, n = 7$). At the system level, R_{voted} is a better evaluation understudy than R_{weak}. Figure 4 shows a comparison between these two possible evaluation measures for automatic evaluation of appropriateness in static context setting.

3.3 Discussion

Different resources are required to build different automatic evaluation measures. For R_{weak}, we need to collect wizard data. When this data is being collected at the surface text level, we need a substantial number of wizards (four or more) each selecting a

large number of appropriate responses for each context. For the automatic evaluation measure R_{voted}, in addition to the wizard data we need resources to estimate the linear regression model. As training data to build a linear regression model, we need human evaluators' appropriateness ratings for responses given the dialogue contexts.

Automatic evaluation for static context setting involves human efforts for collecting wizard data and appropriateness ratings. But since the resources are collected at the surface text level *non-experts* can accomplish this task. An appropriate tool which can ensure a wide variety of appropriate responses proves useful for this task. Moreover since static context setting uses a fixed set of contexts, wizard data collection needs to be performed only once. The resulting automatic evaluation metrics can be used to compare different dialogue models.

When using the *Voted Appropriateness* evaluation method, the training data used for linear regression should represent all possible responses adequately. The data used to fit our model includes relatively well-performing models which results in a rather high intercept value of 3.549. For any model producing responses that are not judged appropriate by any of the wizards, our model would predict the appropriateness value of 3.549 which seems rather high.

4 Conclusion

In this paper, we evaluated a previously proposed automatic evaluation metric for dialogue coherence models, *Weak Agreement* in terms of how closely it correlates with human judgments. We also proposed and evaluated a new metric, *Voted Appropriateness* and showed that it has better correlation with human judgments. We also introduced a taxonomy for evaluation which is useful in understanding how various dialogue model evaluations relate to each other.

Acknowledgments The effort described here has been sponsored by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Bickmore TW, Pfeifer LM, Jack BW (2009) Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: Proceedings of the 27th international conference on Human factors in computing systems, CHI '09. ACM, New York, NY, USA, pp 1265–1274. doi:10.1145/1518701.1518891. http://doi.acm.org/10.1145/1518701.1518891
2. DeVault D, Leuski A, Sagae K (2011) Toward learning and evaluation of dialogue policies with text examples. In: Proceedings of the SIGDIAL 2011 conference. Association for Computational Linguistics, Portland, Oregon, pp 39–48. http://www.aclweb.org/anthology/W/W11/W11-2006
3. Forbes-Riley K, Litman DJ (2006) Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In: Proceedings

of the main conference on human language technology conference of the North American chapter of the association of computational linguistics, HLT-NAACL '06. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 264–271. http://dx.doi.org/10.3115/1220835.1220869
4. Gandhe S, Traum D (2008) Evaluation understudy for dialogue coherence models. In: Proceedings of the 9th SIGdial workshop on discourse and dialogue. Association for Computational Linguistics, Columbus, Ohio, pp 172–181. http://www.aclweb.org/anthology/W/W08/W08-0127
5. Gandhe S, Traum D (2013) Surface text based dialogue models for virtual humans. In: Proceedings of the SIGDIAL 2013 conference. Association for Computational Linguistics, Metz, France, pp 251–260. http://www.aclweb.org/anthology/W/W13/W13-4039
6. Gandhe S, Traum D (2014) SAWDUST: a semi-automated wizard dialogue utterance selection tool for domain-independent large-domain dialogue. In: Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL). Association for Computational Linguistics, Philadelphia, PA, USA, pp 251–253. http://www.aclweb.org/anthology/W14-4333
7. Gustafson J, Bell L, Boye J, Lindström A, Wirén M (2004) The nice fairy-tale game system. In: Strube M, Sidner C (eds) Proceedings of the 5th SIGdial workshop on discourse and dialogue. Association for Computational Linguistics, Cambridge, Massachusetts, USA, pp 23–26
8. Levin E, Pieraccini R, Eckert W (1997) Learning dialogue strategies within the Markov decision process framework. In: Proceedings of the 1997 IEEE workshop on automatic speech recognition and understanding, pp 72–79. doi:10.1109/ASRU.1997.658989
9. Lin CY, Hovy E (2003) Automatic evaluation of summaries using n-gram co-occurrence statistics. In: NAACL '03: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology. Association for Computational Linguistics, Morristown, NJ, USA, pp 71–78. http://dx.doi.org/10.3115/1073445.1073465
10. Papineni KA, Roukos S, Ward T, Zhu WJ (2001) Bleu: a method for automatic evaluation of machine translation. In: Technical report RC22176 (W0109-022), IBM Research Division. http://citeseer.ist.psu.edu/papineni02bleu.html
11. Swartout W, Traum D, Artstein R, Noren D, Debevec P, Bronnenkant K, Williams J, Leuski A, Narayanan S, Piepol D, Lane C, Morie J, Aggarwal P, Liewer M, Chiang JY, Gerten J, Chu S, White K (2010) Ada and grace: toward realistic and engaging virtual museum guides. In: Proceedings of the 10th international conference on Intelligent virtual agents, IVA'10. Springer, Berlin, pp 286–300. http://dl.acm.org/citation.cfm?id=1889075.1889110
12. Traum D, Leuksi A, Roque A, Gandhe S, DeVault D, Gerten J, Robinson S, Martinovski B (2008) Natural language dialogue architectures for tactical questioning characters. In: Proceedings of 26th army science conference
13. Traum D, Swartout W, Gratch J, Marsella S (2005) Virtual humans for non-team interaction training. In: AAMAS-05 workshop on creating bonds with humanoids
14. Traum D, Swartout W, Gratch J, Marsella S (2008) A virtual human dialogue model for non-team interaction. Text, speech and language technology, vol 39. Springer, New York, pp 45–67. doi:10.1007/978-1-4020-6821-8
15. Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460. http://cogprints.org/499/
16. Walker M, Kamm C, Litman D (2000) Towards developing general models of usability with paradise. Natural language engineering: special issue on best practice in spoken dialogue systems. http://citeseer.ist.psu.edu/article/walker00towards.html
17. Williams JD, Young S (2007) Partially observable Markov decision processes for spoken dialog systems. Comput Speech Lang 21:393–422