HDIAL: Dataset for benchmarking dialogue systems on linguistic phenomena

Avishkar Gupta and **Tanaya Babtiwale** and **Chirag Jain** and **Krupal Modi**
Haptik, Mumbai

Abstract

Task-oriented dialogue systems in the wild encounter various linguistic phenomena not present in their domain-specific training sets. In the absence of any linguistic benchmarking, the real-world performance of state-of-the-art dialogue formulations does not live up to expectations set by accuracy metrics. In this paper, we present **HDIAL**, a novel dataset to benchmark dialogue agents' performance on these recurrent characteristics of real-world dialogue. The dataset's value proposition is its test set, containing variations of dialogue generated by experts to introduce key phenomena, with annotations for the type of linguistic device used. It enables a fair evaluation of a dialogue system's comprehension ability for semantically equivalent but grammatically divergent conversations. It also provides for analysis of sequential and contextual capabilities and robustness tests. We report baseline numbers using generative and retrieval formulations and highlight the room for improvement. The annotated dataset, models, and code are released for future work.

1 Introduction

The scope of datasets used in the design and testing of task-oriented dialogue systems is focused on dialogue sub-tasks, sidelining considerations of associated linguistic diversity. These popular datasets (Table 1) are not sufficient to measure how a system would perform on real-world user queries it faces once deployed to the wild, concurrently being unsuitable for measuring contextual capability(Mosig et al., 2020).

To facilitate more granular benchmarking, we present HDIAL, a human-annotated dataset for next response prediction (NRP) with test utterances tagged with recurring linguistic phenomena (Table 2). To analyze the sequential and contextual capabilities of a dialogue system formulation we introduce dialogue characteristics such as ellipsis [3.2.3], anaphora [3.2.1], and sub-dialogue [3.2.5] which requires a model to incorporate conversation history in its prediction. HDIAL also contains annotations and dialogue for behavioral testing of NLP models(Ribeiro et al., 2020) to offer a well-rounded estimation of performance in the real world. We highlight the gaps in existing approaches, present our generation and curation process, highlight the utility of each phenomenon and showcase retrieval (Vlasov et al., 2019) and generative (Raffel et al., 2019) baselines to identify headroom for improvement in dialogue systems. The dataset and our code are both freely accessible at (for submission, 2022).

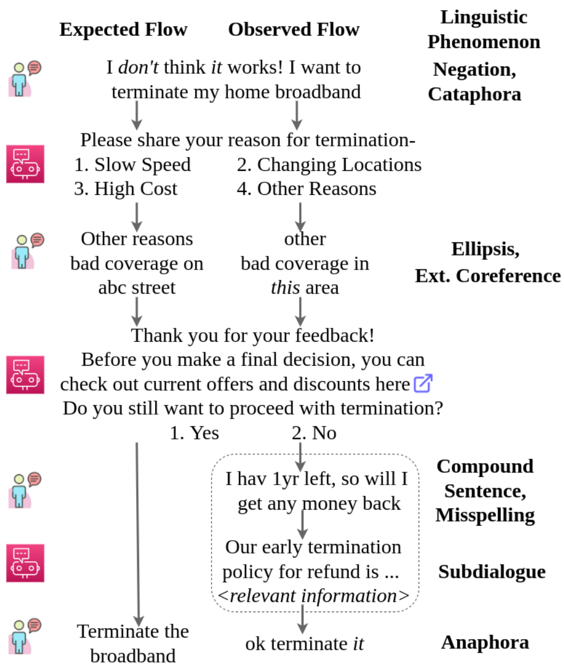

Figure 1: Sample depiction of linguistic phenomenon occurring in dialogue elaborated in 3.2. *Single phenomenon is tagged per data point in HDIAL's test set*

2 Related Work

Linguistic devices such as anaphora and coreference (Sukthanker et al., 2020) are well-studied problems largely evaluated in isolation by datasets

Dataset	Task-Oriented	Collection Methodology	Strengths
MultiWOZ 2.2 (2020)	✓	Wizard-of-Oz (1984) (Human-to-Human) setup	Suitabile for belief tracking and end-to-end dialogue systems with slot filling. Has rich linguistic variety.
MuTual (2020)	✗	Transcription of spoken Human-Human interaction, in a conversational QA setting	Evaluates comprehension and reasoning capabilities of the dialogue system over the open domain.
SGD (2020)	✓	Dialogue simulation via Machines Talking To Machines (2018) setup	Largest available dataset to train scalable dialogue state tracking, intent detection, and slot filling.
CoQA (2018)	✗	Human-to-Human textual interaction in a conversational QA setting	Evaluates support for linguistic phenomena such as coreference, morphology and pragmatic reasoning.
DSTC (2014)	✓	Human speech query to Machine setup	Baseline for spoken dialogue state tracking
HDIAL	✓	Domain expert generated variations of user turns of a human-machine setup	Support for analysis of sequential and contextual capability, behavioral testing, and benchmarking on linguistic phenomena such as anaphora, sub-dialogues, and compound sentences.

Table 1: Review of popular datasets for building dialogue systems

like WinoGrande (Sakaguchi et al., 2019) and in the scope of a related problem such as in CoQA (Reddy et al., 2018).

