INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Human-to-Human Conversation Dataset for Learning Fine-grained Turn-taking Action

Kehan Chen, Zezhong Li, Suyang Dai, Wei Zhou, Haiqing Chen

Alibaba Group, China
kehan.ckh@alibaba-inc.com

Abstract

Conducting natural turn-taking behavior takes a crucial part in the user experience of modern spoken dialogue systems. One way to build such system is to learn those behaviors from real-world human-to-human dialogues, which have the most diverse and fine-grained turn-taking actions than any manual constructed sessions.

In this paper, we propose a Dataset - FTAD which could be used to learn turn-taking policies directly from human. First, we design an annotation mechanism to transform existing human-to-human dialogue session into structural data with most fine-grained turn-taking actions reserved. Then we explored a set of supervised learning tasks on it, showing the challenge and potential of learning complete fine-grained turn-taking policies based on such data.

Index Terms: turn-taking, spoken dialogue systems, dataset

1. Introduction

Building a human-like spoken dialogue system (SDS) requires not only precise dialogue engine but also a smart timing controller, i.e. turn-taking policy. Researches have shown that a good turn-taking policy contributes to the users' experience on dialogues with robots [1, 2, 3].

One of the common usage of SDS is the personal assistants on smart devices like Siri or Amazon Alexa. The major turn-taking problem of such SDS is to take the floor and respond quickly and precisely after user's turn. To achieve this goal, a number of recent studies are done for End-of-Turn (EOT) prediction [4, 5, 6, 7, 8, 9]. While recently, apart from smart device based personal assistant, another form of SDS encountered a rapid expansion in industry - the telephone-based chatbot like Google Duplex for personal or Alime-Hotline of Alibaba for customer service. One major difference between smart device based and telephone-based SDS is the telephone-based sessions are much more active with fast turn switches like human-to-human.

Turn-taking behaviors between humans are complex and fine-grained [10]. A number of turn switches could happen inside one semantic round. Human would act to the mutual silence and overlapping during speech, and generate backchannels or fillers among responses. It is intuitive to build a human-like turn-taking policy supporting all fine-grained actions by learning from human-to-human dialogues directly.

Most of recent studies of turn-taking behaviors are based on datasets constructed basically in the following ways:

- Human-to-human dialogues created in written language,e.g.MultiWOZ [11],Taskmaster [12].
- Generated dialogues from user simulator [13, 14, 15].
- Recorded spoken dialogue by crowd-sourcing workers [7, 9].

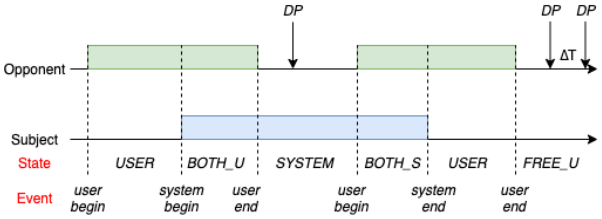

Figure 1: *Showcase of State, Event and DP in dialogue. Note the DP is not just right after IPU because of the VAD gap.*

Each of those ways have some drawbacks: Datasets created in written language have a large gap with spoken scenarios, both in wording and turn-taking behaviors. Quality of generated dialogues are limited by the performance of simulator, which is still remains a challenging area. Record and annotate spoken dialogue has the best quality but also with highest cost, which limits the size of the corpus.

Therefore, in this paper, we propose an automatic annotation mechanism which could construct turn-taking data from existing dialogue audios with very low cost. The annotation pipeline could transform the spoken dialogues with fine-grained turn-taking actions (referred as FTA) into structural data, which enables modeling such behavior by linguistic features as well as acoustic features. Then we applied the pipeline to Switchboard [16] to create and release the *Fine-grained Turn-taking Dataset on Switchboard* (FTAD-sw)[1]. To the best of our knowledge, FTAD-sw is so far the largest structural fine-grained turn-taking corpus released. Through this we hope to give a standard resource for later turn-taking behavior researches. Finally, We explored the methodology of learning a full-functional turn-taking policy, by defining and evaluating a set of tasks over FTAD-sw as well as a self-owned customer service dataset FTAD-zh.

