BOTSTALK: Machine-sourced Framework for Automatic Curation of Large-scale Multi-skill Dialogue Datasets

Minju Kim[1*] **Chaehyeong Kim**[1*] **Yongho Song**[1*] **Seung-won Hwang**[2] **Jinyoung Yeo**[1†]
[1]Department of Artificial Intelligence, Yonsei University
[2]Department of Computer Science and Engineering, Seoul National University
{minnju,cheris8,kopf_yhs,jinyeo}@yonsei.ac.kr seungwonh@snu.ac.kr

Abstract

To build open-domain chatbots that are able to use diverse communicative skills, we propose a novel framework BOTSTALK, where multiple agents grounded to the specific target skills participate in a conversation to automatically annotate multi-skill dialogues. We further present Blended Skill BotsTalk (BS𝔹T), a large-scale multi-skill dialogue dataset comprising 300K conversations. Through extensive experiments, we demonstrate that our dataset can be effective for multi-skill dialogue systems which require an understanding of skill blending as well as skill grounding. Our code and data are available at https://github.com/convei-lab/BotsTalk.

1 Introduction

A considerable progress has been made towards open-domain chatbots with different desirable qualities in conversations. Each of these models is capable of being specialized in one communicative skill, *i.e.*, skill grounding. A number of distinct large-scale datasets targeting a specific conversational skill have recently become available. ConvAI2 (Dinan et al., 2020b) is provided for research work that aims to endow chatbots with personas (Majumder et al., 2020a; Kim et al., 2020b), enabling chatbots to talk about themselves. Wizard of Wikipedia (WoW) (Dinan et al., 2019) is a popular option for recent studies (Lian et al., 2019; Zhao et al., 2020; Kim et al., 2020a) that focus on knowledgeable conversational agents discussing topics in depth. Empathetic Dialogues (ED) (Rashkin et al., 2019) is also commonly used to embody empathy in dialogue systems (Santhanam and Shaikh, 2019; Majumder et al., 2020b). Most of such skill-grounded datasets are designed to improve a single skill, and thus effective when models are asked to demonstrate the targeted conversational skill.

Benefiting from the advances of these conversational agents, recent research focuses on another aspect of open-domain chatbots: the ability to blend various conversational skills into one cohesive flow in a seamless manner, *i.e.*, skill blending. A good open-domain chatbot should be able to weave multiple behaviors and skills in a single conversation, so that it enables to deal with different users and situations appropriately (Shuster et al., 2020; Roller et al., 2021). Towards this goal, there is a need to construct a multi-skill dialogue dataset, which consists of multi-turn dialogues that exhibit multiple skills. While Smith et al. (2020) propose a crowdsourced dataset Blended Skill Talk (BST) of 5K conversations as a reliable benchmark for measuring dialogue systems' ability at the blended objective, it is not sufficient to build a multi-skill chatbot due to its limited scale. Scaling up crowdsourcing is not feasible, as it requires labor intensive manual annotation and verification. Instead, automatic curation shows promising results on large-scale dialogue generation (Mohapatra et al., 2021).

In this paper, we aim to generate a large-scale multi-skill dialogue dataset without additional costs or human efforts. To this end, we introduce an automatic data curation approach named **BOTSTALK**, where multiple dialogue agents grounded to individual skills engage in the conversation to blend all skills together. Based on this framework, we create **Blended Skill BotsTalk (BS𝔹T)**, a large-scale multi-skill dialogue dataset of 300K conversations blended and grounded with a number of skills derived from ConvAI2, WoW, and ED. Our experiments demonstrate that by using our dataset dialogue models successfully yield large performance gains in skill blending while maintaining competitive performance in skill grounding. Furthermore, we validate the quality of BS𝔹T dataset by human evaluation, showing our machine-sourced conversations are consistently preferred over crowdsourced ones from BST by human judges across all metrics.

*Equal contribution
†Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5149–5170
December 7-11, 2022

Dataset	Dialogue episode
ConvAI2	**Skill context** for speaker A: I like to ski; I hate Mexican food; I like to eat cheetos; ... **Skill context** for speaker B: I am an artist; I have four children; I enjoy walking for exercise; ... **Dialogue context** A: How old are your children? B: I have four that range in age from 10 to 21. You?
Wizard of Wikipedia	**Skill context** for speaker A: Armadillo **Skill context** for speaker B: Armadillo are ... "armadillo" means "little armoured one" in ... **Dialogue context** A: I don't think I've ever seen an armadillo in real life! B: I've seen them at the zoo. Armadillo means little armored one in Spanish.
Empathetic Dialogues	**Skill context** for speaker A: My brother jump scared me while I was out playing; Terrified **Skill context** for speaker B: None **Dialogue context** A: Just got scared to death. B: Oh no. What happened?

Table 1: Example dialogues of three single-skill datasets: ConvAI2 provides each speaker persona sentences as skill context; Wizard of Wikipedia provides a topic and knowledge resources as skill context; Empathetic Dialogues provides a situation description and emotion as skill context. We only provide two turns of dialogue contexts due to the limit on the paper length.

2 Related Work

2.1 Skill-grounded Dialogue Datasets

Past research in open-domain chatbots has made solid strides towards dialogue systems with desirable general qualities in a conversation. Generating responses grounded to specific conversational skill has been explored in different axes, as shown in Table 1 (see also Appendix B for details). Dinan et al. (2020b) introduce ConvAI2 dataset which consists of more than 140K utterances of crowdsourced conversations to make chit-chat models more engaging and personalized by conditioning the models on profile information. Wizard of Wikipedia (Dinan et al., 2019) task aims to explore conversation informed by expert knowledge from Wikipedia and provides about 194K utterances of conversations on about 1,250 topics. Rashkin et al. (2019) construct a dataset, Empathetic Dialogues, comprising 50K utterances of crowdworker conversations grounded in an emotional situation for a model to converse with empathy. However, it remains unclear whether models optimized for performance along specific conversational skill can retain the learned skill while blending it with other skills.

Hence, Smith et al. (2020) aim to build a conversational agent who seamlessly blends being personable, knowledgeable, and empathetic. In order to gauge how successful a model is at this blended objective, Smith et al. (2020) collect a new multi-skill dialogue dataset of about 5K conversations, Blended Skill Talk, via crowdsourcing. While this work provides a testbed for future studies, the scale of data could hinder further progress, since training multi-skill chatbots generally requires a large-scale dataset consisting of conversations that involve multiple skills (Shah et al., 2018).

2.2 Automatic Dialogue Data Annotation

Research in dialogue systems has been consistently supported by the development of new dialogue datasets (Williams et al., 2014; Mrkšić et al., 2017). One popular approach is to collect and annotate dialogues via crowdsourcing (Zhang et al., 2018; Smith et al., 2020). However, generating multi-turn dialogues in this manner requires expensive and exhausting human efforts (Shah et al., 2018; Sun et al., 2021; Mohapatra et al., 2021).

Therefore, recent study seeks to facilitate open-domain chatbot development with new datasets automatically constructed by using existing datasets. For instance, Lee et al. (2021) create a 45K multi-modal dialogue dataset by replacing parts of source dialogues from existing text-only dialogue datasets with their semantically relevant images. Sun et al. (2021) propose a Human $\leftrightarrow$ AI collaborative data collection approach for generating diverse chit-chat response to augment task-oriented dialogues and present new chit-chat based annotations to 23.8K dialogues from two popular task-oriented datasets. Kim et al. (2021b) and Vidgen et al. (2021) present a model-based dialogue collection framework and a human-and-model-in-the-loop process for generating datasets respectively.

3 Problem Formulation

In this section, we formulate the problem of multi-skill dialogue annotation and desirable characteristics for the dialogue dataset as a training resource.

3.1 Multi-skill Dialogue Annotation

Our goal is to collect a new large-scale multi-skill dialogue dataset, which seamlessly blends various skills over the course of a multi-turn conversation. Here, inspired by Smith et al. (2020), the inputs of this task are single-skill datasets, which are separately collected on a variety of skills. Let $\mathbb{M}$ be the set of M skill types, *e.g.*, $\mathbb{M} = \{\mathrm{P}, \mathrm{K}, \mathrm{E}\}$, where P, K, E denote personality, knowledge, and empathy derived from ConvAI2, WoW, and ED, respectively. Formally, we refer to $\mathcal{D}_m$ as a dialogue dataset with N_m dialogue episodes for skill $m \in \mathbb{M}$

$$\mathcal{D}_m = \{(stx_{i,m}, dtx_{i,t})\}_{i=1}^{N_m} \tag{1}$$

where $stx_{i,m}$ is a skill-relevant description (*i.e.*, skill context) for skill m and $dtx_{i,t}$ is t dialogue turns (*i.e.*, dialogue context) derived from the skill context, as shown in Table 1. Based on the input datasets $\mathcal{D}_1, ..., \mathcal{D}_M$, we aim to obtain a new dialogue dataset $\tilde{\mathcal{D}}$ for M skills as an output. Formally,

$$\tilde{\mathcal{D}} = \{(\tilde{stx}_i, dtx_{i,t})\}_{i=1}^{\infty} \tag{2}$$

where $\tilde{stx}_i$ is a set of skill contexts for $\mathbb{M}$ and $dtx_{i,t}$ is the dialogue context derived from the multiple skills. We will omit the index i when dealing with a single dialogue episode.

