BotsTalk: Machine-Sourced Framework for Automatic Curation of Large-scale Multi-skill Dialogue Datasets

Anonymous ACL submission

Abstract

Previous work in open-domain chatbots has introduced dialogue corpora and tasks that aim to inject dialogue systems different communicative skills such as being personable, knowledgeable and empathetic. With the advent of conversational agents grounded to specific skills, a new challenge in open-domain chatbots has been posed: A good open-domain chatbot should retain a well-rounded set of skills and seamlessly blend them into a conversation. To this end, a new dialogue dataset Blended Skill Talk is collected via crowdsourcing and commonly used as a benchmark for multi-skill dialogue generation. However, such data construction approach requires labor intensive manual annotation, which severely limits their utility on large-scale learning. In this work, we propose BotsTalk, a novel machine-sourced framework, where several agents participate in a conversation to automatically annotate multi-skill dialogues. We then present Blended Skill BotsTalk (BS𝔹T), a large-scale multi-skill dialogue dataset of 200K conversations. Experimental results show that our dataset can be effectively used as training data for multi-skill dialogue systems which require an understanding of both skill blending and grounding. We also demonstrate the dataset is orthogonally applicable to diverse learning schemes such as fine-tuning and multi-task learning.

1 Introduction

A considerable progress has been made towards open-domain chatbots with different desirable qualities in conversation. 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 (Zhang et al., 2018) is a dataset provided for research work (Kim et al., 2020b; Majumder et al., 2020a; Madotto et al., 2019) that aims to endow chatbots with personas, enabling them to talk about themselves. Wizard of Wikipedia (WoW) (Dinan et al., 2019) is a popular option for recent studies (Kim et al., 2020a; Zhao et al., 2020; Lian et al., 2019) that focus on knowledgeable conversational agents discussing topics in depth. Empathetic Dialogues (ED) (Rashkin et al., 2019) is also commonly used by recent studies (Majumder et al., 2020b; Santhanam and Shaikh, 2019) to embody empathy in dialogue systems. 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 displays, for example, listening with empathy, providing knowledgeable responses, and talking about various topics from everyday life within a conversation (Roller et al., 2020a; Smith et al., 2020). It should be able to adapt to the possibilities of different users and situations and thus use different communicative skills appropriately in a dialogue.

Towards this goal, there is a need to construct a multi-skill dialogue dataset, which comprises 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 (Lee et al., 2021).

In this paper, we propose an automatic data curation approach that repurposes conversational agents with individual skills for generating a large-scale multi-skill dialogue dataset without additional costs or human efforts. Our main contributions are summarized as follows.

- **BotsTalk**, a machine-sourced framework where multiple dialogue agents grounded to individual skills engage in a conversation that blends all skills together.
- **Blended Skill BotsTalk (BS𝔹T)**, a large-scale multi-skill dialogue dataset which contains 200K conversations blended and grounded with a number of skills derived from ConvAI2, WoW, and ED.
- Analysis and evaluation results show that our dataset can be effectively used as training resource for multi-skill dialogue systems which require an understanding of not only skill grounding but also skill blending.

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. Zhang et al. (2018) introduce ConvAI2 dataset which consists of more than 140K utterances of crowdsourced conversations to make chit-chat models more engaging and personalized by conditioning profile information on the models. 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) constructed a dataset, Empathetic Dialogues, comprising 50K utterances of crowdworker conversations grounded in an emotional situation in order to enable 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, there is a clear trend in the research of open-domain chatbots, that single-skill conversation is moving to well-grounded multi-skill conversation (Smith et al., 2020; Shuster et al., 2020; Roller et al., 2020b). In particular, Smith et al. (2020) aims to build a conversational agent who seamlessly blends being engaging and personable (Zhang et al., 2018), knowledgeable (Dinan et al., 2019), and empathetic (Rashkin et al., 2019). 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

Dialogue systems research has been consistently supported by the development of new datasets (Williams et al., 2014; Mrkšić et al., 2017; Budzianowski et al., 2018). One popular approach is to collect and annotate dialogues via crowdsourcing (Zhang et al., 2018; Dinan et al., 2019; Rashkin et al., 2019; Smith et al., 2020). However, generating multi-turn dialogues in this manner requires expensive and exhausting human efforts (Shah et al., 2018; Lee et al., 2021).

