Rating Prediction in Conversational Task Assistants with Behavioral and Conversational-Flow Features

Rafael Ferreira
NOVA University of Lisbon
NOVA LINCS
Lisbon, Portugal
rah.ferreira@campus.fct.unl.pt

David Semedo
NOVA University of Lisbon
NOVA LINCS
Lisbon, Portugal
df.semedo@fct.unl.pt

João Magalhães
NOVA University of Lisbon
NOVA LINCS
Lisbon, Portugal
jm.magalhaes@fct.unl.pt

ABSTRACT

Predicting the success of Conversational Task Assistants (CTA) can be critical to understand user behavior and act accordingly. In this paper, we propose TB-Rater, a Transformer model which combines conversational-flow features with user behavior features for predicting user ratings in a CTA scenario. In particular, we use real human-agent conversations and ratings collected in the Alexa TaskBot challenge, a novel multimodal and multi-turn conversational context. Our results show the advantages of modeling both the conversational-flow and behavioral aspects of the conversation in a single model for offline rating prediction. Additionally, an analysis of the CTA-specific behavioral features brings insights into this setting and can be used to bootstrap future systems.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → *Human computer interaction (HCI).*

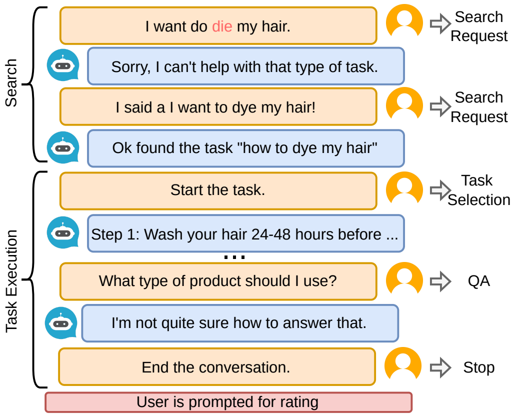
KEYWORDS

Rating Prediction, Conversational Task Assistants, NLP

ACM Reference Format:
Rafael Ferreira, David Semedo, and João Magalhães. 2023. Rating Prediction in Conversational Task Assistants with Behavioral and Conversational-Flow Features . In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592048

1 INTRODUCTION

Recently, Conversational Task Assistants (CTA) that are able to guide users through manual tasks are gathering more attention [6, 11, 24] due to their applicability in everyday routines. These differ and expand from other paradigms, such as conversational search [29] and task-oriented conversational agents [30]. In these paradigms, the user provides information to the assistant, and the system performs a task (e.g., searching or buying a ticket). In a CTA setting, it is the user that completes a task with the help of an assistant [11]. Creating these assistants requires various sub-systems working hand-in-hand to effectively help the user complete a variety of tasks such as "baking a cake" or "fixing a leaky faucet" [11]. Figure 1 illustrates a partial CTA dialog. First, users are prompted to search for the task they want to do, which is done using an IR step. Second, the user selects one of the provided tasks and enters the task-execution phase. Third, the task instructions are presented to the user. The user is then able to follow the task or create conversational sub-flows by asking task-specific or general questions, which the system should answer using domain knowledge. Due to the complex interactions between the user and the system, errors are prone to happen which in turn leads to user dissatisfaction and low ratings. Being able to predict the rating of an interaction is thus a critical step to understand the problems of the system, and act accordingly in both an online and offline setting [5, 26].



I want do die my hair.

Sorry, I can't help with that type of task.

I said a I want to dye my hair!

Ok found the task "how to dye my hair"

Start the task.

Step 1: Wash your hair 24-48 hours before ...

...

What type of product should I use?

I'm not quite sure how to answer that.

End the conversation.

User is prompted for rating

Figure 1: User and CTA example of a low-rated interaction. System and user utterances were emulated from real dialogs.