3.2 Desirable Characteristics of Multi-skill Dialogue Datasets

By the above annotation, we aim to build a multi-skill chatbot that uses all target skills appropriately in a conversation. For that, we lay out two criteria that a multi-skill dialogue dataset should meet as a training resource, namely **skill blending** and **skill grounding**. Skill blending indicates that a multi-skill dialogue dataset should enable dialogue models to exhibit different dialogue skills in a conversation (Smith et al., 2020), while skill grounding emphasizes that dialogue models should learn to maintain each dialogue skill when appropriate (Shazeer et al., 2017). Generally, they have a trade-off relationship as it is insufficient to represent both skill blending and grounding in a conversation of finite length (Madotto et al., 2021). Nevertheless, we note that skill blending and grounding are not contradictory, as some skill-grounded utterances leave room for natural shift between skills. Given an utterance "*I like sneakers because it is comfortable.*" which represents skill type P, it seems reasonable to annotate an utterance with skill type K "*It is because sneakers were primarily designed for sports.*" for next dialogue turn. This example further implies that different skills can be blended naturally so that the chatbots learn to provide reasonable responses in a multi-skill dialogue (Roller et al., 2020).

4 BotsTalk Framework

We now present BotsTalk, a novel framework that automatically annotates multi-skill dialogues based on multiple single-skill dialogue datasets. The focus of our framework is to mimic a natural conversation by featuring both skill blending and grounding within a dialogue episode. Figure 1 illustrates three main phases of the framework. Implementation details are provided in Appendix C.

4.1 Participants in BotsTalk

In our framework, multiple participants engage in a conversation to iteratively generate desirable multi-skill dialogues.

Skill Agents The first participants are multiple single-skill agents who annotate the appropriate skill-grounded utterances to the dialogue. Formally, based on $\mathcal{D}_m$ for skill m, when given skill context stx_m, dialogue context dtx_t, and response space $\mathbb{U}$, a skill agent has dialogue models $f : (stx_m, dtx_t) \mapsto \mathbb{U}$ which return a response

$$res_{m,t} = f(stx_m, dtx_t; \theta^m) \tag{3}$$

where θ^m is the parameters learned for skill m.

We design two main functions of the skill agent, generator model and ranker model, parameterized as θ^m_{gen} and θ^m_{rnk} for skill m, respectively. For θ_{gen}, we aim to generate responses from response space $\mathbb{U}$ in a token-by-token manner, and thus employ a dodecaDialogue (Shuster et al., 2020) model, a modification of a transformer Seq2Seq architecture. On the other hand, for θ_{rnk}, we consider the response space $\mathbb{U}$ as a list of alternatives to pick the correct response, and thus employ a poly-encoder (Humeau et al., 2020) model, a transformer-based retrieval architecture, to score and rank response candidates. Both θ_{gen} and θ_{rnk} are fine-tuned on individual single-skill datasets[1].

[1] On the average, generator and ranker models show around 10 perplexity and 90 accuracy on their respective datasets.

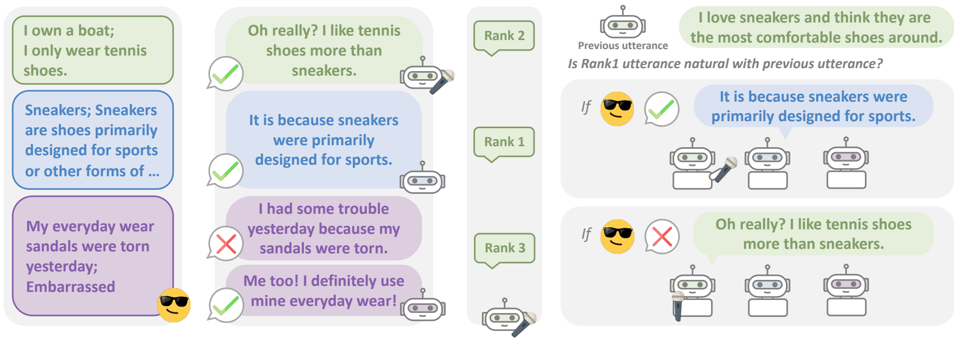


Figure 1: Illustration of BOTSTALK framework. Green, blue, and purple indicate skill types of P, K, E, respectively.

While all skill agents simulate what response to annotate, only one skill agent is given priority over other skill agents, to "speak" the response per dialogue turn for the dialogue annotation, conditioned on a set of skill contexts $\tilde{stx}$ and the dialogue context dtx_t. We call this *active agent*. This priority may be passed to another skill agent such that the current active agent is deactivated, and another skill agent will be newly activated to speak.

Moderator Agent 😎 A critical constraint for skill agents is that neither the generator nor the ranker for skill m is able to read other skill contexts in $\tilde{stx}$ for different skills. For a skill agent, considering all possible skill contexts in multi-skill dialogues is non-trivial. Instead, as an omniscient oracle for all skill contexts $\tilde{stx}$, we aim to develop another participant named moderator agent, which mediates the conversational flow for desirable multi-skill dialogue annotation. To examine the relevance of the response res_t with all skill contexts $\tilde{stx}$ or the dialogue context dtx_t, the moderator agent has a decision function $g : (\tilde{stx}, dtx_t, res_t) \mapsto \mathbb{A}$ where $\mathbb{A}$ is an action space (*i.e.*, approval or refusal) for the given response.

4.2 Phase 1: Simulate what to speak

We integrate different dialogue setups from multiple single-skill datasets as seed information to start a conversation (detailed in Appendix C.3). For a dialogue episode, dialogue context is initialized as an utterance pair (*i.e.*, two dialogue turns) randomly sampled from a single-skill dataset $\mathcal{D}_m$, and the skill agent for skill m becomes the initial active agent. Then, for a generalizable dialogue setup, we retrieve the most relevant skill contexts from each of all input datasets $\mathcal{D}_1, ..., \mathcal{D}_M$ for the seed dialogue context with TF-IDF (Chen et al., 2017)[2].

[2]While we use a simple IR baseline as lower bound since it is not our main focus, one can easily try different IR system.

In the first phase of BOTSTALK, all skill agents simulate their own responses for the next dialogue turn. Formally, given a skill context stx_m and the current dialogue context dtx_t in a dialogue episode, a skill agent for skill m generates a plausible response $res_{m,t}$ as

$$res_{m,t} = \operatorname*{argmax}_{res_t \in \mathbb{U}} P(res_t | stx_m, dtx_t; \theta^m_{gen}) \cdot g(\tilde{stx}, res_t) \tag{4}$$

where $g(\cdot)$ is the function of the moderator agent, which we discuss in the subsequent section.

Depending on individual skills, every skill agent returns its skill-relevant response. For example, as shown in Figure 1, when "*I love sneakers and think they are the most comfortable shoes around.*" is given as dtx, the skill agent for skill P generates a personal response "*Oh really? I like tennis shoes more than sneakers.*" as res_{P} based on a given persona. Meanwhile, the skill agents for skill K and E generate a knowledgeable response "*It is because sneakers were primarily designed for sports.*" as res_{K} and a empathetic response "*Me too! I definitely use mine everyday wear!*" as res_{E}.

4.3 Phase 2: Check dialogue consistency

It is well known that neural dialogue systems lack consistency (Li et al., 2016; Welleck et al., 2019). Furthermore, as a skill agent uses the specific skill context stx_m instead of $\tilde{stx}$ for response generation, the response is more likely to be semantically in conflict with other skill contexts in $\tilde{stx}$. Suppose a stx_{P} is "*I wear sneakers everyday*" and a res_{E} is "*I had some trouble yesterday because my sandals were torn*". This response is inappropriate because "*yesterday because my sandals were torn*" is contradictory to "*I wear sneakers everyday*". Therefore, the moderator agent, who has access to all skill contexts $\tilde{stx}$, filters out conflicting response candidates to preserve dialogue consistency.

Specifically, the moderator agent leverages natural language inference (NLI), a task of determining whether a hypothesis sentence can be inferred from the given premise sentence. The hypothesis sentence is classified into three categories: ENTAIL (true), NEUTRAL (undetermined), and CONTRADICT (false). Based on the NLI classifier, the decision function of the moderator agent is defined as

$$g(\tilde{stx}, res_t) = \begin{cases} 1, & \text{NLI}(\tilde{stx}, res_t) \not\to \text{CONTRADICT} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

which represents approval/refusal of res_t conditioned on $\tilde{stx}$. A skill agent for skill m repeatedly generates new response candidates until its response is approved, as described in Equation 4.

For NLI classifier, we use a RoBERTa (Liu et al., 2019) model trained on MNLI (Williams et al., 2018)[3], which is widely used in fact checking systems (Kim et al., 2021a)[4]. Overall, about 50% of utterances are classified as CONTRADICT by NLI classifier. Out of all utterances classified as CONTRADICT, about 70% are in conflict with other types of skill contexts (Figure 2). The result demonstrates that skill agents indeed generate inconsistent responses due to the restricted access to other skill contexts. We also find that the overall proportion of utterances conflicting with stx_{P} is relatively high, apparently because stx_{P} contains more distinct descriptions than stx_{K} and stx_{E}.

4.4 Phase 3: Speak or pass the mic

The objective of the last phase is to score a set of response candidates and select a final response when given the skill contexts and dialogue context. To this end, we leverage the active agent and the moderator agent, taking into account a balance between skill blending and skill grounding.

