

Projection of Turn Completion in Incremental Spoken Dialogue Systems

Erik Ekstedt

KTH Speech, Music and Hearing
Stockholm, Sweden
erikekst@kth.se

Gabriel Skantze

KTH Speech, Music and Hearing
Stockholm, Sweden
skantze@kth.se

Abstract

The ability to take turns in a fluent way (i.e., without long response delays or frequent interruptions) is a fundamental aspect of any spoken dialog system. However, practical speech recognition services typically induce a long response delay, as it takes time before the processing of the user's utterance is complete. There is a considerable amount of research indicating that humans achieve fast response times by projecting what the interlocutor will say and estimating upcoming turn completions. In this work, we implement this mechanism in an incremental spoken dialog system, by using a language model that generates possible futures to project upcoming completion points. In theory, this could make the system more responsive, while still having access to semantic information not yet processed by the speech recognizer. We conduct a small study which indicates that this is a viable approach for practical dialog systems, and that this is a promising direction for future research.

1 Introduction

One of the most fundamental conversational behaviour of any spoken dialog system (SDS) is that of turn-taking, i.e., to take turns without long response delays or frequent interruptions (Skantze, 2021). To achieve this, the system must be able to correctly identify when the user is yielding the turn, and it is appropriate to make a response, and when the user is simply making a mid-utterance pause.

In their seminal work, Sacks et al. (1974) describe general properties of human-human conversation in which they observe that, overwhelmingly, one speaker talk at a time and the time between consecutive turns (response time) is minimal. For the English language, a typical response time is around 200ms and similar response patterns seem to be consistent across different cultures (Stivers et al., 2009). Contrary to this, current

SDSs typically have response delays of around 700-1000ms. The reason for this is that they typically rely solely on this silence to determine when to take the turn, whereas humans also use other cues, such as prosody, gaze and syntactic completeness (Skantze, 2021). Many studies have investigated how to include such features in turn-taking models for SDSs (Ferrer et al., 2002; Sato et al., 2002; Schlangen, 2006; Raux and Eskenazi, 2008; Meena et al., 2013; Maier et al., 2017; Lala et al., 2019).

Another difference between human turn-taking and SDSs is that humans do not only *react* to turn-yielding cues from the interlocutor. If they were simply waiting for a cue and only then started to formulate a response, psycholinguistic research has estimated that the response time would be around 600-1500ms (Levinson and Torreira, 2015), which is substantially slower than the observed response times. This indicates that humans also *project* turn completions in advance, before the turn is complete (Sacks et al., 1974; Levinson and Torreira, 2015; Garrod and Pickering, 2015).

In this paper, we investigate whether the human ability to project future turn completions could be a viable option for conversational systems to achieve more fluent turn-taking. We constrain our approach to the textual domain using a pre-trained conversational language model to project future words and turn-completions.

The projection of turn-completions in SDSs can have a number of applications. For example, the system could initiate a turn just before the end of the user's utterance to minimize response time, or even take the turn with a small overlap. It could also give the system more time to generate a response, or be used to address the problem of processing delays. For example, SDSs rely heavily on Automatic Speech Recognition (ASR) to extract the text from the user's speech. Most ASR services are associated with a certain latency (Baumann

et al., 2017; Addlesee et al., 2020). For turn-taking, this means that even if the system has detected that the user has stopped speaking, it is hard to determine whether the turn is yielded or not, since the final ASR result is not complete yet.

There has been some previous research on predicting upcoming activity in dialog, such as recognizing NLU intents on incomplete user speech (DeVault et al., 2009), projecting prosodic information and timing (Ward et al., 2010; Baumann and Schlangen, 2011) as well as estimating future voice activity (Skantze, 2017; Roddy et al., 2018; Ward et al., 2018). However, we are not aware of any previous studies of how a SDS could predict upcoming words in the user’s speech, and use this for managing turn-taking.

2 Conversational agent

For our study, we implemented a SDS that performs an interview with a user, talking about past travel memories, similar to Johansson et al. (2016). The reason we chose this domain is that the dialog manager can be implemented in a very simple way, while the turn-taking can be challenging, as pauses within the user’s turn might be more frequent than in, for example, a Q/A system. An example dialog can be found in Appendix A.1.

A general first step for modelling responsive turn-taking is to use an incremental dialog architecture, where the user’s speech is processed incrementally, so that decisions can be made in a more continuous fashion (Schlangen and Skantze, 2009). For this study, we build upon the recent Retico (Michael, 2020) framework (implemented in Python¹), which implements the general, abstract model of incremental dialog processing proposed by Schlangen and Skantze (2009).