Dialogue datasets such as MultiWOZ (Budzianowski et al., 2018) and MultiWOZ 2.2 (Zang et al., 2020) largely lack dependence on conversation history (Mosig et al., 2020), hinting to inadequate support for linguistic devices which need to be resolved ‘in-context‘.

Even the more recent works in the area like Schema Guided Dialogue (SGD) (Rastogi et al., 2020) are focused on measuring accuracy on subtasks like intent detection, dialogue state tracking, etc. In all such dialogue datasets, there is no metadata available to mark the linguistic phenomena associated with the dialogues. In contrast, the design of HDIAL approaches NRP with a focus on linguistic diversity, leading to an accuracy measurement that is more representative of performance on real-world queries.

Available datasets are ideal for agents attempting to mimic human-human interactions, as seen in Table 1. Task-oriented systems conversely, for minimizing user effort elicit concise queries from users while providing detailed responses. This dataset caters to the latter and we have used sample data points of a real-world telecom bot that exhibits this behavior. We provide information on privacy valuing collection in appendix A.

Language phenomena	No of Samples
Anaphora	842
Ellipsis	514
Negation	692
Non-anaphoric co-references	512
Compound sentences	594
Correcting messages	512
Entity variations and Misspellings	583
Total	**4249**

Table 2: HDIAL test set overview by linguistic phenomenon - number of datapoints (utterances)

3 Data

3.1 Dataset Collection

For test set creation, we follow a novel collection approach of a human expert modifying user turns in "seed" transcripts that are synthesized using conversation flow graphs and sample utterances supplied by domain experts on a no-code platform used to build conversational agents. Each seed transcript is simply a sequence of dialogues formed by ordering graph nodes containing user and agent dialogue. We use this data to synthesize 100k transcripts in an automated fashion via a simple DFS traversal. The synthesized transcripts cover two major themes which can be further broken down, giving us 11 different conversation flows. We then select 487 transcripts based on a greedy set-covering logic to use as "seeds" for the test set. We provide the complete set of 100k transcripts to be used as training data, however, we sample from this set for our baseline experiments in section 4.

Next, human experts generated 4492 variations of user utterances present in the seed transcripts to introduce linguistic variation while preserving the semantics and sequential flow of the dialogue.

The test set's quantitative metrics largely mimic attributes of the training set as seen in Table 3. We elaborate on our collection and curation methodology in the data statement in Appendix B.

Metrics	Train Transcripts	Seed Transcripts
Total # of transcripts	100000	487
Avg # of words per user message	5.53	3.8
Avg # of words per bot message	60.88	46.7
Avg # of turns per conversation	13	9.65
# of unique bot responses	75	74

Table 3: HDIAL dataset quantitative information

3.2 Annotated Linguistic Phenomenona

3.2.1 Anaphora

Anaphoric references are frequently occurring cohesive devices in dialogue. Their resolution in the context of discourse (Hirst, 1981) and NLU (Sukthanker et al., 2020) is crucial to perform tasks like NRP. To measure this capability, our dataset tags samples of pronominal anaphora and inferrable-evoked pronouns (Eckert and Strube, 1999) as anaphora. Failure to resolve anaphora may trigger an inaccurate response, like termination response for "TV" instead of "broadband" in Fig. 1.

3.2.2 Cataphora & Coreferences

HDIAL consists of modifications for cataphora to discern the rigidity of the model towards syntax, along with external co-references grounded in real-world knowledge, which may perplex the model, possibly resulting in an incorrect response.

3.2.3 Ellipsis

Ellipsis is widespread in dialogue system interactions (Carbonell, 1983). It's caused by the expectation of imperative information being recovered from local context (Phillips and Parker, 2014), and is HDIAL's most significant indicator of the contextual capabilities of a dialogue system. Analysis of our data in Table 3 confirms short, simplistic utterances instead of precise statements expressing intent as displayed in Fig. 1.

3.2.4 Negation

Modern-day attention models are highly sensitive to negation (Mukherjee et al., 2021). To assert correct negation scope resolution (Khandelwal and Attar, 2021), HDIAL consists of perturbations to introduce negation for behavioral testing of models.

3.2.5 Subdialogue

Human conversations often involve diversions in the dialogue state. Sub-dialogues are introduced in 25% of the seeds, and are synthesized by integration of FAQ flows in Termination flows. A deflection to "Refund FAQ" from a termination flow is demonstrated in Fig. 1. A conversation containing sub-dialogue is only brought to completion provided the system has sequential and contextual capabilities, and thus poor handling will affect tests for every other phenomenon.

3.2.6 Corrective messages

In any real-world dialogue system, users provide inaccurate data accidentally and attempt to correct it in the immediate next turn. This may cause conflicting dialogue flow decisions, the accuracy for which is tested by incorporating common deviations as in Table 4.

Seed Transcript	Test sentence
B: Please enter the PowerTely mobile number for which you want to terminate services.	
U: 001-269-228-0762x7191	001-269-228-0762x7190 <NEXT_TURN> *001-269-228-0762x7191 *(corrective message)*
B: Before we proceed with the postpaid mobile termination, let's check if you have any Early Termination Charges...	