2. Fine-grained Turn-taking Action Dataset

2.1. Structural Representation of FTA

To capture every fine-grained turn-taking action in human-to-human dialogues, we first need a self-contained label system. The behavior of each dialogue can be represented in four elements: *State, Event, Decision Point, Action.* Here reminds that the representation is first-person perspective, which means the four elements describe the session from the perspective of one participant (referred as *Subject* later). And of course we can generate the representation for both sides symmetrically.

State stands for the speaking status of any specific time pieces. Here we use the six-states FST described by Raux et. al. [17] with little adjustment. Here $SYSTEM$ means the

[1]FTAD-sw Dataset: https://github.com/alimehotline/FTAD

http://dx.doi.org/10.21437/Interspeech.2021-994

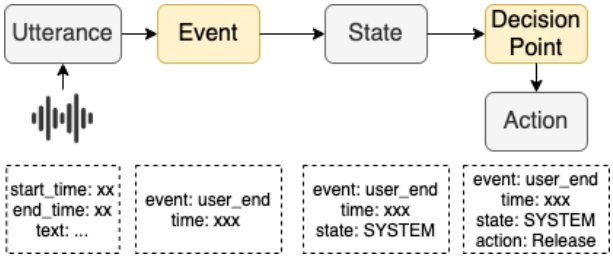

Figure 2: *Annotation pipeline and output of each step.*

$Subject$ speaking and $Opponent$ is listening while $USER$ means the opposite. Overlapping and silence can be noted as $BOTH_S/BOTH_U$ and $FREE_S/FREE_U$ with subscript to distinguish previous state.

Event is the moment of state change, specifically the moment $Subject$ or $Opponent$ start or stop speaking.

Decision Point (referred as DP), in some research called Transition Relevance Point (TRP) [10] , means the place where turn-taking would but not necessarily happen. From a system perspective, decision point is the timing for a robot's turn-taking policy to decide which action to take. A dialogue system can only act in discrete moments, in this paper we use voice activity detection (VAD) of a short silence (200ms) to split continuous speech into Inter-pasual Unit (IPU) [5] , as it's a common solution to get aligned speech with linguistic result by ASR system. Decisions could be made after every $Opponent's$ IPU or every ΔT at silence. Figure1 displays the relation of $State$, $Event$ and DP.

Action is the actual behavior of the $Subject$ at each decision point. The optional actions at decision points are constrained by the state at the moment as described in Table 1.

2.2. Self-contained Turn-taking Annotation

The goal we develop the automatic annotation process is to transform existing dialogue audios into structural data presented above, with least manual work. With such process we can easily construct large turn-taking corpus with any accessible audio data set from various scenarios like face-to-face talk, customer service dialogue over telephone or even from TV-shows.

The process goes in four-steps: First, the audio is cut into IPUs with transcribed text using an ASR system with a small VAD threshold. The IPUs are referred as $Utterances$ here. Then the $Utterances$ are sorted based on start time so that $Events$ and $States$ could be annotated deterministically according to their definitions. After this DPs are generated from the Subject's perspective. And finally $Actions$ are annotated to each DP automatically base on the context. The pipeline and output of each step are shown in Figure2.

The rules of annotation of the $Actions$ is a little trivial. Most of the $Actions$ could be inferred through context, e.g. a long query after opponent's $Utterances$ are annotated as $Grab\_Response$ and DP's on the silence between opponent's $Utterances$ are annotated as $Wait$. We more details as below:

- We shrink short overlapping (less than 1 second) between long $Utterances$, to avoid meaningless turn-switch (shown in Figure3).
- It's guaranteed that each $Utterance$ of subject should be triggered by one $Grab\ Action$ at one DP and each DP should be assigned with a valid $Action$.
- We annotate $Grab\_Backchannel$ by matching a specific vocabulary.
- To align $Utterances$ with discrete DPs, we adjust the timestamp of $Actions$ within a threshold (under 1000ms) of bias (shown in Figure3).
- For the rest of unaligned $Utterances$ are assigned special actions either $Grab\_Response\_Break$ or $Grab\_Backchannel\_Break$ according to the texts.