The system processes incoming user speech and outputs audio. The incoming incremental audio chunks are processed by a local voice activity detection (VAD) component and streamed to a remote incremental ASR service (Google). The VAD triggers on silences of 200ms which defines interpausal units (IPU).

A user turn is started when both the VAD detects ongoing speech and the ASR has provided its first hypothesis. If the VAD module activates during an ongoing agent utterance, an interruption component is triggered. This module checks how much of the planned audio has been transmitted and stops

¹<https://github.com/Uhlo/retico>

the ongoing utterance if less than 80% has been sent. The interrupted utterance is then repeated for the system’s next response. If the agent completed an utterance and the user is inactive for 5 seconds, a fallback is triggered and the agent continues the conversation by producing a new utterance.

For the simplicity of our experiment, the dialog manager is defined by a set of predetermined questions, where the only possible deviation occurs if the user provides a too short utterance. If such a short utterance is recognized, the system randomly chooses from a set of paraphrased responses that encourages the user to elaborate.

In this study, we implement two different turn-taking policies: the **baseline** and the **projection** model. The baseline defines a user turn as complete once the VAD module is inactive and the ASR has produced its final hypothesis.

3 Turn-completion projection model

To make projections, we utilize the TurnGPT model by Ekstedt and Skantze (2020), which is a pre-trained GPT-2 (Radford et al., 2019) language model (LM) fine-tuned on conversational data. The model was trained on data from seven publicly available dialog datasets listed in Appendix A.2. The model trained until the validation loss reached a minimum, resulting in an average validation perplexity of 17.6.

The model includes special tokens that encode speaker shifts, which we will refer to as turn-completions. As shown by Ekstedt and Skantze (2020), the model does not only consider the ongoing user turn, but also benefits from taking the larger dialog context into account (i.e., previous turns by the system and the user).

Given the currently recognized user words (and the dialog context), a set of N possible continuations (of length M) are generated (using a temperature τ and topk sampling). The number of those that include turn-completions are counted, which gives a ratio. This ratio then approximates the probability of an “actual” turn-completion point in the near future. If the ratio is larger than a threshold R , the turn is predicted to be complete.

In this setup we strive towards simplicity and only trigger a projection at the end of each user IPU. However, if new ASR hypotheses are received after this, new projections are made until the system decides to take the turn. The projection model uses a maximum silence threshold T as a fallback, which

triggers a response regardless of the projections.

These different parameters can potentially be fine-tuned for the specific application (or user). This was not done in our study, and we selected values we found reasonable in preliminary tests, which are shown in Table 1.

An example taken from one of the interactions is illustrated in Figure 1

Parameter	Value
IPU	0.2 s
Turn-completion ratio (R)	0.4
Fallback threshold (T)	1.25 s
Sampling	
Continuations (N)	10
Length (M)	3
topk	5
Temperature (τ)	1.0
max context	70

Table 1: The parameters for the model.

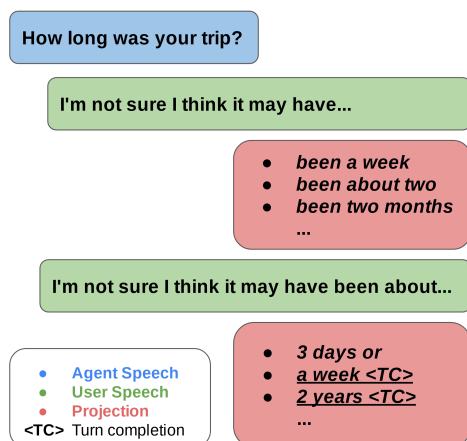


Figure 1: Illustration of language projection. The blue box represents the agent and the green boxes the recognized user words at two projection moments. The red boxes show a subset of projections made by the LM.

4 Experiment

To evaluate the model, we conducted an experiment over Zoom² where ten participants had two conversations each with the agent (testing the two turn-taking policies) about two distinct travel memories. The participants were asked to choose a memory prior to each agent interaction. We used two sets of paraphrased questions, assigned randomly between the two policies. After completing

²<https://zoom.us/>

both interactions, the participants were asked to annotate the recorded dialogues by labeling moments where they felt they had been interrupted by the system. To do this, they were provided with a graphical tool where they could see the waveforms of the dialogs and play them, as well as inserting labels.

The agent interacted directly over Zoom by connecting its microphone to the zoom speakers and vice versa. All audio was recorded directly on the agent side, in the same way as in a live setup.