Table 4: Sample flow for corrective messages

3.2.7 Compound sentences

A rule-based dialogue policy introduces rigidity in the conversation flow. Table 5 illustrates a scenario in which the second user utterance has been supplied as a part of the first sentence itself, which presents a unique challenge in sequential understanding as the dialogue agent is now expected to skip the immediate response in the flow.

Seed Transcript	Expected change and Test sentence
U: Terminate Value Added Services	I would like to discontinue my Broadband VAS
B: Alright, I see that you are looking to terminate your Value Added Services. Just know you can opt in for them again. May I know *for which one you'd like to terminate*? 1. Mobile Value Added Services 2. Broadband Value Added Services 3. TV Value Added Services. Type a number from *1-3* to make a choice.	`should be skipped`
U: broadband service	`should be skipped`
B: Sure, you can *manage and terminate your value added services for broadband* by clicking on the link given below https://example.com/broadband Is there anything else I can help you with? Reply with *yes* or *no*	`pick this response`

Table 5: When responding to compound sentences, agent response should not lead to repetitive dialogue

3.2.8 Grammatical errors and entity replacements

These utterances were added to gauge behavioral understanding (Ribeiro et al., 2020) of the system, specifically the robustness of the model. This includes modifying the original data point with a misspelling like in Fig. 1, or a variation of the entity, or both.

4 Baseline Experiments and Results

We present baseline formulations using generative and retrieval methods. For our retrieval formulation we use TED policy (Vlasov et al., 2019) via RASA(Bocklisch et al., 2017) and the generative formulation is using T5 (Raffel et al., 2019). We also validate the need for dialogue history and vary the model size.

subcategory	T5-small		T5-base		TED-L1-A4		TED-L6-A8	
	Aug	Base	Aug	Base	Aug	Base	Aug	Base
compound sentences	0.40	0.1	0.53	0.15	0.39	0.11	0.36	0.06
anaphora	0.58	0.83	0.63	0.84	0.58	0.74	0.52	0.68
negation	0.42	0.44	0.50	0.40	0.44	0.58	0.38	0.54
correcting messages	0.74	0.73	0.76	0.83	0.75	0.83	0.71	0.77
grammatical errors & entity variations	0.73	0.71	0.79	0.79	0.73	0.80	0.74	0.75
co-references	0.62	0.58	0.69	0.77	0.47	0.63	0.41	0.58
ellipsis	0.74	0.87	0.87	0.92	0.59	0.79	0.63	0.78
Overall accuracy	**0.59**	**0.61**	**0.67**	**0.66**	**0.56**	**0.64**	**0.53**	**0.59**

Table 6: Test accuracy of our best models without (Base) and with (Aug) augmentation

4.0.1 Formulation

Our generative baseline, T5, models each problem as one of mapping a source sequence to an output sequence. Specifically, we solve the problem of generating the first 100 words in the response given conversation history.

In our retrieval formulation with Dialogue Transformer, the sequences are fed as "stories" to RASA(Bocklisch et al., 2017), where each response is mapped to a corresponding intent class for retrieval.

The generative formulation exploits transfer learning and is pre-trained with a language modeling objective, TED policy baseline is trained from scratch. For a detailed explanation refer to appendix C. Our baselines use a sampled subset of the 100k training data points. The experiments are set up to establish a simple evaluation baseline.

Our baselines, when trained without any augmentation cannot handle compound sentences (formulation induced limitation), evident in the "base" column in Table 6. We, therefore, employ augmentations discussed in appendix D and present numbers on the augmented dataset under the "Aug" column.

4.0.2 Evaluation

It is desired of dialogue systems in the wild to produce an exact response, irrespective of their generative capabilities, adhering to which, we use accuracy as our metric of choice. We truncate response length to the average size of a bot turn (60 words) and expect an exact match.

4.0.3 Result

We list the results of our best models at each parameter count in Table 6. Generative formulation exhibits a gain in performance corresponding to an increase in parameter count. In contrast, retrieval has better accuracy when we have fewer parameters to tune- recall that the latter is randomly initialized and trained on a limited dataset. The higher parameter count of the generative formulation is justified

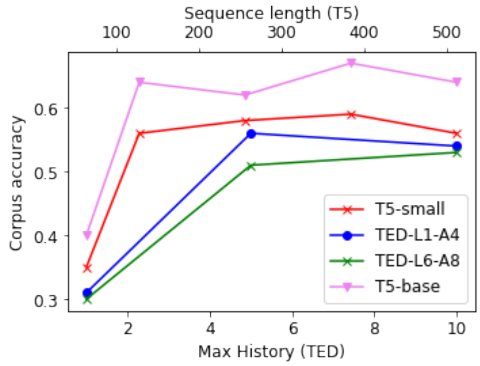

Figure 2: Validating impact of dialogue history

as it solves the problem of sequence generation.

Generative formulation reaches peak accuracy at sequence length 384, roughly in line with the average conversation length of our training set, all results are presented in the sequence length validation in Fig 2.