We visualized the annotation of an audio fragment in Figure4 with white textboxes as $Utterances$, gray as $DP/Action$ and green as unaligned $Utterances$.

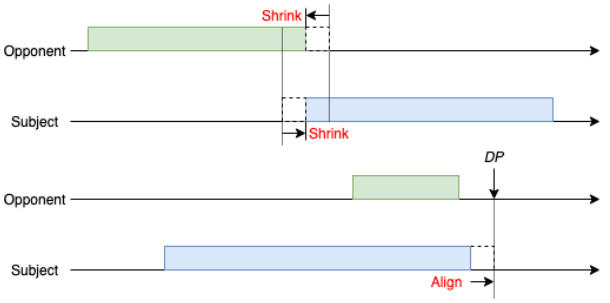

Figure 3: *Upper: shrink short overlapping. Lower: align the subject's utterance with the closest DP as the its is breaked by the opponent's utterance*

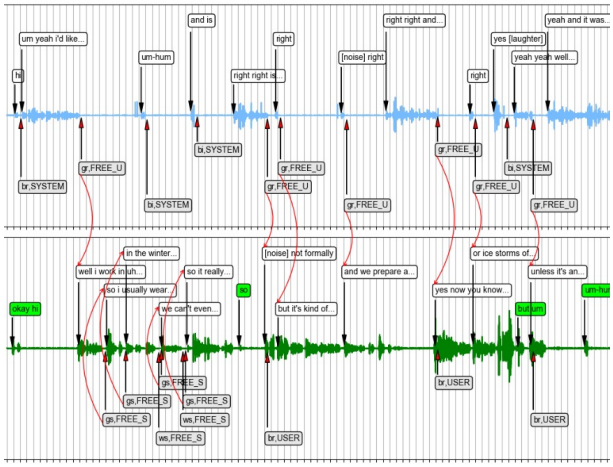

Figure 4: *Visualization of the structural annotation of FTA*

2.3. FTAD Corpus

We applied the annotation pipeline to Switchboard-1 [16], with a more than 2400 sessions over telephone under some casual topics to create FTDA-sw (released this time). As it's transcribed with word alignments for ASR model training, we directly used the word boundary transcripts to cut $Utterances$ with a 400ms threshold. To explore the the turn-taking behaviors between difference languages and domains, we create the FTDA-zh based on 3900 customer service sessions of Alime-Hotline by the same pipeline.

The comparison of statistics about the two datasets are listed in Table2 and Table 3. As dialogue about casual topics,

Table 1: *State and action space*

State	Action	Description
$FREE_S$	$Grab\_Silence$	Take turn if opponent keeps silence, mostly sending reminder or echo.
	$Wait\_Silence$	Release turn and wait opponent's response.
$FREE_U$	$Grab\_Response$	Take turn and respond to opponent.
	$Grab\_Backchannel$	Send a backchannel to opponent (not willing to take turn).
	$Wait$	Keep silence and wait opponent to finish his turn.
$SYSTEM$	$Break\_Ignore$	Opponent tries to break the subject, but ignore and keep speaking.
	$Break\_Release$	Accept opponent's break and stop speaking.
	$Keep$	Subject tries breaking the opponent, and keep speaking at this DP.
	$Release$	Subject tries breaking the opponent, and give up (stop speaking) at this DP.
$USER$	$Grab\_Response\_Break^*$	Subject starts to break the opponent (at any time).
	$Grab\_Backchannel\_Break^*$	Subject sends backchannel to the opponent (at any time).

FTAD-sw tends to have more $Utterances$ with a higher speaking speed and actual turn-taking (DP without $Wait$) appears more frequently.

As for the action annotations, nearly 94% actions could be aligned with a DP with an average bias of 420ms for FTAD-sw. For FTAD-zh the numbers are 98.3% and 351ms. It means that, from the systematic perspective, most of the fine-grained turn-taking actions of human could be simulated by a robot with discrete DPs. FTAD-sw has faster turn-switches with shorter mutual silence, which leads to less $Wait/Wait\_Silence$ actions. 'Break' and 'Backchannel' happens frequently in FTAD-sw, while in FTAD-zh, the customer service consulting dialogues are more organized.