Let $\mathbb{U}_{res}$ be the set of response candidates $res_{1,t}, ..., res_{M,t}$ from all skill agents. The active skill agent identifies the most appropriate response res_t^* in $\mathbb{U}_{res}$ based on its ranker model θ_{rnk}^m, then asks the moderator agent to attach the selected response into the next dialogue context dtx_{t+1} for annotation. Formally, we define such process as

$$res_t^* = \operatorname*{argmax}_{res_t \in \mathbb{U}_{res}} P(res_t | stx_m, dtx_t; \theta_{rnk}^m) \cdot g(dtx_t, res_t) \tag{6}$$

where $g(\cdot)$ is the function of the moderator agent. To compute $g(dtx_t, res_t)$, the moderator agent adopts a skill classifier $\mathcal{P}$ that identifies corresponding skill for the response. We use a BERT (Devlin et al., 2019) model trained on utterances in $\mathcal{D}_m$ and their corresponding skill labels m for all skill types $\mathbb{M}$[5]. Once $\mathcal{P}$ is learned, the decision function of the moderator agent is defined as

$$g(dtx_t, res_t) = \begin{cases} 1, & \text{KL}(\mathcal{P}(res_{t-1}^*) || \mathcal{P}(res_t)) < \alpha \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where res_{t-1}^* is the last utterance of dtx_t and $\mathcal{P}(\cdot) \in \mathbb{R}^M$ outputs the skill distribution of the response. Based on KL divergence between two distributions, $g(dtx_t, res_t)$ is discretized as the approval/refusal decision by a pre-defined threshold α (Figure 3a). Once the moderator agent accepts the candidate res_t from an inactive agent as the final response, the active agent passes the mic, or the priority for annotation, to the inactive agent.

In practice, we compute entropy of the skill distributions of all utterances to investigate whether there is room for shifting between skills. The value of entropy indicates the uncertainty of the skill type of an utterance: utterances with high entropy values are uncertain, generic responses. Figure 3b shows that the number of generic utterances is far from negligible, suggesting that there are opportunities to shift to other skills and thus both skill blending and grounding can be satisfied in a conversation.

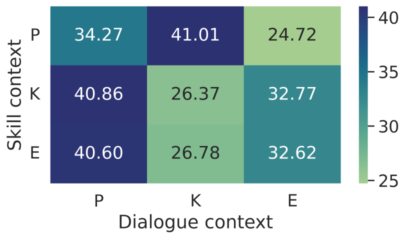
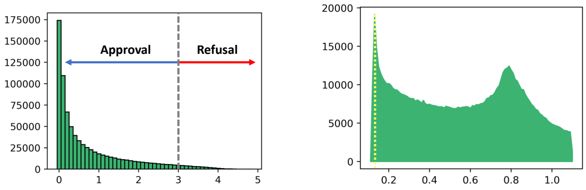


Figure 2: Percentages of utterances which are classified as CONTRADICT via NLI classifier, broken down by the type of skill contexts.



(a) KLD distribution

(b) Entropy distribution

Figure 3: KL divergence between skill distributions of consecutive utterances (left) and entropy of skill distributions for all utterances (right).

[3] Dialogue NLI (Welleck et al., 2019) is biased to ConvAI2.

[4] The RoBERTa model shows 90.59 accuracy on MNLI.

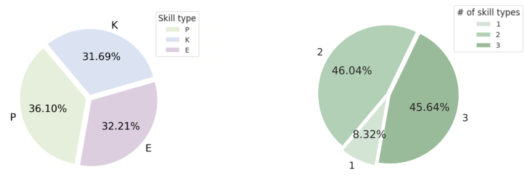
[5] The BERT model shows 81.95 accuracy at inference time.

Skill context from ConvAI2	**Skill context from ConvAI2**
I like to play soccer; I like to read; ...	I have 3 children; I am a karate black belt; ...
Skill context from WoW	**Skill context from WoW**
Nike Inc.	Nike Inc.; ... multinational corporation ... Air Jordan ...
Skill context from ED	**Skill context from ED**
I really like this girl at my job, but I am ... ; Apprehensive	None
Dialogue	
A: Do you have much experience using the different types of cleats? Which do you like best? (P)	
B: I have a little. I also know about Air jordans, a brand of footwear also popular with athletes. (K)	
A: I enjoy Air jordans as well. I like to play soccer, and it's extremely hard to get good footwear. (P)	
B: I agree. Air jordans are good for calves, and it's really easy to pull off. (K)	
A: Air jordans are generally made with the most material so it makes sense they'd be easy to use. (K)	
B: And now shoes can go together with clothes as well, like any other type of material. (K)	
A: Do you know much about aeros then? They have excellent fit and beauty. (K)	
B: I have a hard time finding it but they are great shoes. (P)	
A: I hope you can find ones that are comfortable to you. (E)	
B: Yes. I hope you can get those shoes too. (E)	

Table 2: A conversation from the BS𝔹T dataset. Speaker A is given five personas, one topic, and a situation with an emotion (top left), while speaker B is given five personas, one topic, and seven knowledge resources (top right).

Dataset	Dialogues	Utterances	Turns
ConvAI2	19,893	145,873	7
WoW	22,311	201,999	9
ED	24,850	51,245	2
BST	6,808	58,575	8
BS𝔹T	300,000	3,000,000	10

Table 3: Statistics of dialogue datasets: the number of dialogues, utterances, and the average number of turns.



(a) Utterances (b) Dialogues

Figure 4: Percentages of utterances with respect to the skill type (left) and dialogues broken down by the number of skill types in the dialogue (right).

5 Blended Skill BotsTalk (BS𝔹T)

5.1 Data Statistics

We collect a multi-skill dialogue dataset, namely Blended Skill BotsTalk (BS𝔹T), using BotsTalk. The dataset consists of 300K dialogues with 3M utterances. Each utterance is labeled using the skill classifier with skill annotation (*e.g.*, personality from ConvAI2, knowledge from WoW, or empathy from ED), including skill type and skill distribution. Table 2 presents an example from BS𝔹T. As shown in Table 3, one of the salient features of BS𝔹T is its scalability, mainly because BS𝔹T is composed of bot-bot conversations collected through a machine-sourced approach while other datasets comprise crowdsourced human-to-human conversations.

Skill Blending Figure 4 summarizes the results of skill annotation in BS𝔹T dataset. Overall, the skill annotation percentages are 36.10% for personality, 31.69% for knowledge, and 32.21% for empathy (Figure 4a). Moreover, over 90% of the dialogues demonstrate at least 2 of the 3 skills within a single dialogue (Figure 4b), suggesting the vast majority of conversations feature more than one skill type.

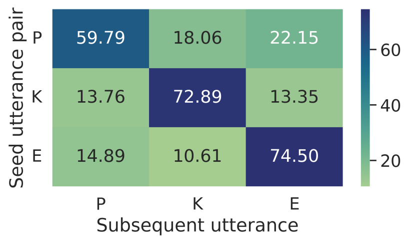


Figure 5: Percentages of the skill type of utterances subsequent to the utterance pairs, broken down by provenance skills of the utterance pairs.

Skill Grounding Although we focus on blending skills, multi-skill dialogue datasets should also contain sufficient sessions grounded to specific skills in conversations for a model to learn the ability of skill grounding. For that, we explore the continuity of skills by investigating skill types of utterances subsequent to seed utterance pairs whose provenance skills are determined by their original datasets. For all skills, more than half of the utterances followed by the utterance pairs are labeled as the same skill types of the utterance pairs (Figure 5).

Model	Retrieval R@1	R@5	MRR	Generative Avg. BLEU
ConvAI2	75.92	94.04	83.96	3.75
WoW	67.48	89.57	77.11	4.08
ED	65.96	88.69	76.10	3.15
BST	75.92	94.76	84.14	4.31
BS𝔹T	**80.68**	**95.79**	**87.39**	**4.38**

Table 4: Automatic evaluation on BST benchmark.

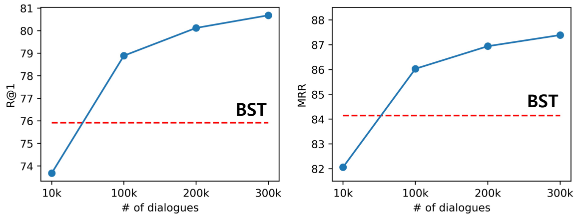


Figure 6: The effect on model performance by varying the number of dialogues, measured by R@1 and MRR.

	BST (Win %)	vs.	BS𝔹T (Win %)
Engagingness	43		57
Interestingness	47		53
Humanness	44		56

Table 5: Human evaluation for pairwise comparison between BST and BS𝔹T datasets.

5.2 Experimental Setups

We conduct a set of experiments to test our BS𝔹T over BST benchmark through automatic and human evaluation. To the best of our knowledge, BST benchmark is the only multi-skill dialogue benchmark which gauges how successful a model is at blended objective as well as grounded objective.

Following Smith et al. (2020), we consider the retrieval task as our primary task and adopt a 256-million parameter poly-encoder (Humeau et al., 2020) pre-trained on pushshift.io Reddit dataset as a base architecture. We further include the generative task as our secondary task and adopt a pre-trained BART (Lewis et al., 2020) as a base architecture. We fine-tune these base architectures on individual datasets, *i.e.*, ConvAI2, WoW, ED, BST, and BS𝔹T, and use them as our baselines. We describe implementation details in Appendix D.

For the retrieval task, we report recall@k (R@k), where each test example has 100 possible candidates to select from, as well as mean reciprocal rank (MRR). For the generative task, we compute the average score of BLEU-1, -2, -3, -4 (Avg. BLEU).