Performance gains of the TED-L1-A4 retrieval model peak at a lower value for max history compared to TED-L6-A8, attributed to its limited capacity. We discuss some interesting examples of data points that result in an inaccurate prediction by both our models in Table 10.

Looking at the linguistic breakdown in Table 6, we see negation is an issue all models struggle with, indicating an inability for negation scope resolution as a prevalent issue as all our formulations use attention. We see models get perplexed by compound sentences. It suggests poor adaptability to variation in dialog structure, a limitation possibly induced by the formulation itself. Performance on ellipsis and anaphora can be considered indicative of context modeling capabilities.

5 Conclusion

Despite the prevalence of commercially deployed task-oriented dialogue systems, there is a lack of comprehensive benchmarks available to test performance on human-machine conversations in the wild. HDIAL is a large, novel conversational dataset proposed to benchmark the performance of dialogue systems on sequential and contextual capability, behavioral testing, with a focus on recurrent linguistic phenomena in human dialogue. Supported by generative and retrieval baselines, we discuss its importance in the context of next response prediction problem and hope to expose researchers to unique problems faced in real-world multi-turn conversations.

References

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Jaime G Carbonell. 1983. Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 164–168.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*.

Miriam Eckert and Michael Strube. 1999. Resolving discourse deictic anaphora in dialogues. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–44.

Anonymous for submission. 2022. Hdial dataset and baseline code. Available online at: https://hdial1.s3.ap-south-1.amazonaws.com/hdial-submission.tar.gz, last modified on 16.05.2022.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Graerne Hirst. 1981. Discourse-oriented anaphora resolution in natural language understanding: A review. *American journal of computational linguistics*, 7(2):85–98.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Aditya Khandelwal and Vahida Attar. 2021. Orthogonal attention: A cloze-style approach to negation scope resolution.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Johannes EM Mosig, Vladimir Vlasov, and Alan Nichol. 2020. Where is the context?–a critique of recent dialogue datasets. *arXiv preprint arXiv:2004.10473*.

Partha Mukherjee, Youakim Badr, Shreyesh Doppalapudi, Satish M Srinivasan, Raghvinder S Sangwan, and Rahul Sharma. 2021. Effect of negation in sentences on sentiment analysis and polarity detection. *Procedia Computer Science*, 185:370–379.

Colin Phillips and Dan Parker. 2014. The psycholinguistics of ellipsis. *Lingua*, 151:78–95.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Vladimir Vlasov, Johannes EM Mosig, and Alan Nichol. 2019. Dialogue transformers. *arXiv preprint arXiv:1910.00486*.

Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.

A Ethical Considerations

Our contributions are completely compliant with ACM's Code of Ethics[1] to the best of our knowledge. The dataset has been curated with various checks in place to control quality and to avoid preexisting biases in manner of gender, culture, or economic standing. We ensure that publishing this dataset results in no direct consequences that may be unjust or negative in nature, and has been proposed in order to minimize possible risks of lack of linguistic understanding that are prevalent with recent neural methods. The work, by itself, does not entail to any potential risks, other than contributing to the imbalance of research work in English language as opposed to other languages.

The work strives to provide a sterner benchmark for assessing models on the properties of human conversation (Jurafsky, 2000), context retention, and linguistic capability. This not only provides novelty, but also a dedication to improve on the quality of the present state of research in dialogue systems, which would not have been possible without compliance to the proposed code of professional conduct. Furthermore, we provide a detailed data statement (Bender and Friedman, 2018) in Appendix B to provide transparency and accountability (Gebru et al., 2021) for the practices adopted, and to supply additional insights into the proposed dataset.

We respect the privacy of our client(s) and consumers alike and thus do not retain any personally identifiable information(PII) belonging to them. Even though the dataset has been inspired by a real-world use-case, we have maintained complete confidentiality and anonymity of the client(s) and have been granted official consent to present this contribution from all parties that own the concerning data.

B Data Statement

B.1 Dataset Curation

Transcripts were generated using methodology discussed in Section 3.1. These were further processed using scripts to ascertain anonymity and remove any identifying information. Bot and user responses in the seed transcripts were manually annotated by a small team comprising of data specialists and machine learning researchers with no access to consumer PII. Any PII present in the dataset is generated using appropriate tools[2,3]. A stringent set of annotation guidelines were maintained, which have been made available in the dataset repository, to preserve uniformity of the data, and to avoid inter-annotator conflict. The dataset is inspired by the dialogues carefully curated by experienced conversation designers, with the help of domain experts, in the year 2021. The final dataset was further reviewed by multiple competent stakeholders to ensure removal of identifying information and gendered references promoting societal biases, and ensuring correctness and consistency of the proposed data.

All annotators, reviewers, and conversation designers have been recruited as employees in the organization, and are thus compensated generously according to market standards with accommodating work conditions and access to multiple resources to ensure their well-being.

B.2 Limitations

The scope of the dataset is restricted to a small number of action use cases in the Telecommunications domain. The proposed problem statement formulation, that the dataset is intended for, does not cover a few common functionalities of dialogue systems such as entity extraction. The intended use of the dataset is primarily to evaluate system capabilities in a real world setting where low training times are as much of a consideration as accuracy peaks.

