Open-Domain Dialog Evaluation using Follow-Ups Likelihood

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, Walter Daelemans
CLiPS Research Center
University of Antwerp, Belgium
maxime.debruyn@uantwerpen.be

Abstract

Automatic evaluation of open-domain dialogs remains an unsolved problem. Moreover, existing methods do not correlate strongly with human annotations. This paper presents a new automated evaluation method using follow-ups: we measure the probability that a language model will continue the conversation with a fixed set of follow-ups (e.g. *Not really relevant here*, *What are you trying to say?*). When compared against twelve existing methods, our new evaluation achieves the highest correlation with human evaluations.

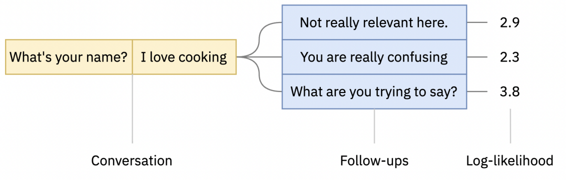

Figure 1: Illustration of our method. We measure the probability (log-likelihood) that a language model will continue the conversation with a set of predefined follow-ups. This paper shows that the sum of the individual log-likelihoods correlates strongly with human evaluations.

1 Introduction

Despite the recent progress in Natural Language Processing, the automatic evaluation of open-domain conversations remains an unsolved problem. It is difficult to establish criteria to measure the quality of a system. Task-oriented dialog systems use metrics such as task success or dialog efficiency. However, these do not apply to open-domain conversational agents (McTear, 2020).

Currently, there are two options for open-domain dialog evaluation: human evaluation and automated evaluation. Thanks to their understanding of natural language, humans are able to digest the entire dialog context in order to meaningfully evaluate a response (Mehri et al., 2022). Human evaluation also has its shortcomings: inconsistency in ratings (the same annotator may give two different scores depending on the mood), lack of reproducibility, and cost (Mehri et al., 2022).

The second option is to use automated evaluation metrics. Methods inherited from sequence-to-sequence machine translation such as BLEU (Papineni et al., 2002) evaluate the generated utterance by comparing it to the ground-truth. By doing so, these methods miss the one-to-many characteristic of conversation: a conversation may evolve in more than one valid direction.

To tackle this problem, researchers came up with reference-free evaluation metrics: the generated utterance is not compared to a ground truth but evaluated on its own.

FED (Mehri and Eskenazi, 2020a) is an unsupervised reference-free evaluation metric. It uses the idea that one can use the next utterance in a conversation to rate the turn before it. When users speak to a system, their response to a given system may implicitly provide feedback for the system. FED uses a set of predefined follow-ups and the log-likelihood from a language model to measure 18 fine-grained attributes in a conversation.

Inspired by the FED metric, we propose a new evaluation method called FULL (Follow-Up Log-Likelihood). We start by explaining our method and how it departs from the original FED metric. Next, we explain our choice of language model and follow-ups. Finally, we demonstrate that our new method achieves the highest correlation with human evaluations compared to 12 automated metrics. We open-source our evaluation code[1] and publish FULL as a Python package[2] for easy usage.

[1] https://github.com/maximedb/full

[2] https://pypi.org/project/full/

Proceedings of the 29th International Conference on Computational Linguistics, pages 496–504
October 12–17, 2022.

2 Related Work

This section reviews the existing literature on evaluation metrics for open-domain conversations. In the interest of space, we limit ourselves to studying reference-free methods (methods that do not require a ground truth). The interested reader is encouraged to read Yeh et al. (2021) for a full review.

GRADE (Huang et al., 2020) and DynaEval (Zhang et al., 2021) use a graph-based structure to model the dialog-level interaction between a user and a system. DynaEval distinguishes between well-formed dialogs from carefully constructed negative samples. MAUDE (Sinha et al., 2020) is also trained to distinguish a correct response from a randomly sampled negative response using a contrastive loss. FlowScore (Li et al., 2021) evaluates the quality of a dialog using the dynamic information flow in the dialog history.

USR (Mehri and Eskenazi, 2020b) trains several models to measure different qualities of dialogs. A masked language modeling head measures the fluency of the conversation, a retrieval model determines the relevance of a response, and a fact-to-response model checks whether a response conditions on knowledge. USL-H (Phy et al., 2020) also has three internal models, although they measure different attributes: grammatical correctness, sensibleness, and the likelihood of a given response. Other notable evaluation methods include Ghazarian et al. (2020); See and Manning (2021); Ghazarian et al. (2022b,a)

FED (Mehri and Eskenazi, 2020a) and HolisticEval (Pang et al., 2020) both use GPT-like (Radford et al., 2019) models to evaluate conversation on several attributes. FED computes the likelihood of manually designed follow-up utterances to measure multiple dialog qualities without supervision. HolisticEval uses a GPT-2 model to measure coherence, fluency, diversity, and consistency.