Therefore, recent research seeks to facilitate open-domain chatbot development with new datasets automatically constructed by utilizing existing datasets. For example, Lee et al. (2021) create a 45K multi-modal dialogue dataset, starting with existing text-only dialogue datasets as source dialogues, and then replacing part of sentences in source dialogues with their semantically relevant images. Yang et al. (2021) propose leveraging both image-context-response triples and large scale of textual conversations for image-grounded response generation. 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. (2020) present a model-based dialogue collection framework and a human-and-model-in-the-loop process for generating datasets respectively.

Motivated by this line of research, in this work, we explore how large-scale multi-skill dialogue datasets can be automatically collected with minimal human efforts for data annotation.

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 contexts; Wizard of Wikipedia provides a topic and knowledge resouces as skill contexts; 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.

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 can be defined as seamlessly blending various skills over the course of a multi-turn conversation. Here, inspired by Smith et al. (2020) and Sun et al. (2021), the inputs of this task are single-skill datasets, which are separately collected on a variety of skills. Let $\mathbb{K}$ be the set of K skill types, *e.g.*, $\mathbb{K} = \{\mathrm{P}, \mathrm{K}, \mathrm{E}\}$ where P, K, E denote personality, knowledge, and empathy derived from ConvAI2, WoW, and ED, respectively. Formally, we refer to $\mathcal{D}_k$ as a dialogue dataset with N_k dialogue episodes for skill $k \in \mathbb{K}$.

$$\mathcal{D}_k = \{(stx_{i,k}, dtx_{i,t})\}_{i=1}^{N_k} \tag{1}$$

where $stx_{i,k}$ is a skill-relevant description (*i.e.*, skill context) for skill k and $dtx_{i,t}$ is t dialogue turns (*i.e.*, dialogue context) derived from the skill context, as shown in Table 1.

Based on input datasets $\mathcal{D}_1, ..., \mathcal{D}_k$, we aim to obtain a new dialogue dataset $\tilde{\mathcal{D}}$ for K 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 $\mathbb{K}$ and $dtx_{i,t}$ is the dialogue context derived from the multiple skills. Table 3 shows a dialogue example in output dataset $\tilde{\mathcal{D}}$. 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 dialogue system 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 dataset should enable dialogue agents to exhibit different dialogue skills in a conversation flow (Smith et al., 2020; Madotto et al., 2021), while **skill grounding** emphasizes that dialogue agents should learn to maintain the same dialogue skill when appropriate (Shazeer et al., 2017). We argue that they have a trade-off relationship as it is often difficult to sufficiently represent both skill blending and grounding in a dialogue of finite length. Therefore, a desirable multi-skill conversation should be composed of short dialogue sessions specific to different skills. We note that skill grounding and blending are not contradictory, as some skill-grounded utterances imply a natural shift in skills. As an example, suppose that given an utterance "*I like sneakers because it is comfortable.*" which demonstrates 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 shifted and blended naturally so that the conversational agents learn to provide reasonable responses in a multi-skill dialogue (Roller et al., 2020a).

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.

4.1 Participants in BotsTalk

In our framework, multiple participants engage in a conversation to iteratively generate the most appropriate response.

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

$$f(stx_k, dtx_t; \theta^k) \triangleq \operatorname*{argmax}_{u \in \mathbb{U}} P(u|stx_k, dtx_t; \theta^k) \qquad (3)$$

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

To determine response space $\mathbb{U}$, we design the two main functions of the dialogue agents, generator model and ranker model, parameterized as θ^k_{gen} and θ^k_{rnk} for skill k, respectively. For θ_{gen}, we aim to generate responses from response space $\mathbb{U}$ in a token-by-token manner, and thus employ the dodecaDialogue (Shuster et al., 2020) model which is a modification of a transformer Seq2Seq architecture. On the other hand, for $\theta = \theta_{rnk}$, we consider the response space $\mathbb{U}$ as a list of alternatives to pick the correct response, and thus employ a transformer-based retrieval architecture to score and rank response candidates in the finite set. Specifically, we use a 256-million parameter poly-encoder (Humeau et al., 2020) pre-trained on the pushshift.io Reddit dataset. Both θ_{gen} and θ_{rnk} are fine-tuned on individual single-skill datasets.

While all skill agents would simulate what response is the most appropriate conditioned on skill context set $\tilde{stx}$ and the current dialogue context dtx_t, only one skill agent is given priority over other skill agents, to "speak" the response per dialogue turn for the dialogue annotation. 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 k can learn to read other skill contexts in $\tilde{stx}$ for different skills. For a single-skill dialogue agent, considering all possible skill contexts in a multi-skill dialogue 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 skill blending and grounding. Suppose that, given an arbitrary dialogue context dtx_t, a skill agent returns a response $res_{k,t} = f(stx_k, dtx_t; \theta^k)$. Based on a set $\tilde{stx}$ of all skill contexts and action space $\mathbb{A}$ (*i.e.*, approval or refusal), a moderator agent uses the decision functions $g : (\tilde{stx}, dtx_t, res_{k,t}) \mapsto \mathbb{A}$ to examine the relevance of the response with the contexts.