The aforementioned problems motivate our work in offline rating prediction, which is a challenging scenario, where the goal is to predict a rating at the end of the interaction taking into account the whole conversation. This task helps discover patterns in user ratings and, more importantly, detect problematic conversations which can be further analyzed to discover avenues for system improvement. In particular, and to the best of our knowledge, we are the first to tackle the problem of rating prediction in a Conversational Task Assistant (CTA) [11] setting. In Figure 1, we show an example of a low-rated CTA conversation. This happens due to various aspects, such as ASR errors recognizing "die hair" instead of "dye hair", or fallback responses, for example, when the system is not able to answer a question. How to use these various signals to predict the rating is one of our goals. To evaluate the rating prediction task, we leverage data collected during the Alexa Prize TaskBot

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592048

Challenge [8, 11], comprised of real human-agent CTA interactions. In this setting, the users interact with Alexa devices, mainly using their voice in a conversational, multi-turn, and multimodal way.

Evaluating conversational assistants is an active and challenging research subject in which the gold-standard metric is human-based evaluation [16, 21, 22]. Many works [3, 5, 15, 26] leverage this human-labeled data to train automatic methods for conversational assistants evaluation. For example, we highlight the design of manual features in [23] for a flight booking system, and [15] for search dialogs. In [3, 5, 26], models are proposed to automatically predict the rating/satisfaction on Alexa's SocialBot Challenge [20]. In particular, Choi et al. [5] show advantages in leveraging both textual and behavioral features. Motivated by this work, we created user behavior features that are specific to the CTA setting. Moreover, we use the recent advances with the use of Transformer-models [7, 19, 25] to create conversational-flow features. Despite the significant differences between chit-chat (SocialBot) and the CTA (TaskBot) setting, we believe that a combination of both types of features can bring improvements in rating prediction. With this, we combine the features into a single model which we call *TB-Rater* (*Transformer-Behavior Rater*) that surpasses the considered baselines. To conclude, we perform an ablation study, showing how the various design decisions influence the model's results, and analyze the importance of the behavior features in this novel setting.

2 TRANSFORMER-BEHAVIOR RATER

In this section, we present our proposed model *Transformer-Behavior Rater* (*TB-rater*), which combines two sets of features.

2.1 Model Architecture

2.1.1 Conversational-Flow Features. The content and flow of the dialog conveys information about the current state and rating. Thus, we propose to use conversational-flow features with the aim of capturing intricate and discriminative dialog flows. To model these features computationally, we use a Transformer-based [25] language model, which is able to capture various patterns in the language and derive a representation of the conversation's state [14, 28].
We represent each turn (T_i) of the dialog as follows:

$$T_i = \text{“}[S]\ [RG_i]\ S_i\ [U]\ [INT_i]\ U_i\text{”}, \tag{1}$$

where S_i and U_i are the system and user utterances, separated by special tokens *[S]* and *[U]* denoting the beginning of a speaker's turn. We go beyond the utterances and include flow-based information in the form of the intent detected $[INT_i]$, which has proven useful in [12, 27], and the response generator selected/activated $[RG_i]$ to provide extra information to the model.
A conversation with n turns is modeled as the sequence:

$$\text{“}[CLS]\ [DEV]\ [DOM]\ T_1 \dots T_i \dots T_n\text{”} \tag{2}$$

The first token of the sequence is a special $[CLS]$ token [7]. Specific to our model, we use additional special tokens denoting the type of device $[DEV]$ (screen/screen-less) and the domain of the user's task $[DOM]$, which can be none, a recipe, or a DIY [11]. We use all turns of the conversation and perform left truncation of the input when it is over the maximum sequence length. Finally, we use the embedding of the *[CLS]* token ($emb_{[CLS]}$) as the representation of the conversation. This first set of features is then complemented with user behavior features.

Table 1: General Features. A / in a feature denotes >1 feature.