Table 2: *Statistics of two FTAD corpuses*

	FTAD-sw	FTAD-zh
# of sessions	2438	2959
Avg. session length	6.33min	1.72min
Word (Character) per minute	201	351
# of $Utterances$ per session	141.20	27.83
# of $Utterances$ per minute	22.27	16.14
# of DPs per session	195.32	118.17
# of DPs per minute	30.85	68.53
# of DPs(w/o $Wait$) per minute	13.43	9.47
Avg. DP alignment error	420ms	351ms

Table 3: *Distribution of Actions*

Actions	FTAD-sw	FTAD-zh
$Grab\_Response$	23.7%	8.79%
$Grab\_Backchannel$	6.1%	0.11%
$Grab\_Silence$	0.03%	0.0%
$Break\_Ignore$	3.77%	1.39%
$Break\_Release$	2.17%	0.97%
$Keep$	0.56%	0.51%
$Release$	0.88%	0.14%
$Grab\_Backchannel\_Break^*$	4.71%	0.74%
$Grab\_Response\_Break^*$	1.34%	1.08%
$Wait\&Wait\_Silence$	56.66%	86.1%

In Table4 we compared FTAD-sw to other datasets which are used for turn-taking modeling in past studies [4] from several perspective. Comparing to native-oral datasets [18, 19, 20], datasets collected from IM-based chat logs [12, 11, 21, 22] have a large gap in both language style and turn-taking behavior. Meanwhile, most of these datasets only contain $Wait\&Response$ actions, with more fine-grained actions absent or ignored. In paper [13, 14], Khouzaimi et. al. used user simulator to generate utterances with turn-taking actions, however the corpus quality and diversity is limited by the simulator's performance. Aldeneh et. al. [23] and Tomer et. al. [24] used Switchboard along with NXT-format[25] annotation to construct turn-switch prediction dataset, but they don't distinguish different turn-taking actions.

Table 4: *Datasets comparison*

Dataset	Style	Domain	Actions
Taskmaster	written	assistant	W&R
MultiWOZ_2	written	assistant	W&R
MetaLWOz	written	assistant	W&R
CCPE [26]	written&oral	assistant	W&R
DailyDialog	written	assistant	W&R
Maptask	oral	assistant	W&R
MRDA	oral	open	W&R
SWDA	oral	open	W&R
Generated session	written	assistant	Fine-grained
FTDA-sw	oral	open	Fine-grained

3. Fine-grained Turn-taking Modeling Task

Based on FTAD, we defined four supervised learning tasks for policy modeling.

End of turn prediction has the similar definition to past studies [4], to decide whether the $Opponent$ has finished his turn and it's time to respond. We generate the task data set by extracting continuous $Opponent's\ Utterances$ before a $Grab_Response$ action, with the last DP (trigger the response action) as positive label and other (trigger $Wait$ actions) as negative.

Respond time prediction models the expected time for the $Opponent$ to respond the $Subject's$ question or statement. This is useful as for SDS, which should be a threshold to determine whether to send a echo/reminder, when the $Opponent$ keeps silence. The prediction model could be used to set proper silence threshold based on the context. We used the gap between $Grab\_Response$ action and the last $Opponent's$

$Utterance$ before it as the prediction target. As the DPs are discrete, we define the task as a three-fold classification problem, using two quantiles: 400ms/800ms for FTAD-sw and 1200ms/2400ms for FTAD-zh.

Break prediction is defined as modeling the behavior of the $Subject$ when he's speech is interrupted by the $Opponent$. We construct the training set by using the $Utterances$ before the DP where the interruption happens, with $Break\_Release$ as positive sample and $Break\_Ignore$ as negative. Note that we only leverage the $Utterances$ before DP because for $Break\_Release$ cases, the complete sentence is unknown. In Figure 5 we illustrate this task.