5 Results

10 subjects interacted with the system, resulting in a total of 20 interactions, with an average duration of 3 minutes and 43 seconds. The number of questions varied by the amount of triggered elaboration requests. The baseline agent asked the users to elaborate 33 times, almost double the amount of 17 for the projection model. A transcript of an interaction is shown in Appendix A.1.

The total number of agent shifts (transitions between the user and the agent) was 220 for the baseline and 210 for the projection model. The duration of these (i.e., *response times*) are shown in the histogram in Figure 2. The average response times were 1.03 and 0.80 seconds for the baseline and projection agent, respectively. While this difference is not very large, it should be noted that the prediction model has a bimodal distribution (as seen in Figure 2), representing early predicted turn shifts and fallbacks. Thus, the model is able to take the turn quickly at some points, while allowing for more time at others.

The users annotated 18 of the agent shifts as interruptions for the baseline, and 28 for the projection model. The estimated average *cut-in rate*, defined as the annotated interruptions divided by the number of agent shifts, was 0.08 for the baseline and 0.13 for the projection model.

When evaluating the performance of a turn-taking model, both response time and cut-in rate should be taken into account (i.e., both should be minimized) (Raux and Eskenazi, 2008). However, there is typically also a trade-off between these two factors. Since both these values were different between the baseline and prediction model, they are difficult to compare directly.

One way of doing that is to perform an analysis of what would happen if we reduce the maximum allowed response time (for the prediction model

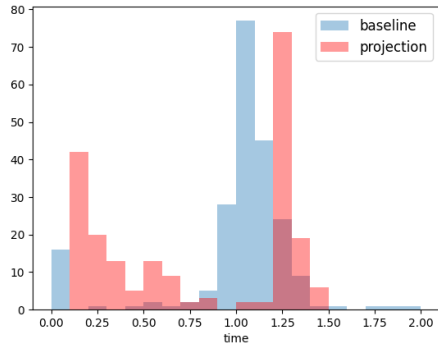


Figure 2: A histogram over the response times for each agent.

this is the parameter T). As we do this, the average response time will also be reduced, while the cut-in rate will increase, since silences in between user IPUs longer than T become both additional cut-ins and agent shifts. The result of this analysis is shown in Figure 3.

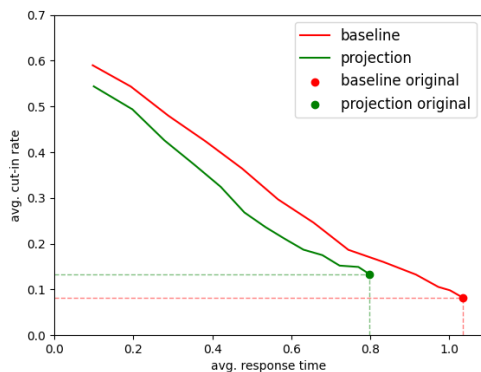


Figure 3: Cut-in rate vs response time. The points represent the aggregate values over the interactions and the lines the estimated performance given varying values of T .

This analysis enables a direct comparison of the agents over values where both lines are defined. The figure shows that the prediction agent is more responsive and produces less interruptions by the fact that the green line is strictly below the red. The greatest difference occurs at around 0.48s on the x-axis, with a cut-in rate difference of 0.1, given threshold values of 0.5 and 0.6 seconds for the baseline and projection agents, respectively.

6 Discussion

To our knowledge, all previous work on end-of-utterance-detection in SDSs have relied on mod-

els that are specifically trained with data from the target domain. Contrary to this, we have used a generic LM (TurnGPT) with a set of basic parameters that were not fine-tuned using domain data. If the LM and the parameters would be fine-tuned, we could expect further improvements. An analysis of the perplexity of the LM on the recorded data shows a rather high perplexity ($ppl \approx 80$). Another obvious improvement would be to also include prosodic features.

An important question we have not addressed here is how good the projections are in terms of predicting the last words more exactly (i.e., not just how well the system predicts whether there will be a turn completion). Depending on the domain of the system, this might be more or less important. In this respect, the comparison of the baseline and prediction models (presented in Figure 3), is somewhat unfair to the prediction model, since we could not reduce the response time of the baseline model without also truncating the ASR result.

The proposed model make turn-completion decisions exclusively in the textual domain, restricted by the latency of the ASR, at the end of user IPUs. In practice, this means that we are more likely to "project" the already spoken words currently being processed by the ASR, as opposed to the actual future activity of the user. This could be mitigated by using a more reactive IPU trigger, increasing the projection events during a user utterance, and to use a longer continuation length, surpassing the latency of the ASR. If so, the system could potentially also start to respond before the user has stopped speaking (i.e., producing overlapping speech).