3 Method

Our metric FULL (Follow-Up Log-Likelihood) is a reference-free evaluation method for dialogs inspired by FED (Mehri and Eskenazi, 2020a). Figure 1 provides an overview.

3.1 Follow-Up Utterance for Evaluation

Our method uses follow-up utterances to evaluate the quality of a conversation (Eskénazi et al., 2019). When interacting with a system, users may provide implicit feedback about the conversation in the semantics of their response. For example, if a user ends a conversation with *It was a pleasure talking to you*, we can reasonably assume it was a pleasant conversation. On the other hand, if a user ends a conversation with *What are you talking about?*, we could conclude that the user is confused about the state of the conversation.

3.2 Log-Likelihood of Follow-Ups

We do not have access to the next utterance in an interactive setting. Instead, we ask a language model to play the role of a human. We ask the model how likely it is to generate a fixed set of follow-ups. For example, if the language model is likely to continue a conversation with the follow-up *I don't understand what you are saying*, we could conclude that the utterance generated by the system does not make sense.

FULL analyzes the quality of a response r in the context of a dialog history h with a language model M and a set of n predefined follow-ups F. For each predefined follow-up, the language model computes the log-likelihood D of a follow-up utterance f_i given the dialog history.

$$\sum_{i=1}^{n} D(h, r, f_i) \qquad (1)$$

The total score is equal to the sum of the individual log likelihoods. It is worth reminding that the metric does not mean anything. It is only useful to *compare* systems together.

3.3 Differences with FED

Our implementation differs from FED (Mehri and Eskenazi, 2020a) in multiple ways. First, we do not consider fine-grained attributes, only the overall quality of the turn or dialog.[3]

Second, FED computes the log-likelihood of the conversation history h, the response r, and the follow-up f_i. Whereas we only compute the conditional log-likelihood of the follow-up f_i. Computing the log-likelihood over the conversation introduces a bias towards the dataset used in training the language model, Reddit, in the case of FED. It also

[3] Whereas FED considers 18 fine-grained attributes (overall quality included). Our initial experiments revealed that follow-ups assigned to a fine-grained attribute (e.g., engaging) often had a higher correlation with another unrelated attribute (e.g., correctness). For that reason, we choose to focus on a single attribute, the conversation's overall quality and leave the study of fine-grained attributes for future work.

favors longer conversations over shorter ones. Our goal is to estimate the likelihood of the follow-up, not the conversation itself.

Third, FED did not justify its choice of follow-ups, while we studied each candidate and only took the most correlated ones making intuitive sense. Fourth, we also study multiple types of language models (conversational and general).

4 Experimental Settings

This section explains our choices of follow-ups, language models, and conversational data. Our goal is to find the combination of language models and follow-ups correlating the most with human evaluations.

4.1 Follow-Ups

A follow-up is an utterance added after a conversation's last turn to evaluate the last turn or the entire dialog. FED defined 63 unique follow-ups in 16 categories (fine-grained attributes) at the turn level and the dialog level. Appendix B list the entire list of follow-ups. The authors did not provide any justification for their choice of follow-ups. Instead of blindly using the list of follow-ups, we attempt to understand which of these follow-ups have the highest correlation with human evaluations.

4.2 Language Models

We experiment with several language models, both general and conversational. The goal of the language module is to compute the conditional log-likelihood of several follow-ups.

BlenderBot v1 is a conversational sequence-to-sequence model (Roller et al., 2020) with three sizes: small, large, and extra-large. A distilled version is also available on HuggingFace.

DialoGPT is a conversational language model (Zhang et al., 2020) with three sizes: small, medium and large. The authors fine-tuned a GPT-2 model on a large corpus of Reddit conversations.

GPT-2 is a general language model (Radford et al., 2019). While it was not trained specifically on conversational data, our experiments revealed its potential to estimate a conversation's quality.