B.3 Speech Situation and Text Characteristics

The scope of the dataset is limited to the Telecommunications domain and has conversations strictly in the English language, following the Latin script.

[1] https://www.acm.org/code-of-ethics

[2] https://github.com/joke2k/faker

[3] https://exrex.readthedocs.io/

The responses are majorly of language variety en-IN, but few cases may also be en-SG, as the foundational data is intended for audience in the region of maritime Southeast Asia. The dataset is intended to mimic spontaneous, asynchronous conversations by written means, wherein the users are aware that they are conversing with a dialogue agent.

B.4 Demographic Information

Since conversations were not picked up from production, and initial data generation was carried out by skilled conversation designers, we may say that the intended audience, or speakers, were not involved in the development of this dataset. The conversation designers, data annotators, and reviewers were in the age range of 20-40 years, and reside majorly in the Indian subcontinent. While all the annotators identify as men, the team of reviewers consisted of men and women in exactly equal proportions.

B.5 Distribution and Maintenance

The dataset will be open-sourced in interest of advancement of science, but will remain an intellectual property of the organization. All original and revised versions of the dataset will be hosted and maintained by the organization. Contributors are encouraged to suggest enhancements to the data, which will be added to the dataset after careful review and validation, under the sole discretion of the organization.

C Experimental Setup, Hyperparameter Choices and Training Procedures

C.1 Common Setup

For all our experiments, we use NVIDIA V100 and T4 GPUs for accelerated training with CUDA 11.2 and cuDNN 8.1. Each experiment uses only a single GPU for the entire run.

A single experiment (one model run on full dataset) depepnding on model size on average took around 5 hours on a T4 and 3 hours on a V100. This excludes our T5-base experiments which each take anywhere from 6-10 GPU hours each due to batch size constraints on a single Nvidia V100. We estimate that our total usage falls in the range of roughly 150 GPU hours of T4 and 50+ GPU hours of V100 with most other work being in dataset collation, pre-processing and evaluation which happened on CPU. We hope our baselines and the scientific artifacts we provide would help reduce this cost for those looking to improve on our work.

C.2 T5

For T5, we start with pretrained PyTorch checkpoints `t5-small`[4] and `t5-base`[5] from HuggingFace Hub[6] and use the Trainer interface implemented in transformers library (Wolf et al., 2020). We mark each user and bot turn as such with `"» User:"` or `"» Bot: "` markers akin to DialoGPT (Zhang et al., 2019). We compress initial bot turns using POS tagging to work around sequence length limitations.

We set the training hyperparameters as follows:

Hyperparameter	Value	
	T5-small	T5-base
Learning Rate	0.0001	0.0003
Epochs	3	3
Max sequence length (output)	100	100
instantaneous batch size (train)	25	8
Gradient accumulation	3	9
Warmup steps	200	200

Table 7: Hypermarameter values for T5 train setup. Batch sizes were kept such that instantaneous fits on 16GB VRAM + gradient accumulation done to round up batch size to $\tilde{7}5$.Numbers above are for input sequence length 512

For both models we run experiments at five different input max sequence lengths - 64, 128, 256, 384 and 512 - leading to a total of 10 experiments. All models are optimized with AdamW (Loshchilov and Hutter, 2019).

We tuned the batch size and learning rate for T5 base and found between the two Huggingface recommendations of $1e-4$ and $3e-4$, $3e-4$ works better with T5-base while $1e-4$ works better with T5-small.

C.3 TED

For TED, we use Rasa v3.0.4. Starting with default settings for all components (obtained with `rasa init`), across all experiments, we remove `EntitySynonymMapper` and `ResponseSelector` from the pipeline and remove all policies except `TEDPolicy`. We then customize configuration settings for `TEDPolicy` as follows to obtain models with two different sizes:

[4]https://huggingface.co/t5-small/tree/a3f81e09b2ac3ff1d0bd6aafde94183f45efec7c

[5]https://huggingface.co/t5-base/tree/7590d914f24f72ee29ffbfb94cf8ba4de8a66189

[6]https://huggingface.co/

Key	Subkey	Value	
		TED-L1-A4	TED-L6-A8
number_of_attention_heads		4	8
number_of_transformer_layers	action_text	1	6
	dialogue	1	6
	label_action_text	1	6
	text	1	6

With these settings TED-L1-A4 has approximately 2.17M parameters (1.3M trainable) and TED-L6-A8 has approximately 10M parameters (5.26M trainable). For each model size, we consider three different `max_history` settings - 1, 5 and 10 leading to six total configurations. In addition, all six TED experiments use the following settings.

Key	Value
batch_size	32
constrain_similarities	true
drop_rate	0.1
drop_rate_attention	0.1
drop_rate_label	0.1
e2e_confidence_threshold	0.0
entity_recognition	false
epochs	3
random_seed	2022
renormalize_confidences	true

Any additional options not listed here are kept with the defaults that Rasa v3.0.4 ships with. Please refer https://rasa.com/docs/rasa/policies/#ted-policy for detailed explanations of these options. All models are optimized with Adam (Kingma and Ba, 2015).