5.3 Automatic Evaluation

The results of retrieval and generative models on BST benchmark are shown in Table 4 (detailed results are in Appendix E). We observe that multi-skill models, *i.e.*, BST and BS𝔹T models, are superior to single-skill models, *i.e.*, ConvAI2, WoW, and ED models on the BST benchmark. As multi-skill dialogues require an understanding of both skill blending and grounding, single-skill models who are only grounded to each of skills struggle to seamlessly blend them over the course of a conversation, whereas multi-skill models are able to not only exhibit individual skills but also combine different skills in a conversational flow. In particular, BS𝔹T model outperforms all of the baselines on all automatic metrics. This indicates that our machine-sourced dataset works properly as the training resource to learn the ability of blending skills as well as grounding to various skills.

To explore the impact of BS𝔹T size on the model performance, we fine-tune the retrieval architecture on the BS𝔹T datasets of varying scales. Figure 6 illustrates the performance of BS𝔹T model in terms of R@1 and MRR when the size of the dataset gradually increases. BS𝔹T300K model achieves a significant performance boost from BS𝔹T10K model, surpassing BST model and showing the best performance. This result not only affirms the importance of large-scale training for building multi-skill chatbots but also indicates the potential of BS𝔹T dataset, as our dataset is collected by automatic BotsTalk framework without human intervention (*i.e.*, no manual annotation or verification).

5.4 Human Evaluation

To assess the quality of BS𝔹T dataset, we perform human evaluation by employing ACUTE-Eval (Li et al., 2019), a popular metric for multi-turn dialogue evaluation (Dinan et al., 2020a; Li et al., 2020). We randomly sample 100 dialogues from the BST and BS𝔹T datasets respectively, and then ask human evaluators to compare each pair of dialogues over three axes: engagingness, interestingness and humanness. We provide more details for the specific settings in Appendix F. Table 5 shows that for all metrics, BS𝔹T dataset achieves comparable and even slightly higher win percentages over BST dataset. This ensures the quality of BS𝔹T and thus validates that BotsTalk framework can be a viable alternative to crowdsourcing when constructing multi-skill dialogue datasets.

Model	Retrieval R@1	R@5	MRR
MTL	78.95	95.37	86.23
BS𝔹T	80.68	95.79	87.39
MTL + BS𝔹T100K	80.94	95.79	86.92
MTL + BS𝔹T200K	82.01	96.37	87.83
MTL + BS𝔹T300K	82.10	96.79	88.04

Table 6: Performance of MTL models on BST benchmark, reported by R@1, R@5, and MRR.

Qualitative Analysis Although most of conversations from BS𝔹T satisfies the desirable characteristics for multi-skill dialogues as evidenced by a set of experiments, any side effect may occur since the dialogues are collected through automatic annotation. Therefore, we select best-case and worst-case examples (each in Table 12 and Table 13) and provide empirical results including three types of error cases. First, in a few cases, speakers repeat greeting (about 3 turns) at the end of conversation. This is mainly because we set dialogues to be of fixed length, while the conversation may end earlier than the given turns. Second, as observed in Figure 4b, we find a number of dialogues that only features one skill type. Lastly, some responses tend to show little relevance with the skill contexts of their corresponding skills, although they are grammatically sound and meaningfully move the conversations forward. We give a deeper analysis in Appendix G.

5.5 Analysis on Multi-task Learning

Given an access to multiple single-skill dialogue datasets, a straightforward approach of developing a multi-skill chatbot is to multi-task on all of them during fine-tuning step (Shuster et al., 2020; Roller et al., 2021). Therefore, we consider MTL model, a poly-encoder (Humeau et al., 2020) pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, WoW, and ED. We further fine-tune the MTL model on BS𝔹T datasets of varying scales sequentially, to probe the effectiveness of BS𝔹T as a training resource for multi-task training scheme. In Table 6, MTL model lags behind BS𝔹T model on BST benchmark, but performs noticeably better when fine-tuned on BS𝔹T dataset in addition. Such improvement in the performance is an encouraging sign that BS𝔹T is orthogonally applicable to multi-tasking strategy. We observe that the performance gain becomes marginal when the size of the dataset increases. We hypothesize that as multi-task learning and BS𝔹T are parameterized and materialized knowledge for multi-skill dialogues respectively, there can be an overlap between the knowledge dialogue systems learn. We leave the mitigation of such overlap for future work. Nevertheless, MTL model achieves the best performance when fine-tuned on BS𝔹T300K dataset, as it compensates the overlap with its size.

Model	ConvAI2	WoW	ED
Retrieval			
ConvAI2	0	−10.95	−14.91
WoW	−30.56	0	−16.95
ED	−27.15	−10.74	0
BST	−14.33	−8.67	−14.70
BS𝔹T	**−2.12**	**−1.92**	**−1.52**
Generative			
ConvAI2	0	−3.27	−2.89
WoW	−3.60	0	−2.51
ED	−3.74	−3.12	0
BST	−2.68	−2.78	**−1.86**
BS𝔹T	**−0.98**	**−2.19**	−1.95

Table 7: Performance on single-skill benchmarks, measured by $\Delta_{\mathrm{R@1}}$ for retrieval models and $\Delta_{\text{Avg. BLEU}}$ for generative models.

5.6 Analysis on Skill Grounding Ability

To gain more insights into individual skill grounding ability, we evaluate the baselines on single-skill benchmarks, *i.e.*, ConvAI2, WoW, and ED benchmarks (detailed results are in Appendix E). We compute a relative performance drop $\Delta_{\mathrm{R@1}}$ for retrieval models and $\Delta_{\text{Avg. BLEU}}$ for generative models over the best performing model on the respective benchmark, which gives us upper bound we aim for with our model. Results are summarized in Table 7. As expected, each of single-skill models perform best on their original benchmarks but not as well on other benchmarks, whereas the multi-skill models show more well-rounded performance across all benchmarks. In particular, BS𝔹T model outperforms BST model on most cases and even achieves comparable performance to single-skill models on corresponding single-skill benchmarks. This suggests that BS𝔹T is effective to not only inject the ability of blending various skills but also maintain the ability for grounding specific skill.

6 Conclusion

To build multi-skill chatbots, we construct a large-scale dialogue dataset BS𝔹T through automatic BotsTalk framework. We validate its efficacy as training resource by experiments and analyses.

7 Limitations

We summarize the error patterns in BSBT and discuss potential directions to improve BOTSTALK framework. First, our framework *always* produces a fixed-length conversation, even when the conversation ends earlier than the given turns. As shown in Example 4 from Table 13, this often results in generic and repetitive responses at the end of a dialogue. In future work, we are interested in training dialogue models that understand *when* to end a conversation based on the context. Second, as observed in Figure 4b, a few dialogues fail to cover explicit transitions between multiple skills. For instance, Example 5 from Table 13 only features one skill type and lacks the nature of skill blending, which may hinder models from learning diverse communicative skills. Third, some responses, particularly those near the end of conversations, tend to show little relevance with their corresponding skill contexts. We conjecture that as a conversation flows, skill agents condition their responses more on the dialogue context, which is likely to be longer than the pre-defined skill contexts. To enhance the model's ability of skill grounding, one can employ conditional training by modeling the progress of the conversation and controlling input contexts. We leave these issues for future research.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments. This research was partially supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022-11-0941).

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of EMNLP*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020b. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of ICLR*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of ICLR*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020a. Sequential latent knowledge selection for knowledge-grounded dialogue. In *Proceedings of ICLR*.

Byeongchang Kim, Hyunwoo Kim, Seokhee Hong, and Gunhee Kim. 2021a. How robust are fact checking systems on colloquial claims? In *Proceedings of NAACL*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020b. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of EMNLP*.

Sungdong Kim, Minsuk Chang, and Sang-Woo Lee. 2021b. Neuralwoz: Learning to collect task-oriented dialogue via model-based simulation. In *Proceedings of ACL*.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyun Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. In *Proceedings of ACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of ACL*.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of ACL*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of IJCAI*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2021. The adapter-bot: All-in-one controllable conversational model. In *Proceedings of AAAI*.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian J McAuley. 2020a. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of EMNLP*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020b. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of EMNLP*.

Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2021. Simulated chats for building dialog systems: Learning to generate conversations from instructions. In *Findings of EMNLP*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL*.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of EACL*.

Sashank Santhanam and Samira Shaikh. 2019. Emotional neural language generation grounded in situational contexts. In *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*.

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*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of ICLR*.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and J. Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of ACL*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of ACL*.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chats to enhance task-oriented dialogues. In *Proceedings of NAACL*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of ACL*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of ACL*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*.

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.

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 ACL*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of EMNLP*.

A Overview

In the following sections, we explore more details on BOTSTALK framework and BSBT dataset. In Appendix B, we lay out the details of single-skill dialogue datasets and how they are incorporated into BOTSTALK framework to construct BSBT dataset.

We provide implementation details in Appendix C for all component models of participants in BOTSTALK framework. We also provide implementation details of baselines used for experiments in Appendix D. The evaluation results on all benchmarks used in this paper are in Appendix E. The specific settings for human evaluation are in Appendix F. Finally, we present a number of conversation examples from BS$\mathbb{B}$T and further analyze its strengths and weaknesses in Appendix G.

B Single-skill Datasets into BOTSTALK

We describe details on the single-skill dialogue datasets used to construct BS$\mathbb{B}$T and elaborate on how they are incorporated into BOTSTALK framework to construct our dataset. Example dialogues from the single-skill dialogue datasets *i.e.*, ConvAI2, WoW, ED, are shown in Table 9, 10, 11.