4.2 Phase 1: Simulate what to speak

We integrate different dialogue setups from multiple single-skill datasets as a seed information to start a conversation. Specifically, for a dialogue episode, dialogue context is initialized as an utterance pair (*i.e.*, two-turn dialogue) via random sampling with a single-skill dataset $\mathcal{D}_k$, and the skill agent for skill k becomes the initial active agent. Then, for a generalizable dialogue setup, we retrieve the most relevant skill contexts from each of all input datasets $\mathcal{D}_1, ..., \mathcal{D}_K$ by querying the seed dialogue context with a widely used IR system.[1]

In the first phase of BotsTalk, all skill agents simulate their own responses for the next dialogue turn. Formally, given a skill context set $\tilde{stx} = \{stx_1, ..., stx_K\}$ and the current dialogue context dtx_t in a dialogue episode, a skill agent for skill k generates a plausible response $res_{k,t}$ as:

$$res_{k,t} = f(stx_k, dtx_t; \theta^k_{gen}) \qquad (4)$$

where stx_k is a skill context for skill k in $\tilde{stx}$.

Depending on individual skills, every skill agent returns its skill-relevant response. For example, when an dtx "*I love sneakers and think they are the most comfortable shoes around.*" is given, the skill agent for skill P generates a response "*Oh really? I like tennis shoes more than sneakers.*" as res_{P}, which personalizes the dialogue agent by grounding the response to a given persona. Meanwhile, the skill agents for skill K and E generate a knowledgeable response "*It is because sneakers*

[1] We use the implementation of Chen et al. (2017).

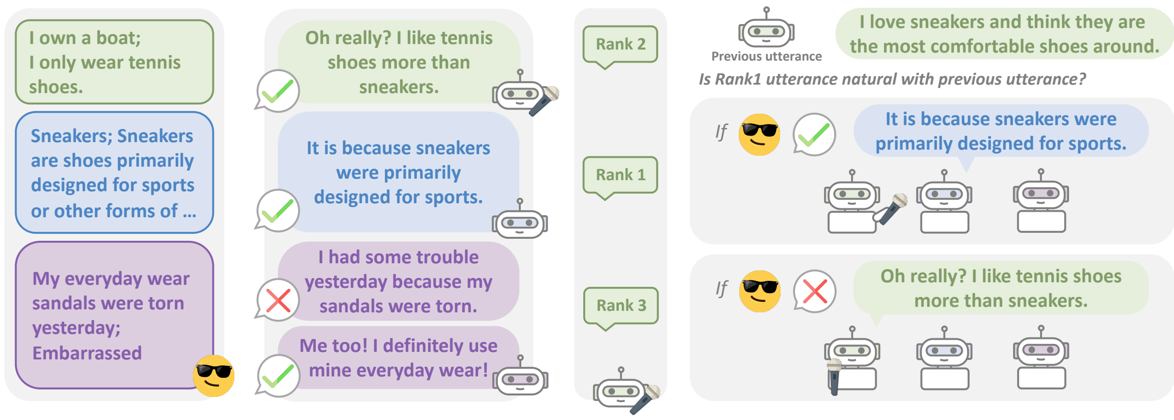


Figure 1: Illustration of BotsTalk framework. Green, blue, and purple indicate skill types of P, K, E.

were primarily designed for sports." as res_{K} and a empathetic response "*Me too! I definitely use mine for everyday wear!*" as res_{E}, respectively.

Note that a skill agent uses the specific skill context stx_k instead of $\tilde{stx}$ for response generation. We observe that skill agents suffer from inconsistent and generic responses when conditioned on $\tilde{stx}$, as they refer less to the dialogue context dtx_t and more to skill contexts that they are not trained on. Since a skill agent aims to generate a skill-grounded response, it should take its own skill context as an input.

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 we consider different skills together in a conversation, the response generated by a skill agent is more likely to be semantically in conflict with other skill contexts in $\tilde{stx}$. Therefore, the moderator agent, who has access to all skill contexts $\tilde{stx}$, is designed to maintain dialogue consistency by filtering out conflicting response candidates. A skill agent repeatedly generates new response candidates until its response $res_{k,t}$ is not contradictory to $\tilde{stx}$ anymore.