General Feature	Description
SessionDuration	Duration in seconds
TurnDuration	Turn i duration in seconds
Avg/Max TurnDuration	Avg/Max turn duration
Turns	# Turns
Utterance Pos/Neg[1]	# User positive/negative utterances
AvgUtterance Pos/Neg	Avg user positive/negative utterances
Offensive/Sensitive[2]	# Turns w/ offensive/sensitive content
User/System WordOverlap	Word overlap ratio btw consecutive user/system
UserSystemWordOverlap	Word overlap ratio btw S_{i-1} and U_i
Avg User/System WordOverlap	Avg Word overlap ratio btw user/system
AvgUserSystemWordOverlap	Avg Word overlap ratio btw user and system
Words User/System	Total # words in U_i / S_i
AvgWords User/System	Avg # words in user/system
Unique User/System Words	Unique words btw consecutive user/system turns

Table 2: CTA-specific. A / in a feature denotes >1 feature.

CTA-Specific Feature	Description
StepsRead	# Steps read
Repeated User/System Utterance	# Repeated user/system utterances
Resumed	User resuming session
HasScreen	User is using a device w/ screen
Screens	# Screens visited
Searches	# Search requests
RepeatedSearches	# Repeated search requests
ResultPages	# Result pages seen
Started/Finished Task	User started/finished a task
FallbackExceptions	# System fallback responses
Domain	Domain of task (recipe, DIY, none)
Curiosities Accepted/Denied	# Curiosities user accepted/rejected
CuriositiesSaid	# Curiosities the system said
Phase-based	
Greeting/Search/Task Overview/ Ingredients/Steps/Step's Detail/Conclusion	# Turns in a particular phase
Intent-based	
Search/None of These/Cancel/Yes/No/ Ingredients/Start Cooking/Start Steps/Next/ Next Step/More Detail/Terminate Task/ Help/Repeat/Fallback	# Intents of particular type

2.1.2 Behavior Features. Taking inspiration from Choi et al. [5], which showed a performance increase when combining text and behavior features. We follow a similar pattern and add manually engineered features specific and unique to the CTA domain, with the aim of providing more domain context to the model. In particular, we use the last turn of the conversation (T_n) to get the behavior features (B_n). In total, we created 70 features divided in General, System-Induced, and CTA-Specific features.
General. Table 1 presents general conversational features, where we can see a large overlap with the features in Choi et al. [5].
System-Induced. We consider the values for a particular turn, the avg, and the max across the conversation for user latency, system latency, and scores given by the ASR model, as in [5].

[1] We created a threshold-based method to identify positive/negative utterances based on an internal Amazon Alexa algorithm that uses the audio of the utterance.
[2] Offensive and sensitive content is identified by an internal Amazon Alexa classifier and by matching with a list of special words.

Table 3: Alexa TaskBot Dataset statistics.

	Train	Validation	Test
# Conversations	1344	168	169
# Turns	12784	1390	1567
Avg # Turns	9.5	8.3	9.2
Rating 1	263 (19.6%)	34 (20.2%)	30 (17.8%)
Rating 2	158 (11.8%)	19 (11.3%)	23 (13.6%)
Rating 3	179 (13.3%)	23 (13.7%)	24 (14.2%)
Rating 4	231 (17.2%)	26 (15.5%)	30 (17.8%)
Rating 5	513 (38.2%)	66 (39.3%)	62 (36.7%)

CTA-Specific. In Table 2, we propose CTA-specific features, such as the number of searches or steps read, the number of turns in a phase, which indicates the depth the user is going into the conversation, or the counts of a specific intent as predicted by another model. These features were designed based on real-world interactions and thus can serve as a basis for other works in this setting.

2.1.3 *Features Combination.* First, we use two feed-forward neural networks (FFNN), $FFNN_T$ and $FFNN_B$ (with ReLu activations), that take as input the $emb_{[CLS]}$ and the behavior features B_n, respectively. After this, the resulting representations are concatenated and passed through a final $FFNN_{TB}$ that combines the two streams:

$$FFNN_{TB}(FFNN_T(emb_{[CLS]}) \oplus FFNN_B(B_n)). \tag{3}$$

With this approach, we make predictions benefiting from both information streams, as shown in [5, 17] in different domains. The model is then trained using the cross-entropy loss.