To integrate different dialogue setups from the single-skill dialogue datasets, we follow the basic settings for constructing a dialogue dataset, assuming a multi-turn, one-to-one conversation between two speakers. We simulate turn-taking in a conversation by switching two different sets of skill contexts for the input skill context stx to a dialogue model f in a skill agent.

B.1 ConvAI2

ConvAI2 (Dinan et al., 2020b) is a dataset based on PersonaChat (Zhang et al., 2018). ConvAI2 dataset contains of more than 140K utterances from conversations in which each of paired crowdworkers is given a role based on their persona description and gets to know their partner. Specifically, the speaker pairs are each assigned profiles from a set of 1155 possible personas, each consisting of at least 5 profile sentences. The personas are collected through crowdsourcing, where the workers are asked to create natural, descriptive profiles that contain typical topics of human interest. Workers are also asked to keep each profile sentence short, *i.e.*, no longer than 15 words.

Following the setting from ConvAI2, we define a skill context stx_{P} as a profile comprising 5 distinct persona sentences. We then provide two different skill contexts as the input to the dialogue model f in an alternating manner to simulate turn-taking.

B.2 Wizard of Wikipedia

Wizard of Wikipedia (Dinan et al., 2019) task involves discussing a given topic in depth, where the goal is to both engage the partner as well as display expert knowledge. The dataset consists of 194K utterances over 1250 topics, where each conversation begins with a randomly chosen topic. A retrieval system over Wikipedia is used to retrieve articles from which the dialogues are grounded during the human-human crowdsourced conversations. The topics are also crowdsourced and range from commuting to Gouda Cheese to Arnold Schwarzenegger. Each conversation in the dataset involves two speakers named the apprentice and the wizard: the apprentice aims at delving deeply into a topic whereas the wizard uses knowledge in articles retrieved from Wikipedia to craft a relevant reply. Specifically, given a topic derived from the dialogue context, the apprentice keeps the conversation engaging and talks eagerly about a topic, while the wizard responds to the apprentice based on the first paragraphs of 7 relevant Wikipedia articles provided by the retrieval system.

In our setting, we use a simpler version of Wizard of Wikipedia task, which ignores the retrieval aspect of the task. We first specify the topic of the conversation, which is the same for the apprentice and wizard. The skill context stx_{K} of the apprentice is thus defined as the given topic, while the skill context stx_{K} of the wizard is defined as a topic and 7 relevant knowledge sources.

B.3 Empathetic Dialogues

Empathetic Dialogues (Rashkin et al., 2019) is a dataset includes 50K utterances of crowdworker conversations grounded in an emotional situation. In the conversation, one speaker describes a personal situation based on an emotion label and the other speaker, named listener, displays empathy in their response. Specifically, a pair of workers (*i.e.*, speaker and listener) are asked to choose an emotional word each, depict a situation in 1-3 sentences based on the label, and engage in a short conversation of 4-8 utterances about each of the situations. Neither of the workers, whether they be the speaker or the listener, can see the emotion label and the situation description of their partner, so that they must refer only to cues within the conversation for their response.

In our setting, we define the situation description and its corresponding emotion label as the skill context stx_{E} of the speaker. Note that we do not define the skill context stx_{E} of the listener for our framework, so that the dialogue system is trained to show empathy based solely on the conversation.

C Implementation Details of BotsTalk

For the implementation of BotsTalk framework, we employ ParlAI[6] toolkit, which is specialized in training and evaluating dialogue systems. We will release our agents and dataset for public use.

C.1 Skill Agents

In our BotsTalk framework, a skill agent leverages both generator model and ranker model.

Given a stx_k and dtx as input, a generator model of skill agent generates a response for the next dialogue utterance. For the generator model, we employ a dodecaDialogue (Shuster et al., 2020). The dodecaDialogue model is a modification of transformer seq2seq architecture, which has a 8-layer encoder, 8-layer decoder with 512 dimensional embeddings and 16 attention heads. We fine-tune the dodecaDialogue models on ConvAI2, WoW, and ED, respectively. We use nucleus sampling as the decoding strategy for generative models at inference time. The deodecaDialogue model shows 11.19, 8.46, and 11.08 perplexity on ConvAI2, WoW, and ED.

Given a stx_k and dtx as input, a ranker model of skill agent selects the next dialogue utterance by scoring a large set of candidate responses and outputting the one with the highest score. For the ranker model, we employ the poly-encoder architecture of Humeau et al. (2020). The poly-encoder encodes global features of the context using multiple representations, which are attended to by each possible candidate response. This final attention mechanism gives improved performance over a single global vector representation whilst still being tractable to compute compared to simply concatenating input and output as input to a Transformer. The poly-encoder has state-of-the-art performance on a number of dialogue tasks when compared to other retrieval models, and also gives comparable performance to the winning generative models on the ConvAI2 competition task in terms of human evaluation.

More specifically, we consider a 256M parameter poly-encoder model, which has 12 layers, 12 attention heads, and a hidden size of 768. We pre-train our poly-encoder on pushshift.io Reddit dataset and then fine-tune on ConvAI2, WoW, and ED, respectively. We use a large number of negatives by considering the other batch elements as negative training samples, avoiding recomputation of their embeddings. We use the Adamax optimizer without weight decay, a learning rate of 5e-5 with batch size 128, epoch 8. The learning rate decays by a factor of 0.4 upon plateau of the loss evaluated on the valid set every half epoch. The best parameters are chosen based on R@1 score. The poly-encoder model shows 89.41, 91.01, and 63.26 R@1 on ConvAI2, WoW, and ED, respectively.

C.2 Moderator Agent

In BotsTalk framework, the moderator agent leverages NLI classifier and skill classifier.

Given a response $res_{k,t}$ from a skill agent of skill k and the set of skill contexts $\tilde{stx}$, the NLI classifier is designed to determine whether a response candidate contradicts any of the skill contexts. For NLI classifier, we employ the public HuggingFace[7] implementation of a RoBERTa-Large model (Liu et al., 2019) fine-tuned on the Multi-Genre NLI dataset (Williams et al., 2018). The RoBERTa model shows 90.59 accuracy on MNLI validation set. We regard each response candidate res_k as a hypothesis sentence and each skill context $stx_k \in \tilde{stx}$ as a premise sentence, then conduct unidirectional NLI classification between stx_k and res_k, determining whether a hypothesis sentence res_k can be inferred from the given premise sentence stx_k for all response candidates.

Given a response res_t, the skill classifier identifies the skill of the response among all skills represented in the skill context set $\tilde{stx}$. For skill classifier, we employ a BERT-base (Devlin et al., 2019) model. We trained the model on utterances from ConvAI2, WoW, ED train sets and their corresponding skills k as labels. The model was trained with a batch size of 16, a learning rate of 2e-5 and epoch 3. The BERT model shows 81.95 accuracy on utterances from ConvAI2, WoW, ED test sets.

C.3 Skill Context Retrieval

In Appendix B, we explore how we define the skill context for each skill considering the original settings from single-skill dialogue datasets. We now describe how we construct the seed information (*e.g.*, skill contexts for each of skil types and dialogue context of 2 turns) for skill agents to start a conversation. We first collect consecutive utter-

[6] https://github.com/facebookresearch/ParlAI

[7] https://github.com/huggingface

Model	Retrieval				Generative			
	R@1	R@5	R@10	MRR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Evaluation on BST benchmark								
ConvAI2 (Zhang et al., 2018)	75.92	94.04	97.19	83.96	10.97	2.88	0.86	0.32
WoW (Dinan et al., 2019)	67.48	89.57	94.33	77.11	12.00	3.20	0.80	0.31
ED (Rashkin et al., 2019)	65.96	88.69	93.80	76.10	9.36	2.47	0.61	1.72
BST (Smith et al., 2020)	75.92	94.76	97.83	84.14	12.19	3.65	1.06	0.37
BS$\mathbb{B}$T (Ours)	80.68	95.79	98.16	87.39	11.92	3.74	1.28	0.57
Evaluation on ConvAI2 benchmark								
ConvAI2 (Zhang et al., 2018)	88.46	98.92	99.71	93.03	17.69	7.21	2.96	1.15
WoW (Dinan et al., 2019)	57.90	86.85	95.80	70.59	10.97	2.82	0.64	0.18
ED (Rashkin et al., 2019)	61.31	89.44	96.69	73.53	10.64	2.60	0.63	0.20
BST (Smith et al., 2020)	74.13	95.64	98.80	83.37	12.42	4.00	1.40	0.48
BS$\mathbb{B}$T (Ours)	86.34	98.00	99.42	91.46	16.12	5.88	2.12	0.81
Evaluation on WoW benchmark								
ConvAI2 (Zhang et al., 2018)	79.84	96.97	98.84	87.62	7.88	1.89	0.53	0.15
WoW (Dinan et al., 2019)	90.79	99.28	99.66	94.67	14.85	5.63	2.02	1.02
ED (Rashkin et al., 2019)	80.05	96.25	98.37	87.34	8.22	2.17	0.48	0.18
BST (Smith et al., 2020)	82.12	97.57	98.99	89.11	9.05	2.58	0.60	0.18
BS$\mathbb{B}$T (Ours)	88.87	98.84	99.28	93.44	10.13	3.31	0.97	0.35
Evaluation on ED benchmark								
ConvAI2 (Zhang et al., 2018)	47.90	76.14	85.87	60.60	8.66	1.79	0.60	0.32
WoW (Dinan et al., 2019)	45.86	74.79	85.15	58.94	9.65	2.20	0.71	0.34
ED (Rashkin et al., 2019)	62.81	88.91	94.58	74.18	14.86	5.14	1.95	1.00
BST (Smith et al., 2020)	48.11	77.09	86.96	61.04	11.22	2.64	1.06	0.58
BS$\mathbb{B}$T (Ours)	61.29	87.39	93.59	72.70	10.45	2.74	1.21	0.70

Table 8: Performance of retrieval and generative dialogue systems on all benchmarks used in this paper.

ance pairs from ConvAI2, WoW, and ED as seed utterance pairs and define it as dtx. We then follow the convention of past research, which inject a target communicative skill to dialogue systems by providing an extra description about the specific skill, *i.e.*, skill context. As we aim to build a multi-skill chatbot, there is a need for integrating different dialogue setups from multiple single-skill datasets. For a generalizable dialogue setup, we retrieve relevant skill context for each skill by querying dtx which is the seed utterance pair. Here, to gather seed information as much as possible, we match top-5 relevant skill contexts per utterance pair, which we give us five times the seed information. We use TF-IDF (Chen et al., 2017) to find relevant skill contexts and a SQLite database for storing the sparse TF-IDF matrix. Note that we use a simple IR baseline as a lower bound since it is not our main focus. One can easily try other IR systems for more sophisticated setting.