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: Entailment (true), Contradiction (false), and Neutral (undetermined). A sound response $res_{k,t}$ should not be contradictory to all skill contexts $\tilde{stx}$. 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 removed because "*yesterday because my sandals were torn*" is contradictory to "*I wear sneakers everyday*".

As the moderator agent, we use a RoBERTa model (Liu et al., 2019) trained on MNLI (Williams et al., 2018)[2], which is widely used in fact checking systems (Kim et al., 2021a). The RoBERTa model shows 90.59% accuracy on MNLI validation set. Overall, about 50% of utterances are classified as Contradiction by NLI classifier. The result demonstrates the skill agents indeed generate inconsistent responses due to the restricted access to other skill contexts. Figure 2 breaks down the result by the types of skill contexts (*i.e.* P, K, E). Out of all utterances classified as Contradiction, about 70% are in conflict with other types of skill contexts. We also find that the overall proportion of utterances conflicting with stx_{P} is relatively high. This tendency results from the difference between skill contexts, *e.g.*, stx_{P} contains more distinct descriptions than stx_{K} and stx_{E}, and thus is more likely to contradict the utterance. The moderator agent filters out such contradictory candidates to preserve dialogue consistency.

4.4 Phase 3: Speak or pass the mic

Given the dialogue context and the skill contexts, the objective of the last phase is to score a set of response candidates and select a final response. 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 a set of response candidates $res_{1,t}, ..., res_{k,t}$ 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^k_{rnk}, then asks to attach the selected response into the next

[2] Dialogue NLI (Welleck et al., 2019) is biased to ConvAI.

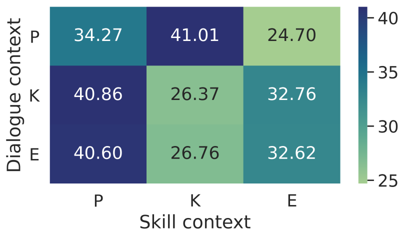
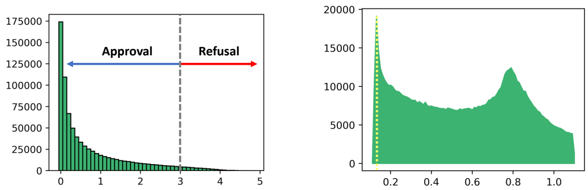


Figure 2: Percentages of utterances which are classified as CONTRADICTION via NLI classifier.



(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).

dialogue context dtx_{t+1} for annotation. Formally, we define such a process as:

$$res_t^* = \underset{res_t \in \mathbb{U}_{res}}{\text{argmax}}\ P(res_t | stx_k, dtx_t; \theta^k_{rnk}) \cdot g(dtx_t, res_t) \tag{5}$$

where $g(dtx_t, res_t) \in \{0, 1\}$ is a function of the moderator agent which determines whether res_t is approved or not.

For computing $g(dtx_t, res_t)$, the moderator agent adopts a skill classifier $\mathcal{P}$ of identifying the corresponding skill for a response. We use a BERT (Devlin et al., 2019) model trained on utterances in $\mathcal{D}_k$ and their corresponding label k for any skill k. Once $\mathcal{P}$ is learned, we compute $g(dtx_t, res_t)$ as:

$$g(dtx_t, res_t) = \begin{cases} 1, & \text{KL}(\mathcal{P}(res^*_{t-1}) || \mathcal{P}(res_t)) < \alpha \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where res^*_{t-1} is the last utterance of dtx_t and $\mathcal{P}(\cdot) \in \mathbb{R}^K$ outputs a skill distribution of an utterance/response. As the difference between two distributions (*i.e.*, KL divergence) is larger, $g(dtx_t, res_t)$ is likely to have a smaller value, which is further discretized as the approval/refusal decision with a pre-defined threshold α (Figure 3a). Once the moderator agent accepts the candidate res_t^* from an inactive agent as the response, the active agent passes the mic, or the priority for annotation, to the inactive agent.

In practice, we compute entropy of all utterances based on their skill distributions in order to investigate whether there is a room for shifting between skills. The value of entropy indicates the uncertainty of the skill type of an utterance, *e.g.*, the value of entropy becomes lower when the utterance is grounded to the specific skill and higher when the utterance is generic. As shown in Figure 3b, not all utterances are grounded to specific skill. It is evident that there are a significant amount of opportunities to shift to other skills when the entropy value of an utterance is high. This result suggests that while skill grounding, skill blending is also available in natural and reasonable way.