3 EXPERIMENTS

3.1 Experimental Setting

3.1.1 *Alexa Prize TaskBot Dataset.* To evaluate our models, we use internal data collected in the first Alexa Prize TaskBot challenge [8, 11]. This challenge focuses on developing a CTA that helps users perform real-world manual tasks in the cooking and DIY domains. It is also the first multimodal challenge of this type, combining both voice-only and voice-and-screen interactions. In this setting, the system interacted with thousands of users, and for each conversation, at the end of the interaction, they are asked to provide an optional rating on a 1 to 5 scale. However, only about 10% of the users provide a rating, making it hard to pinpoint which conversations require more attention, further motivating our work.

We used a stable version of the system to collect ratings and considered only rated conversations with a minimum of 3 turns. In total, we used 1681 conversations which we separated into training (90%), validation (10%), and test (10%) sets. The statistics of the dataset are in Table 3. We observe that, on average, a dialog has 8 to 9 turns, with a standard deviation of 6.8, indicating a large variety of conversation lengths. In terms of the ratings, we see a larger concentration in 1 and 5, with a standard deviation of 1.55, indicating that the users generally have a strong opinion about the system's performance, as also noticed in the SocialBot domain [3].

3.1.2 *Task and Metrics.* In this work, we define the task of rating prediction at the end of the interaction. This makes this task challenging due to the need for a model capable of understanding the entire conversation, and identify the non-trivial subtleties that contribute to the rating.

Following a similar approach to Choi et al. [5], we use a binary classification task by separating ratings 1-3 into 0 and 4-5 into 1, instead of using the original 1-5 rating scale. In terms of metrics, we considered accuracy (Acc), precision (P), recall (R), and F1.

3.2 Methods and Baselines

Behavioral-only - we tested the following methods *RandomForest* [2], *AdaBoost* [9], *Bagging* [1], *GradientBoosting* [10], *XGBoost* [4], and *LogisticRegression.* All methods are implemented using *sklearn* [18] and use the behavior features of the last turn.
Conversational-Flow-only - To encode the dialog features, we used a BERT model [7] with a classification head. We also adapted a T5 [19] model for classification.
Conversational-Flow and Behavior - we implemented *ConvSat* [5], which combines text features at an utterance and character levels using BiLSTMs [13], which are combined with behavioral features. We also present the results of the proposed *TB-Rater* model.[3]

3.3 Results

3.3.1 *General Results.* We present the results of the various methods on the Alexa TaskBot Dataset in Table 4. First, we observe that the best behavior-only method is the *SVM.* Regarding conversational-flow-only methods, the *BERT-Base* model achieves the best results, surpassing the enc-dec model *T5.* This might be explained by BERT having a specific and pre-trained classification token [7], while T5 is adapted to classification using a text-to-text paradigm [19]. The *BERT-Base* approach also surpasses the best behavior-only method (*SVM*), showing that only using conversational-flow information may be a good alternative for rating prediction, avoiding the need for the design of domain-specific features. Comparing the conversational-flow and behavior models, we see that the best results are achieved by the proposed *TB-Rater* model, surpassing all of the considered baselines. This result is in line with previous work [5, 17] that showed advantages in combining text and behavior features. However, *ConvSat* [5], which also uses both types of features, did not perform as well. We believe this may be due to having a small amount of training data to effectively train the character and word level embeddings, making Transformer-based models a more robust approach. To conclude, the results show that it is possible to have conversation-flow-only models that are on par with classic approaches based on manually engineered features. We also show that combining both types of features in *TB-Rater* brings an improvement in performance.