D Additional Dialogue Preprocessing

D.1 Data augmentations to support compound sentences

Both our generative and retrieval baseline when trained without any further augmentations to the data, fail to skip repetitive bot turns to shorten conversation length when compound sentences are encountered, a property that directly follows from our formulations.

- We combine all user messages in a sample required to generate a response and treat this simulation of a “compound message” as a single user turn that generates it.
- To simulate compound sentences that occur mid-conversation, we exclude one bot turn at random from the aforementioned user pairs and combine the user utterances before and after said message.

D.2 Actions With External Dependencies

In our dataset, we have five actions whose actual utterance depends on external factors that cannot be inferred from the conversation context. For e.g. a user may or may not have to pay extra charges for termination of their postpaid service depending on their agreement with the telecom provider. As such a machine learning model by itself can only predict that the system has to check external conditions to resolve the actual utterance. Such actions are explained in 8 To handle such actions we perform additional preprocessing which is explained below. We provide an example in 9

D.2.1 T5

For T5, We replace the bot utterances of these actions with special placeholders only in the output action in the training data. We do not perform these replacements in the input context fed to the T5 encoder. We do this to allow the encoder to exploit any useful information from the raw bot utterance in the context. As such during prediction, for these actions, T5 decoder is expected to generate these placeholders instead of the actual bot utterances themselves. We replace utterances for the first two actions in 8 with just a single bot turn `SET_ETC_FLOW` and the other three with just a single bot turn `VERIFY_NRIC__DOB`. We also perform similar mapping on the expected action during the evaluation process.

D.2.2 TED

For TED, even though we are using end to end formulation, we replace the bot utterances of these actions with special placeholders everywhere in the Rasa training stories. As such during prediction, for these actions, the model is then expected to generate these placeholders instead of the bot utterances themselves i.e. we replace utterances for the first two actions in 8 with just a single bot turn `SET_ETC_FLOW` and the other three with just a single bot turn `VERIFY_NRIC__DOB`. We also perform similar mapping on the input context and expected action during the evaluation process.

D.3 Bot Turns Summarization and Context Truncation in Input Context for T5

As mentioned in 4.0.1 we are feeding the complete dialogue context delimited by turn markers as a single input to T5. But since the T5 encoder has a max input sequence length defined, we compress and truncate the input context while trying to maximize the information we can provide. To do this first if the total number of turns in the dialogue transcript is N and if N greater than 10, we compress up to N - 10 bot turns beginning from the start of the transcript. We do not compress any user turns. To

Scenario	External condition	Bot utterances
action_id: `test bot dialogue policy::914::SET_ETC_FLOW_V_ae5822ad_b7ee_4cf1_a905_e682c99dc0470+61cc9fa3f163a202109cbec3`		
User wants to terminate their service	User has no termination charges	B: Before we proceed with the postpaid mobile termination, let's check if you have any Early Termination Charges. I can see that as of now there aren't any Early Termination Charges for 9999999999 B: Please note that once your number [[phone]] is terminated — You will not be able to receive calls and SMS's on this number You will not be able to retrieve or reissue this number again You cannot reactivate this number once terminated Kindly make the final confirmation: 1.* Terminate 2.* Don't terminate
action_id: `test bot dialogue policy::914::SET_ETC_FLOW_V_ae5822ad_b7ee_4cf1_a905_e682c99dc0471+61cc9fa3f163a202109cbec3`		
User wants to terminate their service	User has to pay termination charges	B: Before we proceed with the postpaid mobile termination, let's check if you have any Early Termination Charges. I checked and found that the early termination charge for 9999999999 would be $50 as of today. B: Please note that once your number [[phone]] is terminated — You will not be able to receive calls and SMS's on this number You will not be able to retrieve or reissue this number again You cannot reactivate this number once terminated Kindly make the final confirmation: 1.* Terminate 2.* Don't terminate
action_id: `test bot dialogue policy::255::VERIFY_NRIC__DOB_V_49bcd33a_e86f_4d3a_98e8_1c3ba75fec560`		
User is experiencing internet issues	There is an outage	B: Looks like there is an outage in your area right now. We are working on fixing this as soon as possible. We post update on our Facebook page after major outages are fixed. Tap on the button below and one of our teammates will be able to assist you right away
action_id: `test bot dialogue policy::255::VERIFY_NRIC__DOB_V_49bcd33a_e86f_4d3a_98e8_1c3ba75fec560+61cc9f8df163a202109cbd5e`		
User is experiencing internet issues	The problem is likely on user's end	B: We do not detect any problems in your internet connection. The problem is likely to be with the router. Let's confirm that by removing the ethernet cable from the router and plugging it directly into your PC/laptop. B: Let me know if your internet is working once this is done. Are you able to surf the internet now?
action_id: `test bot dialogue policy::255::VERIFY_NRIC__DOB_V_49bcd33a_e86f_4d3a_98e8_1c3ba75fec560+61cc9f92f163a202109cbdd2`		
User is experiencing internet issues	Telecom provider takes an action on their end	B: We've just refreshed the signal to your TV service. Please check if you are able to watch your channels now. B: Did that resolve your issue?