D Implementation Details of Baselines

We conduct a set of experiments with retrieval and generative tasks to cover diverse baselines. We provide training details of these baseline models.

Retrieval Task We adopt a 256-million parameter poly-encoder (Humeau et al., 2020) pre-trained on pushshift.io Reddit dataset as a base architecture for the retrieval task. We fine-tune this base architecture on individual datasets, *i.e.*, ConvAI2, WoW, ED, BST, and BS$\mathbb{B}$T for 8 epochs with batch size 128 and learning rate 5e-5. The learning rate decays by a factor of 0.4 upon plateau of the loss evaluated on the valid set every half epoch. The best parameters are chosen based on R@1 score.

Generative Task As a base architecture for the generative task, we adopt a pre-trained BART (Lewis et al., 2020) with 12 layers in each of the encoder and decoder. We fine-tune this base architecture on individual datasets, *i.e.*, ConvAI2, WoW, ED, BST, and BS$\mathbb{B}$T for 3 epochs with batch size 32 and learning rate 1.0. The learning rate decays by a factor of 0.3 upon plateau of the loss evaluated on the valid set every half epoch. The best parameters are chosen based on accuracy score. We use greedy sampling as a decoding strategy.

E Additional Performance

We provide the evaluation results on all dialogue benchmarks used in this paper, *i.e.*, BST benchmark for multi-skill benchmark, and ConvAI2, WoW, ED behcmarks for single-skill benchmarks. We report

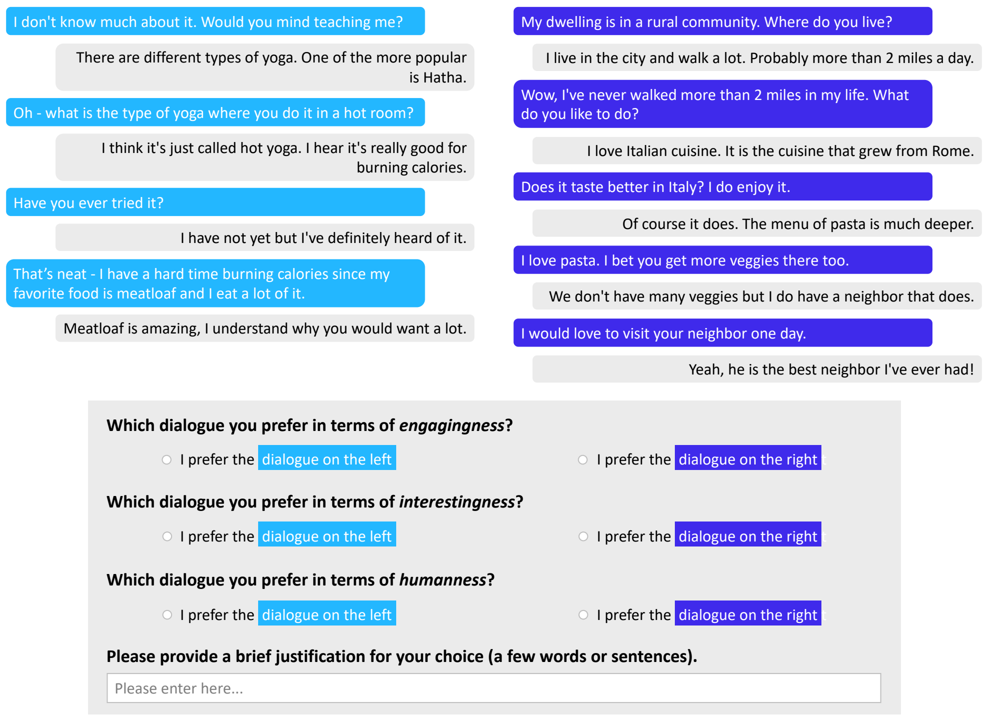


Figure 7: Interface for human evaluation, which provides a pair of dialogues from BST (left) and BS$\mathbb{B}$T (right).

R@1, R@5, R@10, MRR for evaluating retrieval dialogue models, and BLEU-1, BLEU-2, BLEU-3, BLEU-4 for evaluating generative dialogue models. Table 8 presents the performance of all of the baselines on all dialogue benchmarks.

F Settings of Human Evaluation

We describe the specific settings for human evaluation that we perform to validate the quality of BS$\mathbb{B}$T dataset. Specifically, we employ ACUTE-Eval (Li et al., 2019), a widely used metric for multi-turn dialogue evaluation (Dinan et al., 2020a; Li et al., 2020). We randomly sample 100 dialogues from BST and BS$\mathbb{B}$T datasets respectively. We only include dialogue contexts and exclude skill contexts for anonymity, since the skill contexts of BST and BS$\mathbb{B}$T are distinguishable. Figure 7 shows the interface used for human evaluation. We ask judges to compare each pair of conversations over three axes: engagingness, interestingness and humanness. The wording of questions is presented as follows:

- **Engagingness:** Who would you prefer to talk to? Which version is more likely to hold your attention and make you want to hear more?
- **Interestingness:** Who would you say is more interesting? Which version arouses your curiosity or tells you something new or useful?
- **Humanness:** Who would you say sounds more human? Which version is more natural and personable?

G Qualitative Analysis

In what follows we conduct qualitative analysis on BS$\mathbb{B}$T dataset based on diverse samples. In each dialogue episode from BS$\mathbb{B}$T, one speaker is given five personas as stx_{P}, one topic as stx_{K}, and a situation description and emotion as stx_{E}, while another speaker is given five personas as stx_{P}, the topic and seven knowledge resources as stx_{K}, and nothing for stx_{E}. Each speaker is conditioned on their corresponding set of skill contexts, and annotates the response turn by turn. To give a deeper analysis, we select best-case (*i.e.*, cherry-picked) and worst-case (*i.e.*, lemon-picked) examples from BS$\mathbb{B}$T dataset. Table 12 and Table 13 shows cherry-picked examples and lemon-picked examples, respectively. We also present more dialogue examples randomly sampled from BS$\mathbb{B}$T in Table 14 and Table 15.

Cherry-picked Dialogues In Table 12, we provide cherry-picked dialogue samples from the BS$\mathbb{B}$T dataset. Example 1 and 2 both show that the conversation contains a sufficient amount of utterances to learn skill grounding for skill type P and E. In Example 3, each skill type is sustained for more than 2 turns, which allows dialogue systems to learn the ability of skill grounding. Example 3 also demonstrates all of the skill types are blended in a natural and reasonable way, which enables dialogue systems to weave multiple skills in a single conversation properly.

Lemon-picked Dialogues In Table 13, we present lemon-picked conversations from BS$\mathbb{B}$T dataset, which fail to satisfy the key criteria of multi-skill dialogues. In Example 4, speaker A and B repeats greeting such as "*You as well. Have a good weekend as well*" and "*You too. Have a good weekend as well*" for about 3 turns. This is mainly because we set conversations to be of fixed length (*i.e.*, 10 turns), while the conversation may end earlier than the given turns. To prevent these error cases, it will be helpful to have termination condition of conversations. We leave this for future work. Example 5 only features one skill type, which we already observe a few cases in Figure 4b. While the conversation may make sense, lack of skill blending in the conversation prevents a dialogue model from learning multiple skills. Meanwhile, it is widely known that dialogue agents suffer from generic responses (Li et al., 2016), as shown in Example 6. We observe that while shifts between skill are frequent, responses tend to show little relevance with the skill contexts of their corresponding skills. They often rephrase or repeat what has been discussed during conversation.