Another important aspect is that the interactions were all conducted over Zoom which introduces added latencies. This also makes the probability of cut-ins even greater than it would have been in a live setup.

7 Conclusion

In conversation, humans project future turn-completion points in order to achieve faster response times. In this paper, we have investigated whether it is possible to implement this ability in a SDS. The projections are done in the textual domain by generating future dialog continuations with a conversational LM (TurnGPT). We conducted a small study and show, as a proof-of-concept, that this approach is viable. We note that there is room for improvements, such as optimizing

the hyperparameters, train and use a task specific LM, project turn-completion at finer increments, and add prosodic features. However, the idea to use a text-based LM to project turn-completions, as a way to improve the turn-taking abilities of a SDS, is something we believe will be common and useful for the future of conversational systems.

Acknowledgements

This work is supported by the Swedish research council (VR) project "Prediction and Coordination for Conversational AI" (2020-03812) and the Bank of Sweden Tercentenary Foundation (RJ) project "Understanding predictive models of turn-taking in spoken interaction" (P20-0484).

References

- Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020. [A comprehensive evaluation of incremental speech recognition and diarization for conversational AI](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3492–3503, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. [Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There](#), pages 421–432. Springer Singapore, Singapore.
- Timo Baumann and David Schlangen. 2011. [Predicting the micro-timing of user input for an incremental spoken dialogue system that completes a user's ongoing turn](#). In *Proceedings of the SIGDIAL 2011 Conference*, pages 120–129, Portland, Oregon. Association for Computational Linguistics.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. [MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. [Taskmaster-1: Toward a realistic and diverse dialog dataset](#).
- David DeVault, Kenji Sagae, and David Traum. 2009. [Can I finish? learning when to respond to incremental interpretation results in interactive dialogue](#). In *Proceedings of the SIGDIAL 2009 Conference*, pages 11–20, London, UK. Association for Computational Linguistics.
- Erik Ekstedt and Gabriel Skantze. 2020. [TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online. Association for Computational Linguistics.
- Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. [Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines](#). *CoRR*, abs/1907.01669.
- L. Ferrer, E. Shriberg, and A. Stolcke. 2002. [Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody](#). pages 2061–2064. Cited By 56.
- Simon Garrod and Martin J. Pickering. 2015. [The use of content and timing to predict turn transitions](#). *Frontiers in Psychology*, 6:751.
- Martin Johansson, Tatsuro Hori, Gabriel Skantze, Anja Höthker, and Joakim Gustafson. 2016. [Making turn-taking decisions for an active listening robot for memory training](#). In *Proceedings of the International Conference on Social Robotics*, volume 9979 LNAI, pages 940–949.
- Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. [Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues](#). In *2019 International Conference on Multimodal Interaction, ICMI '19*, page 226–234, New York, NY, USA. Association for Computing Machinery.
- Sungjin Lee, Hannes Schulz, Adam Atkinson, Jianfeng Gao, Kaheer Suleman, Layla El Asri, Mahmoud Adada, Minlie Huang, Shikhar Sharma, Wendy Tay, and Xiujun Li. 2019. [Multi-domain task-completion dialog challenge](#). In *Dialog System Technology Challenges 8*.
- Stephen C. Levinson and Francisco Torreira. 2015. [Timing in turn-taking and its implications for processing models of language](#). *Frontiers in Psychology*, 6:731.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Angelika Maier, Julian Hough, and David Schlangen. 2017. [Towards deep end-of-turn prediction for situated spoken dialogue systems](#). In *Proc. Interspeech 2017*, pages 1676–1680.
- Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2013. [A data-driven model for timing feedback in a map task dialogue system](#). In *Proceedings of the SIGDIAL 2013 Conference*, pages 375–383, Metz, France. Association for Computational Linguistics.