4.3 Conversational Data

We use the FED dataset (Mehri and Eskenazi, 2020a) for evaluating the set of follow-ups. It consists of 372 turn-level (124 dialog-level), originally collected by Adiwardana et al. (2020). The dataset consists of human-system conversations (Meena and Mitsuku) and human-human conversations. Mehri and Eskenazi (2020a) asked annotators to evaluate turn-level and dialog-level conversations on several attributes. In this work, we only use the evaluation of the overall quality of the turn or dialog.

Follow-up	**Correlation**	
	Turn	**Dialog**
Not really relevant here.	0.48	0.65
You're really confusing.	0.46	0.67
I don't understand what you're saying.	0.46	0.58
That's not really relevant here.	0.45	0.70
You are so confusing.	0.45	0.64
You're really boring.	0.44	0.65
That's not very interesting.	0.44	0.60
That was a really boring response.	0.43	0.63
You don't seem interested.	0.43	0.61
I am so confused right now.	0.43	0.57

Table 1: Top 10 follow-ups ranked by Spearman correlation to human evaluations. All follow-ups exhibit a positive relationship, meaning that the likely presence of the follow-up (low log-likelihood) entails a low human evaluation and vice-versa.

5 Results

Our objective is to find the best combination of language models and follow-ups. We start by analyzing which language model correlates the most with human evaluation. In the second step, we look for the best set of follow-ups.

5.1 Choice of Language Model

We are looking for a language model whose log-likelihood of generating the follow-ups correlates highly with human evaluations. We do so both on a turn-level and dialog-level. We compare the average absolute correlation of each follow-up with human judgments. The results are displayed on Figure 2 in Annex A. The model standing out is the large Blender model (Roller et al., 2020). It has the highest correlation with humans both on a turn-level and dialog-level. The difference in performance between Blender-3B and Blender-400M is small. For these reasons, we choose Blender-400M as our default language model.

5.2 Choice of Follow-ups

Now that we have identified our model of choice (Blender-400M), we wish to identify the follow-

	Turn Level	**Dialog Level**
QuestEval	0.09	0.08
MAUDE	-0.09	-0.28
DEB	0.19	-0.01
GRADE	0.12	-0.06
DynaEval	0.32	0.55
USR	0.12	0.06
USL-H	0.19	0.15
DialoRPT	-0.09	-0.21
HolisticEval	0.12	-0.30
PredictiveEngage	0.09	0.15
FED	0.09	0.32
FlowScore	-0.05	-0.00
FULL (ours)	**0.51**	**0.69**

Table 2: Comparison of our evaluation method FULL with other automated methods. FULL achieves the highest correlation on turn-level and dialog-level, followed by DynaEval. Except for FULL, results are copied from Yeh et al. (2021).

ups correlating the most with humans. We compute the Spearman correlation between each follow-up and human evaluation (turn-level and dialog-level). We present the top-10 follow-ups (by absolute correlation) in Table 1. The full table is available Appendix B.

The follow-up correlating the most on a turn-level basis is *Not really relevant here* with a Spearman correlation of 0.48. The least correlated follow-up is *Wow! That's really cool!* with correlations of 0.04. The follow-up correlating the most on a dialog-level basis is *That's not really relevant here* with a correlation of 0.70. The least correlated follow-up on a dialog level is *Cool! That sounds super interesting!* with a correlation of 0.01.

Most follow-ups exhibit a positive relationship, meaning that the likely generation of the follow-up by the language model (low log-likelihood) entails a low human rating and vice-versa. However, all the top follow-ups are *negative* follow-ups (e.g., *You're really confusing*), and their likely presence indicates a negative conversation. On the other hand, the *positive* follow-ups (e.g., *Great talking to you*) are not as highly correlated. On average, negative follow-ups correlate with 0.39, while positive follow-ups correlate with 0.24. These results indicate that the language model evaluates a good conversation by the likely absence of negative follow-ups.

Each follow-up brings another forward pass of the model, so ideally, we want to restrict the number of follow-ups in the final evaluation method. For the final selection of follow-ups, we combine the rank of the turn-level and dialog-level correlations and take the top 5.[4] The final selection of follow-ups is the following: *Not really relevant here. You're really confusing. You're really boring. What are you trying to say? You don't seem interested.*

5.3 Comparison

Yeh et al. (2021) compared 12 evaluation methods on the FED dataset (Mehri and Eskenazi, 2020a). We compare our method FULL against these 12 other methods in Table 2. The results are clear, FULL achieves the highest correlation both on a turn-level and dialog-level while being fully unsupervised (except in the choice of follow-ups). By combining the log-likelihood from 5 follow-ups, the average correlation on turn-level increases to 0.51, while the average of the individual correlation equals 0.45.