Table 8: Actions depending on external conditions

compress a bot turn we obtain the POS tags for the words in the turn using spaCy (Honnibal et al., 2020), drop any punctuations and stop words and then only keep the words with one of the following tags: `ADP`, `PROPN`, `ADJ`, `NOUN`, `VERB`. If the compressed input context is still larger than the max input sequence length we start dropping turns from the beginning from the start of the transcript one by one until it fits within the max input sequence length. We do this transformation both during training and evaluation. We provide an example in 9

E Interesting mis-classifications/generations

In 10 we present some examples that all models got wrong

Example Transcript with 14 turns

>>User: I want to cancel the service as I have moved to Australia
>>Bot: It's sad to know that you are looking to cancel your PowerTely services Which *service do you wish to terminate*? *1.* Postpaid mobile *2.* Broadband *3.* PowerTely TV *4.* PowerTely TV+ *5.* Both TV & Broadband *6.* Digital Voice *7.* Value Added Services Type a number between *1-7* to proceed further.
>>User: postpaidusing sim only
>>Bot: Got it! I'd like to know the reason for Mobile termination so that I can serve you better in the future. *Would you want to share your concerns?* Reply with a *Yes* or *No* to proceed further.
>>User: yah
>>Bot: Thank you for choosing to share your feedback. I am all ears. *Select an option* which best describes your reason to terminate the Postpaid Mobile service. *1*. Network coverage Issue *2*. Prefer other service provider *3*. Moving overseas *4*. Other reasons Type a number from *1-4* to select an option
>>User: one
>>Bot: Thank you for your essential feedback. We will continue to improve our services and hope that you'll opt for PowerTely as a provider in future. Please enter the PowerTely mobile number for which you want to terminate services.
>>User: 638-613-6766
>>Bot: Before we proceed with the postpaid mobile termination, let's check if you have any Early Termination Charges. I checked and found that the early termination charge for 9999999999 would be $50 as of today.
>>Bot: Please note that once your number [[phone]] is terminated —
You will not be able to receive calls and SMS's on this number
You will not be able to retrieve or reissue this number again
You cannot reactivate this number once terminated

Kindly make the final confirmation:
1.* Terminate
2.* Don't terminate
>>User: cancel
>>Bot: Alright! Please wait, while I connect you with one of my teammates for further assistance.
>>Bot: Transfer to agent

For T5, First 4 bot turns compressed and external dependent action with placeholder

context	**action**
>>User: I want to cancel the service as I have moved to Australia >>Bot:sad know looking cancel powertely services service wish terminate postpaid mobile broadband tv tv+ tv broadband digital voice value added services type number proceed >>User: postpaidusing sim only >>Bot:got like know reason mobile termination serve future want share concerns reply proceed >>User: yah >>Bot:thank choosing share feedback ears select option best describes reason terminate postpaid mobile service network coverage issue prefer service provider moving reasons type number select option >>User: one >>Bot:thank essential feedback continue improve services hope opt provider future enter mobile number want terminate services >>User: 638-613-6766 	>>Bot: SET_ETC_FLOW

For T5, First 4 bot turns compressed. External dependent action appears in the context but it is retained in utterance form

context	**action**
>>User: I want to cancel the service as I have moved to Australia >>Bot:sad know looking cancel powertely services service wish terminate postpaid mobile broadband tv tv+ tv broadband digital voice value added services type number proceed >>User: postpaidusing sim only >>Bot:got like know reason mobile termination serve future want share concerns reply proceed >>User: yah >>Bot:thank choosing share feedback ears select option best describes reason terminate postpaid mobile service network coverage issue prefer service provider moving reasons type number select option >>User: one >>Bot:thank essential feedback continue improve services hope opt provider future enter mobile number want terminate services >>User: 638-613-6766 >>Bot: Before we proceed with the postpaid mobile termination, let's check if you have any Early Termination Charges. I checked and found that the early termination charge for 9999999999 would be $50 as of today. >>Bot: Please note that once your number [[phone]] is terminated — You will not be able to receive calls and SMS's on this number You will not be able to retrieve or reissue this number again You cannot reactivate this number once terminated Kindly make the final confirmation: *1.* Terminate *2.* Don't terminate >>User: cancel 	>>Bot: Alright! Please wait, while I connect you with one of my teammates for further assistance. >>Bot: Transfer to agent

For Rasa, story steps with external dependent action replaced with placeholder

```
- user: I want to cancel the service as I have moved to Australia
- bot: It's sad to know that you are looking to cancel your PowerTely services Which
service do you wish to terminate*? *1.* Postpaid mobile *2.* Broadband *3.*
PowerTely TV *4.* PowerTely TV+ *5.* Both TV & Broadband *6.* Digital Voice
7.* Value Added Services Type a number between *1-7* to proceed further.
- user: postpaidusing sim only
- bot: Got it! I'd like to know the reason for Mobile termination so that I can
serve you better in the future. *Would you want to share your concerns?* Reply
with a *Yes* or *No* to proceed further.
- user: yah
- bot: Thank you for choosing to share your feedback. I am all ears. *Select an
option* which best describes your reason to terminate the Postpaid Mobile service.
1*. Network coverage Issue *2*. Prefer other service provider *3*. Moving overseas
4*. Other reasons Type a number from *1-4* to select an option.
- user: one
- bot: "Thank you for your essential feedback. We will continue to improve our services\
\and hope that you'll opt for PowerTely as a provider in future. \nPlease enter\
\the PowerTely mobile number for which you want to terminate services."
- user: 638-613-6766
- bot: SET_ETC_FLOW
- user: cancel
- bot: Alright! Please wait, while I connect you with one of my teammates for further
assistance.
- bot: Transfer to agent
```