- Thilo Michael. 2020. [Retico: An incremental framework for spoken dialogue systems](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52, 1st virtual meeting. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.
- Antoine Raux and Maxine Eskenazi. 2008. [Optimizing endpointing thresholds using dialogue features in a spoken dialogue system](#). In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Columbus, Ohio. Association for Computational Linguistics.
- Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. [Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs](#). In *Proceedings of Interspeech*, Hyderabad, India.
- H Sacks, Emanuel Schegloff, and G Jefferson. 1974. [A simplest systematics for the organization of turn-taking for conversation](#). *Language*, 50:696–735.
- Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. 2002. [Learning decision trees to determine turn-taking by spoken dialogue systems](#). In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*. ISCA.
- David Schlangen. 2006. From reaction to prediction: experiments with computational models of turn-taking. In *Proceedings of Interspeech 2006, Pittsburgh, PA, USA, 2010-2013*.
- David Schlangen and Gabriel Skantze. 2009. [A general, abstract model of incremental dialogue processing](#). In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece. Association for Computational Linguistics.
- Gabriel Skantze. 2017. [Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany. Association for Computational Linguistics.
- Gabriel Skantze. 2021. [Turn-taking in Conversational Systems and Human-Robot Interaction : A Review](#). *Computer Speech & Language*, 67:101178.
- Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. [Universals and cultural variation in turn-taking in conversation](#). *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.
- N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes. 2018. [Turn-taking predictions across languages and genres using an lstm recurrent neural network](#). In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 831–837.
- Nigel G. Ward, Olac Fuentes, and Alejandro Vega. 2010. [Dialog prediction for a general model of turn-taking](#). pages 2662–2665. International Speech Communication Association.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

The dialog in A.1 shows a transcript of a user interaction with the projection agent. We highlight a negative turn-completion projection by [H] (hold) and a positive with [S] (shift). Note that these transcriptions and highlights are shown from the agent’s perspective (i.e., based on what the ASR has reported), and not on what was actually said. Given a continuation length of $M=3$, a [S] indicates that the model estimates that there is less than 3 tokens left in the turn.

The blue text shows words that are recognized by the ASR after the decision to take the turn was made. That is, blue text to the right of a [S] had been spoken by the user at the time of projection, but was not accessible to the model. The orange [...] are used to comment on specific events.

A.1 Example dialog

Agent: *Hi there, in this session I want to talk about one of your past travels. Where did you travel?*

User: I travel to the United States[H] of[H] America[H]

Agent: *What mode of transportation did you use to get to your travel destination?*

User: we went there by[S] airplane

Agent: *Would you say that it was a summer or winter trip?*

User: it was a summer[S] trip

Agent: *Let’s see, did you try any new foods that you could tell me about?*

User: and yeah we tried a lot of foods then like[H] a[H] burritos and[H] vegan pizza[H]

Agent: *What is your best memory from the trip?*

User: the best memory from the Westland and me and my husband went to Central Park and it was a sunny day and we sat there and ate some[H] noodles[S]

Agent: *So, did you visit any tourist attractions?*

User: yeah we[H] did a lot of tourist attractions like the World Trade Center[H]

Agent: *Can you tell me [interrupted by user cough]*

Agent: *Can you tell me anything interesting about your trip?*

User: yes it was very interesting to go by car to the[H] different states we went to with a car[H] from from Denver in Colorado all the way to Los Angeles in[H] California[S]

Agent: *Do you have any recommendations for people traveling there?*

User: yeah I would definitely definitely recommend them to go to New York and[H] the Grand Canyon[H]

Agent: *Let’s see, how was the weather during your trip?*

User: it was very warm[H] it was in the summer[H]

Agent: *So, could you talk about who you traveled with?*

User: yeah I traveled with my husband[H] Rasmus[S]

Agent: *Yeah, would you recommend others to go there?*

User: absolutely[H]

Agent: *Please, tell me more about that. [elaboration request]*

User: well because America is[H] such[H] a big country it’s a lot of things to see[H] something[H] that suits[S] all yeah[H]

Agent: *Let’s [interrupted by ”yeah”]*

Agent: *Let’s see, do you think that such a trip is suitable for families?*

User: yeah[H] I[H] think[H] so[S]

Agent: *Did you do anything that you have not done before?*

User: yeah we were hiking in the mountains near Grand Canyon I hadn’t done[S] that before

Agent: *Do you think you are going to revisit any time soon?*

User: yes[H] I think so[H]

Agent: *So, was it a vacation or work related trip?*

User: it was a[H] vacation[H]

Agent: *For how long were you away?*

User: we were away for three weeks three weeks[H] I[H] think I almost[S] 4 weeks

Agent: *How was the journey home?*

User: it was a long journey home[H] and it was good[S]

Agent: *Thank you for answering my questions. This session is over. Goodbye.*

A.2 Datasets

We trained TurnGPT on seven publicly available dialog datasets namely, Taskmaster (Byrne et al., 2019), MetaLWOZ (Lee et al., 2019), Multiwoz (Budzianowski et al., 2018; Eric et al., 2019), The Coached Conversational Preference Elicitation (Radlinski et al., 2019), Persona (Zhang et al., 2018), DailyDialog (Li et al., 2017), and Empathetic dialogs (Rashkin et al., 2019).