6 Conclusion

This short paper introduces a new automated evaluation method (FULL) for open-domain conversations. FULL measures the quality of a conversation by computing the probability that a language model will continue the conversation with a set of follow-ups (e.g., *Not really relevant here*, *What are you trying to say?*). FULL achieves the highest correlation with human evaluations compared to twelve other existing methods.

Our experiments revealed that negative follow-ups (e.g., *Not really relevant here*) have a higher correlation with human evaluations than positive follow-ups (e.g., *Wow, interesting to know*). It is easier for the model to evaluate a conversation from its bad angles rather than its good ones.

Future work is needed to know which fine-grained attribute can be measured using the same technique. Using ever-large models such as GPT-3 (Brown et al., 2020) or OPT (Zhang et al., 2022) could be a direction for future research, although the resulting model will likely need to be distilled to be of practical use.

[4] We arbitrarily choose the number 5. We also removed close duplicates. For example *Not really relevant here.* and *That's not really relevant here.*

Acknowledgement

We thank the reviewers for their helpful feedback. This research received funding from the Flemish Government under the *Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen* programme.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Maxine Eskénazi, Shikib Mehri, Evgeniia Razumovskaia, and Tiancheng Zhao. 2019. Beyond turing: Intelligent agents centered on the user. *CoRR*, abs/1901.06613.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022a. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. *arXiv*, abs/2203.13927.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7789–7796.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022b. Deam: Dialogue coherence evaluation using amr-based semantic manipulations. *arXiv*, abs/2203.09711.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Michael McTear. 2020. Conversational ai: dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3):1–251.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *arXiv*, abs/2203.10012.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *CoRR*, abs/2004.13637.

Abigail See and Christopher Manning. 2021. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *arXiv*, abs/2205.01068.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

A Appendix: Comparison of Models

We present in Figure 2 the average absolute correlation to human evaluations per model.

B Appendix: List of Candidate Follow-ups

Table 3 list the entire list of follow-ups considered.

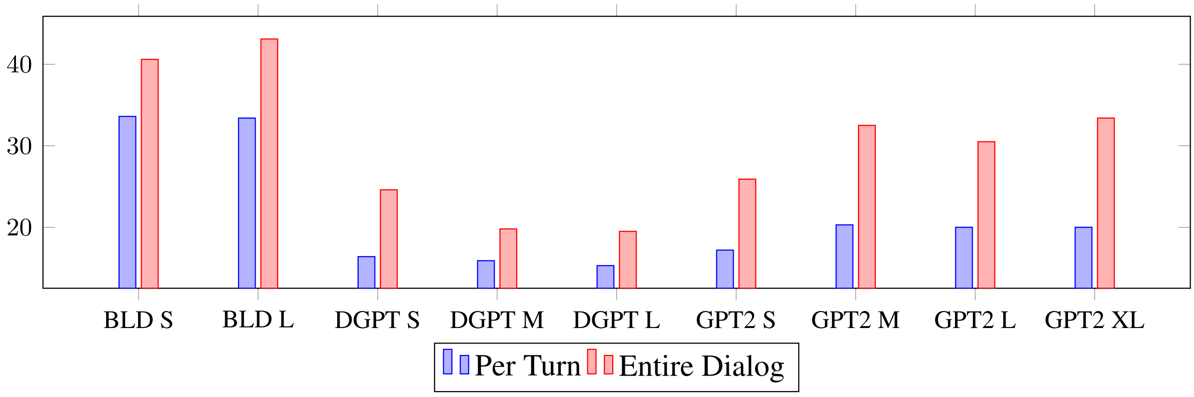

Figure 2: Average absolute correlation with human evaluations for several language models. We use Blender-400 (BLD S) as language model because of its high correlation with human evaluations. For space reasons, Blender is abbreviated as BLD and DialoGPT as DGPT.