Example 1	
Persona 1 I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	**Persona 2** I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones
PERSON1: Hi PERSON2: Hello ! How are you today? PERSON1: I am good thank you, how are you. PERSON2: Great, thanks ! My children and I were just about to watch Game of Thrones. PERSON1: Nice ! How old are your children? PERSON2: I have four that range in age from 10 to 21. You? PERSON1: I do not have children at the moment. PERSON2: That just means you get to keep all the popcorn for yourself. PERSON1: And Cheetos at the moment! PERSON2: Good choice. Do you watch Game of Thrones? PERSON1: No, I do not have much time for TV. PERSON2: I usually spend my time painting: but, I love the show.	
Example 2	
Persona 1 I enjoy going to museums I am married I live in New york city I have ten nieces and nephews I am a writer	**Persona 2** I like to party My major is business I am in college I love the beach I work part time at a pizza restaurant
PERSON1: Hi. How are you doing today? PERSON2: Hi, I am great just finishing up some homework. How are you? PERSON1: I am alright. I just got done writing. PERSON2: Do you write for a living or as a hobby? PERSON1: It is my living. I like culture. PERSON2: That sounds like a fun job. I am a business major but have a part time job. PERSON1: What are you going to school for? PERSON2: I am trying to get my ba in finance. PERSON1: Do you own your own company? PERSON2: No still in school work at pizza hut part time. PERSON1: Do you get free pizza? PERSON2: No but I do get a good deal.	
Example 3	
Persona 1 I work at McDonald's My parents did not want me I was raised by my uncle and aunt I am considering going to college I am looking for something different this year	**Persona 2** I want to be successful I have got two more years in college I have no siblings I am a student I study law
PERSON1: My mom had me in McDonald's bathroom when she was 12. PERSON2: That is really sad. Do you have any siblings? PERSON1: No. My uncle is the manager and took me home. PERSON2: That was nice of him. I am almost done with college. PERSON1: He gave me a job at McDonalds. I was considering college but I am good now. PERSON2: I am almost done. I only have two years left in law school. PERSON1: You be smart. I would like to do something different but I team leader now. PERSON2: That sounds important. Maybe one day you will go back to school. PERSON1: I dropped out to work. Maybe you can help me sure my mom for abandoning me? PERSON2: I do not know if we would be able to do anything because is has been so long. PERSON1: Oh I am only 15 right now. Is that too old? PERSON2: Do not you have to be older than 15 to work at McDonald's? That sounds illegal.	

Table 9: 3 sample conversations from ConvAI2 dataset. In dialogue episodes, each speaker is given five persona sentences respectively.

Example 1	
Topic:	Armadillo
Wizard: Apprentice: Wizard: Apprentice:	I love animals and think armadillos are awesome with their leathery shell. I don't think I've ever seen an armadillo in real life! I've seen them at the zoo. Armadillo means little armored one in Spanish. Are they native to a Spanish-speaking part of the world?
Knowledge:	Armadillos are New World placental mammals in the order Cingulata ... The word "armadillo" means "little armoured one" in Spanish. It is usually made from dairy products, such as milk and cream, and ... ... The nine-banded armadillo ("Dasypus novemcinctus"), or the nine-banded, long-nosed armadillo, is a medium-sized mammal found in North, Central, and South America.
Wizard:	Yes, they are most commonly found in North, Central, and South America
Example 2	
Topic:	Ice cream
Wizard: Apprentice:	I just love ice cream. I love the types with fruits and flavours. Do you like ice cream? I love Ice cream as much as any one. I especially like Gelato, foreign ice cream!
Knowledge:	Ice cream is a sweetened frozen food typically eaten as a snack or dessert ... It is usually made from dairy products, such as milk and cream, and ... ... Bacon ice cream (or bacon-and-egg ice cream) is an ice cream generally created by adding bacon to egg custard and freezing the mixture.
Wizard: Apprentice:	Me too. There are some strange combinations though, have you heard of bacon ice cream? where they add bacon and even egg custard to the freezing mixture! Surprisingly bacon ice cream doesn't surprise me. That doesn't sound appealing to me, but perhaps it could be delicious ...
Example 3	
Topic:	Lifeguard
Apprentice: Wizard: Apprentice:	So I am a lifeguard. Know anything about saving lives in water? I'm impressed! It's a big responsibility to supervise other people's safety in the water! Tell me more. Well, I help make sure people do not drown or get injured while in or near the water!
Knowledge:	A lifeguard is a rescuer who supervises the safety and rescue of swimmers, surfers, ... Lifeguards are strong swimmers and trained in CPR/AED first aid, certified in water ... ... In some areas, the lifeguard service also carries out mountain rescues, or may function as the primary EMS provider.
Wizard: Apprentice: Wizard:	I've heard that in some places, lifeguards also help with other sorts of emergencies, like mountain rescues! Is that part of your job too? I have! I feel like you know much about this! What brings you to know so much? Oh, that's about the extent of my knowledge. I've just been around beaches and I've always admired lifeguards. I'm not a super strong swimmer myself.

Table 10: 3 sample conversations from Wizard of Wikipedia dataset. In each dialogue episode, apprentice is given a topic, while wizard is given the same topic and access to an information retrieval system over Wikipedia. For each utterance, knowledge retrieval is performed based on dialogue history, giving about 61 knowledge candidates per turn.

Example 1	
Emotion:	Terrified (label)
Situation:	My brother jump scared me while I was out playing. It was crazy bad.
Speaker:	Just got scared to death.
Listener:	Oh no. What happened?
Speaker:	My brother jumped scared me.
Listener:	lol is he younger or older?
Example 2	
Emotion:	Proud (label)
Situation:	My little dog learned to sit!
Speaker:	I finally tough my new little puppy his first trick!
Listener:	What trick did you teach him?
Speaker:	I tought him to sit for a treat, it's so cute.
Listener:	That is good, do you plan to teach him more tricks?
Example 3	
Emotion:	Apprehensive (label)
Situation:	I have to call my landlord about being late on the rent. I really don't want to have this conversation.
Speaker:	I have to make a dreadful phone call tomorrow.
Listener:	Oh no, about what?
Speaker:	I'm late on my rent and I need another week. I don't want to because my landlord isnt very nice.
Listener:	Oh no, I've been there done that too many times.
Speaker:	I don't want her to make a big deal.
Example 4	
Emotion:	Content (label)
Situation:	Eating my favorite meal makes me happy.
Speaker:	I am at my best when I have my favorite meal.
Listener:	Nice.
Speaker:	I love enchiladas.
Listener:	Really?
Example 5	
Emotion:	Joyful (label)
Situation:	I have had a great week!
Speaker:	I have had a great start to my week!
Listener:	That's great. Do you think the rest of the week will be as great?
Speaker:	I hope so! It looks promising!!
Listener:	Lucky you. Are you always a positive person or it's just been an amazing week really?
Example 6	
Emotion:	Proud (label)
Situation:	I was proud when my brother finished college. He worked so hard at it.
Speaker:	I was proud of my brother when he finished school. He worked so hard at it.
Listener:	Nice, tell him congrats. What did he major in?
Speaker:	It was English.
Listener:	He should become an English teacher.

Table 11: 6 sample conversations from Empathetic Dialogues dataset. In each dialogue episode, speaker is given a situation with an emotion, while listener is given nothing as input.

Example 1

Skill context from ConvAI2	**Skill context from ConvAI2**
I like to make cookies; I like bagels with everything on; ...	I love the smell of beans; ...
Skill context from WoW	**Skill context from WoW**
Obesity in the United States	Obesity in the United States; Obesity in the ...
Skill context from ED	**Skill context from ED**
My girlfriend told me she's pregnant; Surprised	None

Dialogue
A: Oh wow what kinds of videos do you have? (P)
B: I do mainly make up tutorials. Do you have any side hobbies? (P)
A: I like making cookies and eating cookie dough, I love bagels too. (P)
B: I love bagels as well! What else do you like to do? (P)
A: I also love playing sports, but I'm too busy at this moment. (P)
B: You should try joining a fitness team! It will make you feel better about you being busy. (E)
A: LoL I'll, but I'd feel like I wouldn't be ready. (P)
B: You could always try it out! It's always fun to try out new things! (E)
A: I would if I could I really appreciate new things and learn new things from people like you. (E)
B: I wish you could someday. Thanks for chatting. (E)

Example 2

Skill context from ConvAI2	**Skill context from ConvAI2**
I like to read books; I can play the piano; ...	Halloween is my favorite holiday; ...
Skill context from WoW	**Skill context from WoW**
Vitamin C	Vitamin C; Vitamin C is a vitamin found in food ...
Skill context from ED	**Skill context from ED**
I love walking my dogs on late summer mornings ...; Content	None

Dialogue
A: I'm not nearly as energetic. My dogs take up most of my time unfortunately. (E)
B: Well, at least you take care of them. (E)
A: I don't mind them since they're my dogs. Just wish they were more active. (E)
B: Well, what do you like to do to stay active? (P)
A: I play the piano as my hobby! Are you interested in basketball? (P)
B: Of course. I work a lot so basketball is important. (P)
A: Do you like to read? I love to read about chinese history. (P)
B: I do but mostly I read about halloween on the weekends. (P)
A: That's fun! Halloween is such a fun holiday. (E)
B: Yes I agree. (E)

Example 3

Skill context from ConvAI2	**Skill context from ConvAI2**
I enjoy skiing; My favorite music is rock and roll; ...	I've started working as a freelance designer; ...
Skill context from WoW	**Skill context from WoW**
Pecan pie	Pecan pie; Pecan pie is a pie of pecan nuts mixed ...
Skill context from ED	**Skill context from ED**
I wanted to do something to show my Mom some ... ; Caring	None

Dialogue
A: Amazing to think some places divorce is illegal. I shudder to think what it must be like to be trapped like that. (K)
B: At least in the US the courts can step in and help figure out all the child custody, support, and visitation issues. (K)
A: Yeah, I agree. It's probably hard to make sure kids are all kept by both parents. (E)
B: I'm sure that's certainly a big part of it, too. (E)
A: Do you travel often? I really love skiing, but am considering going to the us for business. (P)
B: I really like Canada. I'm thinking about going there for a business trip. (P)
A: Canada is very beautiful ! Do you like music, too? I really enjoy rock and roll. (P)
B: Yeah, I really love music. It's one of my hobbies. (P)
A: That is good. What do you like to do for fun? (P)
B: I like to draw, and be a good animator. You? (P)

Table 12: Cherry-picked dialogue examples from BS$\mathbb{B}$T dataset.