Backchannel prediction is to imitate the $Subject's$ behavior during a long speech of the $Opponent$. We'd like to know the best timing to trigger a backchannel. Unlike other tasks, as a part of backchannel actions cannot be aligned with DPs, we construct backchannel data set in a word-level sequential labeling format, as shown in Figure5. The word right before the backchannel in $Opponent's\ Utterances$ is labeled as 1 (trigger word) while other labeled as 0, and the pause between utterances is treated as a special token. As the sparsity of backchannel in FTAD-zh, we only generate task data for FTAD-sw.

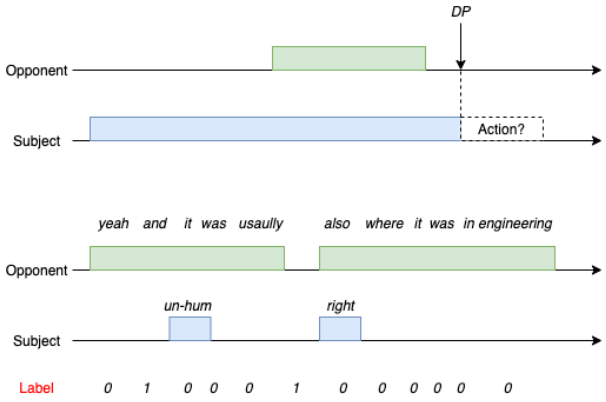

Figure 5: *Upper: Break action prediction task. Lower: Sequence Labeling task for Backchannel action.*

Some details of the four tasks are illustrated in Tabel 5. Note the way of learning turn-taking policy from FTAD is not limited in the tasks mentioned above. An alternative way is to learn a unified end-to-end policy model. And in this research we focus on linguistic features only, but it's easy to get aligned acoustic features as all the $Utterances$ and $Actions$ are labeled with timestamp.

Table 5: *Task summaries*

Task	Task type	Size(sw)	Size(zh)
EOT	classification	149618	40693
Respond time	classification	52099	10260
Break	classification	28321	8288
Backchannel	sequence labeling	29294	-

4. Experiment Result

We give a baseline for the *End of Turn*, *Respond time prediction* and *Break prediction* tasks using the context-aware self-attention model from [27]. Each $Utterance$ is encoded through a bi-GRU, with token embedding from an pre-trained Roberta [28]. Then the token-wise encoding is aggregated into a single vector through a context-ware self-attention, and a classification layer is applied on it (illustrated in Figure 6). The context utterances are concatenated along with a role token. For backchannel prediction, we simply used a LSTM-CRF based model to generate label for each token.

Table 6 shows the model performance over different tasks. For EOT and Break prediction, the user behavior is more divinable in FTAD-zh than FTAD-sw, as the later one has more complex and frequent turn-switches. However the respond time is more correlated with the context in the dialogues from FTAD-sw. And specially, for backchannel, the behavior is more arbitrary and most unpredictable, which has the lowest accuracy of positive sample.

We also conduct experiments of single utterance as model input as well as using three previous utterances as context for the classification tasks. And the result of three classification tasks the effectiveness of involving more context in predict turn-taking actions.

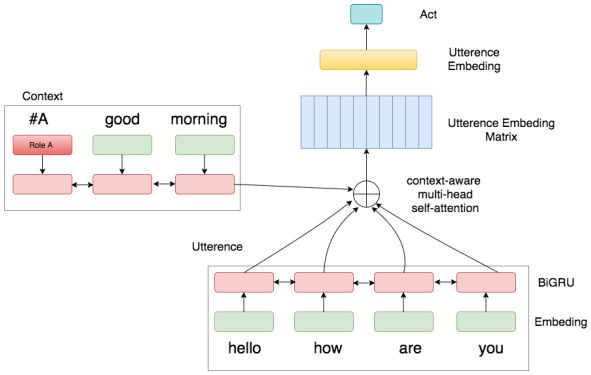

Figure 6: *Contextual attentive model for turn-taking action prediction.*

Table 6: *Model performance over tasks*

Task	Acc.(sw)	Acc.(zh)
EOT	0.6829	0.8273
EOT (with context)	0.6835	0.8398
Break	0.7142	0.7704
Break (with context)	0.7552	0.7737
Respond time	0.7230	0.5322
Respond time (with context)	0.7303	0.5928
Backchannel	0.3609	-

5. Conclusion

In the paper, we propose a structural representation of fine-grained turn-taking behaviors in human-to-human dialogues. Then we introduce an automatic annotation mechanism to transform dialogue audios into structural data and release FTAD-sw constructed on Switchboard. And finally we explore ways to model turn-taking policy based on FTAD. In the future, we want to evaluate the effectiveness of acoustic features in turn-taking modeling of FTAD and another option is to learn a unified end-to-end turn-taking model supporting all actions based on FTAD.