	Follow-up	Category	Level	Type	Level	Dialog
X	Not really relevant here.	specific	turn	neg	0.48	0.65
X	You're really confusing.	error recovery	dialog	neg	0.46	0.67
	I don't understand what you're saying.	correct	turn	neg	0.46	0.58
	That's not really relevant here.	specific	turn	neg	0.45	0.70
	You are so confusing.	coherent	dialog	neg	0.45	0.64
X	You're really boring.	informative	dialog	neg	0.44	0.65
	That's not very interesting.	interesting	turn	neg	0.44	0.60
	That was a really boring response.	interesting	turn	neg	0.43	0.63
X	You don't seem interested.	inquisitive	dialog	neg	0.43	0.61
	I am so confused right now.	error recovery	dialog	neg	0.43	0.60
	I'm so confused!	understandable	turn	neg	0.43	0.59
	I don't really care. That's pretty boring.	engaging	turn	neg	0.43	0.61
	I want to talk about something else.	engaging	turn	neg	0.43	0.65
	That's not even related to what I said.	relevant	turn	neg	0.42	0.58
X	What are you trying to say?	understanding	dialog	neg	0.42	0.68
	I am so confused right now!	correct	turn	neg	0.42	0.57
	That makes no sense!	semantically appropriate	turn	neg	0.42	0.56
	I don't understand at all!	understandable	turn	neg	0.41	0.54
	That's really boring.	interesting	turn	neg	0.41	0.54
	I don't like you.	likeable	dialog	neg	0.40	0.58
	I'm so confused right now!	fluent	turn	neg	0.40	0.56
	Don't change the topic!	relevant	turn	neg	0.40	0.58
	You're not understanding me!	correct	turn	neg	0.40	0.62
	That's a very generic response.	specific	turn	neg	0.39	0.50
	You don't really know much.	informative	dialog	neg	0.39	0.52
	You're not very nice.	likeable	dialog	neg	0.38	0.56
	You're not very fun to talk to.	likeable	dialog	neg	0.37	0.55
	Is that real English?	fluent	turn	neg	0.37	0.49
	That's a lot of questions!	inquisitive	dialog	pos	0.36	0.52
	Why are you repeating yourself?	diverse	dialog	neg	0.35	0.50
	You're making no sense at all.	coherent	dialog	neg	0.35	0.43
	You ask a lot of questions!	inquisitive	dialog	pos	0.35	0.54
	Let's change the topic.	engaging	turn	neg	0.35	0.45
	You don't ask many questions.	inquisitive	dialog	neg	0.35	0.54
	Why are you changing the topic?	relevant	turn	neg	0.34	0.51
	Stop saying the same thing repeatedly.	diverse	dialog	neg	0.34	0.50
	Do you know how to talk about something else?	flexible	dialog	neg	0.33	0.49
	You're changing the topic so much!	coherent	dialog	neg	0.33	0.47
	You know a lot of facts!	informative	dialog	pos	0.32	0.48
	Tell me more!	engaging	turn	pos	0.32	0.34
	I like you!	likeable	dialog	pos	0.31	0.43
	Wow that's a lot of information.	informative	dialog	pos	0.31	0.38
	Stop changing the topic so much.	depth	dialog	neg	0.31	0.44
	What does that even mean?	understandable	turn	neg	0.30	0.35
	I don't want to talk about that!	flexible	dialog	neg	0.29	0.50

Table 3 continued from previous page

Follow-up	**Category**	**Level**	**Type**	**Level**	**Dialog**
That’s not what you said earlier!	consistent	dialog	neg	0.29	0.37
You have a good point.	semantically appropriate	turn	pos	0.29	0.43
I see, that’s interesting.	specific	turn	pos	0.28	0.31
Stop contradicting yourself!	consistent	dialog	neg	0.28	0.36
You’re very easy to talk to!	flexible	dialog	pos	0.28	0.40
Stop repeating yourself!	diverse	dialog	neg	0.27	0.40
That’s good to know. Cool!	specific	turn	pos	0.25	0.30
That’s a good point.	specific	turn	pos	0.25	0.34
Wow you can talk about a lot of things!	flexible	dialog	pos	0.23	0.27
I’m really interested in learning more about this.	engaging	turn	pos	0.22	0.26
That makes sense!	semantically appropriate	turn	pos	0.21	0.21
Thanks for all the information!	informative	dialog	pos	0.21	0.15
You’re super polite and fun to talk to	likeable	dialog	pos	0.17	0.23
Wow that is really interesting.	interesting	turn	pos	0.17	0.14
That’s really interesting!	interesting	turn	pos	0.16	0.11
Great talking to you.	likeable	dialog	pos	0.15	0.10
Cool! That sounds super interesting.	interesting	turn	pos	0.08	- 0.01
Wow! That’s really cool!	engaging	turn	pos	0.04	- 0.08

Table 3: List of candidate follow-ups along with their category (fine-grained attribute), positivity (negative of positive follow-up) and correlation with a human evaluation of the overall quality of the turn/dialog. All follow-ups and static data is from Mehri and Eskenazi (2020a).