Table 9: Example of transformations for T5 and TED

Turn of conversation	Compound sentence	Anaphora	Negation	Ellipsis
Turn 1	U: I would want to go ahead with my TV+ cancellation coz I found some better service provider	U: i would like to know how much would it cost if i terminate my plan with 1 year left	U: Can I ask if there is an issue with the PowerTely TV singapore wide right now?	U: My broadband is not working. What do I do?
Turn 2		B: It's sad to know that you are looking to cancel your PowerTely services. Which *service do you wish to terminate*? *1.* Postpaid mobile *2.* Broadband *3.* <snip>	B: If you are experiencing issues related your Set-Top Box, PowerTely TV+ or PowerTely Streaming Player, you may begin with our self troubleshooting steps <snip>	B: If experiencing Internet Connection issues, intermittent disconnection, slow connection or wireless issues, you may begin with our self troubleshooting steps or choose to visit more troubleshooting guides for other issues: I0 - Self Troubleshoot I1 - More Guides
Turn 3		U: I'll go with the first one	U: Self troubleshooting	U: Troubleshooting
Turn 4			B: Just a quick reminder! In order to help you troubleshoot the issue, we need to collect and store a few personal details. This will be used by PowerTely to assist you with your query. <snip>	B: Just a quick reminder! In order to help you troubleshoot the issue, we need to collect and store a few personal details. This will be used by PowerTely to assist you with your query. Here's the privacy policy: https://www.example.com/legal-notices-and-terms/personal-data-protection.html Will this be fine with you? QR0 - Yes QR1 - No QR2 - View Menu
Turn 5			U: I don't see any concerns in you saving my details	U: proceed
Turn 6				B: Please enter your NRIC number/FIN below
Turn 7				U: t5640640X
Turn 8				B: And what's your date of birth? The DD/MM/YYYY (e.g. 31/10/2019) format works best!
Turn 9				U: 13/01/1961
Turn 10				B: We do not detect any problems in your internet connection. The problem is likely to be with the router.\r\n\r\nLet's confirm that by removing the ethernet cable from the router and plugging it directly into your PC/laptop. Let me know if your internet is working once this is done.\r\n\r\nAre you able to surf the internet now?
Turn 11				U: Unable to access
Expected Response	B: I see! Please know that your feedback is highly appreciated. Before you make your final decision regarding TV+ termination don't forget to check out all our exciting offers here: <snip>	B: Got it! I'd like to know the reason for Mobile termination so that I can serve you better in the future. *Would you want to share your concerns?* Reply with a *Yes* or *No* to proceed further.	B: Please enter your NRIC number/FIN below	B: Let's try this next: 1. Turn off your Optical Network Terminal (this is the black or white coloured NucleusConnect device shown below). 2. Wait for a minute before turning it back on again. 3. Check to see if you have a connection after that. \nAre you able to surf the internet now? Pick one from the below to continue QR0 - Yes QR1 - No
Predicted Response (T5)	B: I see! As we continue to improve our services, we hope that you'll opt for PowerTely as a provider in future. Kindly follow the instructions in the following messages to terminate your services. *Here's how you can terminate your TV+ plans: <snip>	B: Thank you for your essential feedback. We will continue to improve our services and hope that you'll opt for PowerTely as a provider in future. Please enter the PowerTely mobile number for which you want to terminate services.	B: No problem! Tap on the button below and one of our teammates will be able to assist you right away I0 - Chat with us	B: Sorry to hear that. We will arrange for technicians to check your issue in person.
Predicted Response (TED)	B: I see! Please know that your feedback is highly appreciated. Before you make your final decision regarding TV and Broadband termination <snip>	B: Thank you for your essential feedback. We will continue to improve our services and hope that you'll opt for PowerTely as a provider in future. \n *Here's how you can terminate your Broadband service <snip>	B: No problem! Tap on the button below and one of our teammates will be able to assist you right away I0 - Chat with us	B: Since you're able to surf now, we can confirm that the router is faulty. If your router is within warranty, please approach the router manufacturer for further support.\nIf you have more queries, tap on the button below to see what else I can do for you
Comments	T5 generates a turn from the correct flow however skips to the turn after and doesn't show offers. TED gets a response from a different flow.	TED generates a response from the wrong flow, T5 picks up the correct flow but yet again skips a turn and assumes the user has shared their concerns.	Neither model can handle negation and goes to the opposite response that would've been given had the user refused to share details.	TED gives the opposite response while T5 gives a valid but incorrect response.

Table 10: Interesting mis-classifications/generations and discussion