6. References

[1] R. Meena, G. Skantze, and J. Gustafson, "The map task dialogue system: A test-bed for modelling human-like dialogue," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 366–368.

[2] A. Cafaro, N. Glas, and C. Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 2016, pp. 911–920.

[3] D. Lala, S. Nakamura, and T. Kawahara, "Analysis of effect and timing of fillers in natural turn-taking." in *INTERSPEECH*, 2019, pp. 4175–4179.

[4] E. Ekstedt and G. Skantze, "Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 2981–2990.

[5] D. Lala, K. Inoue, and T. Kawahara, "Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 226–234.

[6] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, "Prediction of turn-taking using multitask learning with prediction of backchannels and fillers," *Listener*, vol. 162, p. 364, 2018.

[7] ——, "Turn-taking prediction based on detection of transition relevance place." in *INTERSPEECH*, 2019, pp. 4170–4174.

[8] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi, and T. Kawahara, "Attentive listening system with backchanneling, response generation and flexible turn-taking," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 127–136.

[9] S. Z. Razavi, B. Kane, and L. K. Schubert, "Investigating linguistic and semantic features for turn-taking prediction in open-domain human-computer conversation." in *INTERSPEECH*, 2019, pp. 4140–4144.

[10] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn taking for conversation," in *Studies in the organization of conversational interaction*. Elsevier, 1978, pp. 7–55.

[11] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," *arXiv preprint arXiv:1810.00278*, 2018.

[12] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, D. Duckworth, S. Yavuz, B. Goodrich, A. Dubey, A. Cedilnik, and K.-Y. Kim, "Taskmaster-1: Toward a realistic and diverse dialog dataset," *arXiv preprint arXiv:1909.05358*, 2019.

[13] H. Khouzaimi, R. Laroche, and F. Lefèvre, "Reinforcement learning for turn-taking management in incremental spoken dialogue systems." in *IJCAI*, 2016, pp. 2831–2837.

[14] ——, "A methodology for turn-taking capabilities enhancement in spoken dialogue systems using reinforcement learning," *Computer Speech & Language*, vol. 47, pp. 93–111, 2018.

[15] T. Zhao, A. W. Black, and M. Eskenazi, "An incremental turn-taking model with active system barge-in for spoken dialog systems," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 42–50.

[16] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," https://doi.org/10.35111/sw3h-rw02.

[17] A. Raux and M. Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 629–637.

[18] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller *et al.*, "The hcrc map task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.

[19] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The icsi meeting recorder dialog act (mrda) corpus," INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA, Tech. Rep., 2004.

[20] D. Jurafsky and E. Shriberg, "Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13 daniel jurafsky*, elizabeth shriberg+, and debra biasca** university of colorado at boulder &+ sri international," 1997.

[21] S. Kim, M. Galley, C. Gunasekara, S. Lee, A. Atkinson, B. Peng, H. Schulz, J. Gao, J. Li, M. Adada *et al.*, "The eighth dialog system technology challenge," *arXiv preprint arXiv:1911.06394*, 2019.

[22] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[23] Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6159–6163.

[24] T. Meshorer and P. A. Heeman, "Using past speaker behavior to better predict turn transitions." in *Interspeech*, 2016, pp. 2900–2904.

[25] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, no. 4, pp. 387–419, 2010.

[26] F. Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi, "Coached conversational preference elicitation: A case study in understanding movie preferences," in *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, 2019.

[27] V. Raheja and J. Tetreault, "Dialogue act classification with context-aware self-attention," *arXiv preprint arXiv:1904.02594*, 2019.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.