3.3.2 *Ablation Study.* In Table 5, we analyze how our design decisions influence the model's results. As we saw previously, removing behavioral features negatively affects the results. In *w/o Step Token*, we keep the text of the task's step instead of replacing it with a special token $[STEP]$. We see a decrease in performance, which we attribute to the step text not being especially important for the rating. Adding to this, keeping the text of a step also decreases the number of turns inputted into the Transformer model due to steps typically being long. In *w/o Additional Tokens*, we remove

[3] Code available at https://github.com/rafaelhferreira/cta_rating_prediction

Table 4: Avg. result of 3 runs on the Alexa TaskBot test set.

Method	Acc	P	R	F1
Behavior-Only				
RandomForest	66.3	66.0	65.8	65.8
AdaBoost	64.5	64.2	63.7	63.7
Bagging	67.1	66.8	66.8	66.8
GradientBoosting	66.3	66.0	66.0	66.0
XGBoost	64.5	64.2	64.1	64.1
LogisticRegression	67.5	67.5	66.4	66.4
SVM	68.6	68.4	68.2	68.3
Conversational-Flow-Only				
BERT-Base	69.0	69.2	69.1	68.8
T5-Base	66.9	67.2	67.1	66.8
Conversational-Flow and Behavior				
ConvSat [5]	63.4	61.9	61.4	63.3
TB-Rater (Ours)	**69.6**	**70.0**	**70.0**	**69.6**

Table 5: *TB-Rater* ablation study on Alexa TaskBot test set.

Method	Acc	P	R	F1
TB-Rater	**69.6**	**70.0**	**70.0**	**69.6**
w/o Behavior (i.e., BERT BASE)	69.0	69.2	69.1	68.8
w/o Step Token	66.7	68.0	67.3	66.3
w/o Additional Tokens	65.9	66.4	66.0	65.5
Right Side Truncation	63.9	64.2	64.2	63.9

the special tokens pertaining to the device, domain, intent, and response generator, but we keep the special [$STEP$]. Again, we see that adding extra information in the form of these tokens increases performance. Finally, we test the *TB-Rater* model but truncate inputs larger than the maximum input size from the *right side* (end of the conversation) instead of the left side. Here, we observe the worse results out of all methods. This result shows that focusing on the end of the conversation is more important to predict the rating, this can be attributed to the last turns having more impact than the ones at the beginning, indicating a possible recency bias.

3.3.3 Error Analysis. While user subjectivity plays an important role [3, 26], we believe that a portion of the model's errors can be categorized. Thus, we analyze *TB-Rater*'s 50 error cases (counts of error types are given between parentheses). We noticed that the model generally gives a low rating if the interaction is stopped early, but the user is able to find and/or start a task (12). Another mistake is when the user starts a task that is different from the one the user is looking for but still goes further into the task, usually with consecutive dull responses (e.g., next step). In this case, the model predicts a high rating despite the user giving a low one (9). There were also cases where despite the system giving a considerable number of fallback answers, the conversation still moves forward, however, the model predicts this as an unsatisfactory conversation (10). Finally, user ratings have a lot of variability, and some do not seem to reflect how the interaction went, for example, "throw-away"/bad interactions that returned high ratings (7), or interactions where the user is not impressed with the system, returning a low rating despite the system responding to every request correctly (12). These results reaffirm the volatility of user ratings [3, 5] and the difficulty of the task, shedding light on the most common error cases.

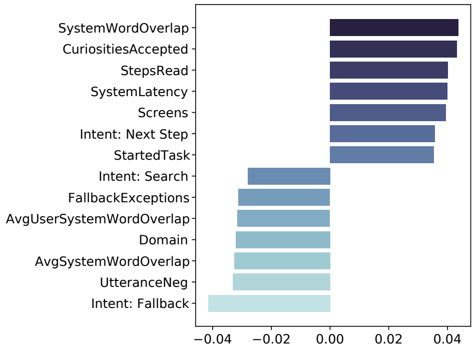


Figure 2: Logistic Regression Top-14 absolute coefficients.