Example 4	
Skill context from ConvAI2 I love spicy food; I like swimming; I am an only child; ... **Skill context from WoW** Marathon **Skill context from ED** At work sometimes I put a lot of faith ... ; Trusting	**Skill context from ConvAI2** My favorite food is jello; I cannot whistle; ... **Skill context from WoW** Marathon; The marathon is a long-distance running ... **Skill context from ED** None
Dialogue A: Lucky. I've to get things ready for tonights group event. (P) B: What's your event that you're doing? (P) A: Weekly conference where I can talk to people from my field. (E) B: That sounds like a lot of fun. Is this your first one? (E) A: This is the first one of my own. I hope to make it to the next conference. (E) B: Good luck with that. I hope you have fun. (E) A: Thank you. I hope you have fun as well. (E) B: I hope you have a great weeked and good luck. (E) A: You as well. Have a good weekend as well. (K) B: You too. Have a good weekend as well. (K)	
Example 5	
Skill context from ConvAI2 My friends and I watch the games together; We drink beer; ... **Skill context from WoW** Veganism **Skill context from ED** My neighbor's above ground swimming pool ...; Confident	**Skill context from ConvAI2** I like to swim; My favorite drink is chocolate milk; ... **Skill context from WoW** Veganism; Veganism is both the practice of ... **Skill context from ED** None
Dialogue A: I meant actually swimming in the pool. (P) B: I really only swim in natural ponds (NSP) because I have an autoimmune disease. (P) A: Oh wow. Did you eat some bad foods? (P) B: Well I have a nasty bald beard which is a major problem. What do you do? (P) A: I'm an architect. I love the money I get from it. (P) B: What is your favorite dessert? I like tacos. (P) A: I like chocolate so much. What is your favorite food? (P) B: I love all chocolate milk drinks. It is my favorite. (P) A: I love chocolate too. Especially vanilla. You? (P) B: I love all the vanilla flavors. (P)	
Example 6	
Skill context from ConvAI2 I listen to podcasts; I like to read ... **Skill context from WoW** Cake Decorating **Skill context from ED** Been making a lot of money off of my hobby lately ...; Surprised	**Skill context from ConvAI2** I like makeup; many people want to be my friend; ... **Skill context from WoW** Cake Decorating; Cake decorating is one of the ... **Skill context from ED** None
Dialogue A: That sounds like an exciting hobby. (E) B: It is. Super fun and good exercise. What are your hobbies? (P) A: I like to listen to podcasts. My favorite is free piano. (P) B: I play the piano too. And makeup and exercise. I really enjoy those things. (P) A: It's hard to do fun things when you're not having fun. (E) B: Very true. I can be kind to my friends when I am having fun. (P) A: I am sure your friends will be happy to have you as a friend. (K) B: They sure will. They want to be my friend too! (P) A: That is so sweet of you! You have great friends! (K) B: They have a lot of a nice smile so I try to be nice. (P)	

Table 13: Lemon-picked dialogue examples from BS𝔹T dataset.

Example 1	
Skill context from ConvAI2 I like to go for runs; I live by myself; ... **Skill context from WoW** Yoga **Skill context from ED** I felt really guilty when I had to leave my dog ...; Guilty	**Skill context from ConvAI2** I like to listen to music; I have a boxer dog; ... **Skill context from WoW** Yoga; Yoga is a group of practices or disciplines ... **Skill context from ED** None
Dialogue A: Hey, do you have any kids? (P) B: No. I shop in my spare time. Do you? (P) A: Yes, I do. Not the only shopper here. (P) B: Nice! Do you like to swim? I am all about that exercise. (P) A: I love swimming. I wish I could swim more though. (P) B: I wish I could swim too. I am also much older. (E) A: I'm not as young as I used to be. I lost weight then. I am wondering how old you are. (E) B: I am 50 years old. Are you in school? (P) A: Yep. I think its time to start. You don't like to wait long. (P) B: Definitely. That is the life stage for me. I hope I can get back to age. (E)	
Example 2	
Skill context from ConvAI2 I'm going on a cruise next month.; I love to cook.; ... **Skill context from WoW** Cooking **Skill context from ED** I tried a new recipe the other day. It did not ...; Apprehensive	**Skill context from ConvAI2** I like going barefoot.; I can knit.; ... **Skill context from WoW** Cooking; Cooking is the art of preparing food for ... **Skill context from ED** None
Dialogue A: Did you do any sight seeing when you last went to Dallas? (K) B: Not really, we were there for a friends wedding. Mind you, we did go to some great restaurants. With a population of over a million. People the city has a huge number of quality places to go! (K) A: I would have thought the same, but then again we have so many great dishes to eat. Did you enjoy any of them? (P) B: I enjoyed them all! We stayed for more than one night, so I know how to do some great things! (P) A: I am excited because I'm going on a cruise. (E) B: I hope you have a great time! Cruise is such a great destination! (E) A: There are so many different places to go on a cruise. Did you visit any other restaurants near where you were at? (P) B: There were a lot! I saw a restaurant on my cruise tour that I had been to before. (P) A: Was it expensive? Did you have a special restaurant? (P) B: No, it wasn't expensive at all. We had hot dogs and rice, which I enjoyed most of the time. (P)	
Example 3	
Skill context from ConvAI2 I go to high school.; I play ice hockey; ... **Skill context from WoW** Laziness **Skill context from ED** I participated in my high school runners' marathon ...; Proud	**Skill context from ConvAI2** I have a wife and two kids; I am a vegetarian; ... **Skill context from WoW** Laziness; Laziness is disinclination to activity or ... **Skill context from ED** None
Dialogue A: One job would be better then my three. (P) B: What do you like to eat? I'm vegetarian. (P) A: I love beef broth, grilled cheese and eggs. (P) B: Yum! My wife and 2 kids do the same. (P) A: Family is also my life style. Maybe we should have a child! (E) B: But what if you didn't want a child? (E) A: If I didn't want a child I probably wouldn't have them. (E) B: That's true! Our kids would just be us. (E) A: That's what I thought. (E) B: I understand that. (E)	

Table 14: Randomly-sampled dialogue examples from BS$\mathbb{B}$T dataset.

Example 4

Skill context from ConvAI2 I drive an SUV; I've 3 dogs; I have a law degree; ... **Skill context from WoW** Mexico City **Skill context from ED** New apartment, family and friends I love, this is ...; Content	**Skill context from ConvAI2** I've 3 small dogs; I tutor kids in my spare time; ... **Skill context from WoW** Mexico City; Mexico City is the capital and most ... **Skill context from ED** None

Dialogue
A: How modern is your capital city? (K)
B: It's been around a minute, capital and most populous city is Algiers, it's located in the far north of the country. (K)
A: Very cool. How do you feel about it? Do you like your city or your country? (P)
B: I love it and love the people. What do you do for a living? (P)
A: I currently have a B.A. in law. Do you have any pets? (P)
B: I have 3 little ones, and 2 girls. (P)
A: Nice! Will you move? (P)
B: Yes, I would. The cost of living in Algiers is so high that I do not know if I would be able to afford it. (P)
A: It is expensive. Do you ever vacation? I like to visit the mountains. (P)
B: Yes. I visit a lot of places there. (P)

Example 5

Skill context from ConvAI2 My favorite band is Bon Iver; I am from Seattle; ... **Skill context from WoW** Marduk (band) **Skill context from ED** My best friend's husband cheated on her with ...; Disgusted	**Skill context from ConvAI2** I like to attend wine tours; I enjoy visiting museums; ... **Skill context from WoW** Marduk (band); Marduk is a black metal band from ... **Skill context from ED** None

Dialogue
A: I have never heard of a chupacabra. Would you like to live forever? (P)
B: No. However I would like to listen to Oasis' hit "Live Forever" since you mention the idea. (K)
A: That is true, I was hoping I could be of some help. (E)
B: Anything that helps would be great. Thank you very much. (E)
A: You're welcome. There is also a band called CTCR that plays a bit of rap as well. (K)
B: I'll have to give them a listen. (E)
A: There is an album you should check out. It's called the on the court, you should really listen to it. (K)
B: I'll have to take a listen to it. (E)
A: I have to go now, but I hope you can tell them how much you love it and how you felt in this situation. (E)
B: Yes, I will do that as soon as possible. (E)

Example 6

Skill context from ConvAI2 I love coffee with cream; I went to New York city once; ... **Skill context from WoW** Dog **Skill context from ED** We were living in Hawaii ...; Content	**Skill context from ConvAI2** I tutor kids in my spare time; I am very religious; ... **Skill context from WoW** Dog; The domestic dog is a member of "Canis", ... **Skill context from ED** None

Dialogue
A: That is funny. What does jaja mean? (K)
B: Same as ll but in Spanish! (K)
A: Well, that is interesting, I guess. (E)
B: Well it's used for many cultures, Spanish comes after. What we get from it is also used in most western countries. (K)
A: Okay, that is cool. Thank you again, I am learning Spanish also! (P)
B: I'm learning English as well! I'm so happy for you! (P)
A: Thanks, that is interesting. I like learning a foreign language. (P)
B: That's a really cool skill to have. I bet you get to be learning all sorts of different languages. (E)
A: I sure do, because it's an important skill. (E)
B: Did you take it in high school? I know a lot of native Spanish speakers. (P)

Table 15: Randomly-sampled dialogue examples from BS𝔹T dataset.