3.3.4 Behavior Feature Importance. In Figure 2, we present the top-14 abs. feature coefficients for the *Logistic Regression* model. Here positive/negative scores indicate a feature that predicts a positive/negative rating. Starting with the *system word overlap* on the last turn, this indicates that the last two system utterances share a large number of words. This feature is relevant because when the user finishes a task there is a large token overlap. The higher *system latency* on the last turn also appears to have importance in a positive rating, which at first seems counter-intuitive. After a closer analysis, we attribute this to the last turn of a finished task having a larger latency while an abrupt stop has a latency value of zero. In practice, these two features indicate that finishing a task is an important signal for predicting the rating. Other features such as the number of *steps read*, *next step*, and *started task* suggest that the user is engaged with the system and going deeper into a task.

Regarding the negative coefficients, we see that a larger number of *fallbacks* leads to a lower rating. The *average system overlap* denotes that the system is saying a similar response in multiple turns, which might indicate that the user is stuck. Finally, a higher value of *domain* indicates that the user did not search for a task, and in opposition, a high *number of searches* indicates that the user is struggling to find a task, resulting in a lower rating. It is also worth noting that out of the 14 features, 9 are from the CTA-specific set, showing the relevance of the proposed features.

4 CONCLUSION

In this paper, we propose *TB-Rater*, a model that combines conversational flow and behavioral features to perform rating prediction in the novel CTA setting. We show the advantages of combining both types of features by evaluating on human-agent interactions collected in the Alexa TaskBot challenge. Moreover, we provided a comprehensive set of CTA-specific features and measured their importance. The model proposed can be used to estimate a rating, which may allow for the discovery and prioritization of system errors. In future work, we intend to apply the model in an online setting, using its predictions to change the course of a conversation.

ACKNOWLEDGMENTS

This work has been partially funded by the Amazon Science - TaskBot Prize Challenge 2021, by the NOVA LINCS project Ref. UIDP/04516/2020, by FCT Ref UI/BD/151261/2021, and by the CMU| Portugal iFetch project LISBOA-01-0247-FEDER-045920.

REFERENCES

[1] Leo Breiman. 1996. Bagging Predictors. *Mach. Learn.* 24, 2 (1996), 123–140. https://doi.org/10.1007/BF00058655

[2] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Alessandra Cervone, Enrico Gambi, Giuliano Tortoreto, Evgeny A. Stepanov, and Giuseppe Riccardi. 2018. Automatically Predicting User Ratings for Conversational Systems. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018 (CEUR Workshop Proceedings, Vol. 2253)*. CEUR-WS.org. http://ceur-ws.org/Vol-2253/paper32.pdf

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 785–794. https://doi.org/10.1145/2939672.2939785

[5] Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and Online Satisfaction Prediction in Open-Domain Conversational Systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. ACM, 1281–1290. https://doi.org/10.1145/3357384.3358047

[6] Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. Wizard of Tasks: A Novel Conversational Dataset for Solving Real-World Tasks in Conversational Settings. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*. International Committee on Computational Linguistics, 3514–3529. https://aclanthology.org/2022.coling-1.310

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Rafael Ferreira, Diogo Silva, Diogo Tavares, Frederico Vicente, Mariana Bonito, Gustavo Goncalves, Rui Margarido, Paula Figueiredo, Helder Rodrigues, David Semedo, and Joao Magalhaes. 2022. TWIZ: A conversational Task Wizard with multimodal curiosity-exploration. In *Alexa Prize TaskBot Challenge Proceedings*.

[9] Yoav Freund and Robert E. Schapire. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* 55, 1 (1997), 119–139. https://doi.org/10.1006/jcss.1997.1504

[10] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[11] Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, Prerna Dwivedi, Hangjie Shi, Lucy Hu, Andy Huang, Luke Dai, Bofei Yang, Varun Somani, Pankaj Rajan, Ron Rezac, Michael Johnston, Savanna Stiff, Leslie Ball, David Carmel, Yang Liu, Dilek Hakkani-Tur, Oleg Rokhlenko, Kate Bland, Eugene Agichtein, Reza Ghanadan, and Yoelle Maarek. 2022. Alexa, let's work together: Introducing the first Alexa Prize TaskBot Challenge on conversational task assistance. In *Alexa Prize TaskBot Challenge Proceedings*. https://www.amazon.science/publications/alexa-lets-work-together-introducing-the-first-alexa-prize-taskbot-challenge-on-conversational-task-assistance

[12] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Measuring User Satisfaction on Smart Speaker Intelligent Assistants Using Intent Sensitive Query Embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. ACM, 1183–1192. https://doi.org/10.1145/3269206.3271802

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/e946209592563be0f01c844ab2170f0c-Abstract.html

[15] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. ACM, 45–54. https://doi.org/10.1145/2911451.2911521

[16] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016*. ACM, 121–130. https://doi.org/10.1145/2854946.2854961

[17] Yue Ma, Zengfeng Zeng, Dawei Zhu, Xuan Li, Yiying Yang, Xiaoyuan Yao, Kaijie Zhou, and Jianping Shen. 2019. An End-to-End Dialogue State Tracking System with Machine Reading Comprehension and Wide & Deep Classification. *CoRR* abs/1912.09297 (2019). arXiv:1912.09297 http://arxiv.org/abs/1912.09297

[18] 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 (2011), 2825–2830.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[20] 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. 2018. Conversational AI: The Science Behind the Alexa Prize. *CoRR* abs/1801.03604 (2018). arXiv:1801.03604 http://arxiv.org/abs/1801.03604

[21] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding User Satisfaction with Task-oriented Dialogue Systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM, 2018–2023. https://doi.org/10.1145/3477495.3531798

[22] Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI, ConvAI@ACL 2022, Dublin, Ireland, May 27, 2022*. Association for Computational Linguistics, 77–97. https://doi.org/10.18653/v1/2022.nlp4convai-1.8

[23] Stefan Steidl, Christian Hacker, Christine Ruff, Anton Batliner, Elmar Nöth, and Jürgen Haas. 2004. Looking at the Last Two Turns, I'd Say This Dialogue Is Doomed - Measuring Dialogue Success. In *Text, Speech and Dialogue, 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004, Proceedings (Lecture Notes in Computer Science, Vol. 3206)*. Springer, 629–636. https://doi.org/10.1007/978-3-540-30120-2_79

[24] Carl Strathearn and Dimitra Gkatzia. 2022. Task2Dial: A Novel Task and Dataset for Commonsense-enhanced Task-based Dialogue Grounded in Documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc@ACL 2022, Dublin, Ireland, May 26, 2022*. Association for Computational Linguistics, 187–196. https://doi.org/10.18653/v1/2022.dialdoc-1.21

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[26] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing open domain dialog systems. *arXiv preprint arXiv:1801.03625* (2018).

[27] Liu Yang, Minghui Qiu, Chen Qu, Cen Chen, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, and Haiqing Chen. 2020. IART: Intent-aware response ranking with transformers in information-seeking conversation systems. In *Proceedings of The Web Conference 2020*. 2592–2598.

[28] Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards Fully End-to-End Task-Oriented Dialog System with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 14230–14238. https://ojs.aaai.org/index.php/AAAI/article/view/17674

[29] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808* (2022).

[30] Zheng Zhang, Ryuichi Takanobu, Minlie Huang, and Xiaoyan Zhu. 2020. Recent Advances and Challenges in Task-oriented Dialog System. *CoRR* abs/2003.07490 (2020). arXiv:2003.07490 https://arxiv.org/abs/2003.07490