	Dataset	Train	Valid	Test
# Dialogues	ConvAI2	17,878	1,000	1,015
	WoW	18,430	1,948	1,933
	ED	19,533	2,770	2,547
	BST	4,819	1,009	980
	BS$\mathbb{B}$T	200,000	-	-
# Utterances	ConvAI2	131,438	7,801	6,634
	WoW	166,787	17,715	17,497
	ED	40,252	5,736	5,257
	BST	41,475	8,678	8,422
	BS$\mathbb{B}$T	2,000,000	-	-

Table 2: Statistics of dialogue datasets.

5 Blended Skill BotsTalk (BS$\mathbb{B}$T)

5.1 Data Statistics

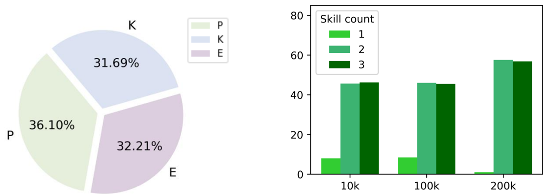
We collect Blended Skill BotsTalk (BS$\mathbb{B}$T), a multi-skill dialogue dataset, using BOTSTALK framework. The dataset consists of 200K conversations with 2M utterances, and there are 10 utterances in each conversation. Each utterance is labeled using a skill classifier with skill annotation, which is personality from ConvAI2, knowledge from WoW or empathy from ED, including both hard label (*i.e.* skill type) and soft label (*i.e.* skill distribution). An example from BS$\mathbb{B}$T is shown in Table 3.

One of the salient features of BS$\mathbb{B}$T dataset is its scalability. The overall statistics of datasets used in this paper are summarized in Table 2. The BS$\mathbb{B}$T dataset overwhelms other datasets with its size, since it is composed of bot-bot conversations collected through a machine-sourced approach while other datasets comprise crowdsourced human-to-human conversations. Theoretically, the size of the BS$\mathbb{B}$T dataset can increase infinitely as long as seed information is provided. We ran our framework on 20 Titan RTX GPUs for 4 days to collect 200K conversations, and the data construction of 2M conversations is on progress.

Skill context from ConvAI2	**Skill context from ConvAI2**
I attend book club every week; My mom is my best friend; ...	I love riding roller coasters; I am a factory worker; ...
Skill context from WoW	**Skill context from WoW**
Reading process	Reading is a complex cognitive process ...
Skill context from ED	**Skill context from ED**
I saw Stephen King on his book tour last ... & Disappointed	None

Dialogue
A: Wow that's quite a few. That's cool. (K)
B: It's so good. It won a Pulitzer Prize. (K)
A: Wow, what made you read it? What are you into now? (P)
B: It was the oldest work of literature in history, because it was written in broad enough mediums. (K)
A: Yeah, well you are probably lucky because I love reading too. (P)
B: Yes, it is very handy to keep reading and to enjoy the parts that you like. (P)
A: I want to read it. Hopefully I'll get a chance to see it some day. (P)
B: I hope you will as well. Good luck! (E)

Table 3: Sample 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, seven knowledge resources (top right). In the dialogue, P, K, E denotes the skill types corresponding to ConvAI2, WoW, and ED, respectively.



(a) Skill percentage (b) Skill count

Figure 4: Illustration of skill annotation: skill percentage (left) for all utterances and the skill count per dialogue by varying the number of dialogues (right).

	Skill of seed utterances		
% annotated as:	P	K	E
P	**59.79**	13.76	14.89
K	18.04	**72.89**	10.60
E	22.15	13.33	**74.50**

Table 4: Percentages of utterances followed by seed utterances annotated by the skill classifier as coming from P, K, E, broken down by provenance skills of the seed utterances.

Skill Blending Figure 4 summarizes the results of skill annotation for all utterances in BS𝔹T dataset. Overall, the skill annotation percentages are 36.10% for personality, 31.68% for knowledge, and 32.21% for empathy, as shown in Figure 4a. Figure 4b also shows that over 90% of the conversations demonstrate at least 2 of the 3 skills within a single conversation. This supports that the vast majority of conversations feature more than one skill, where skills of utterances are defined based on skill distribution. In addition, the tendency of skill blending in Figure 4b is stable at varying data size (10K, 100K, 200K), suggesting the efficacy of BotsTalk on multi-skill dialogue generation.

Skill Grounding Although we focus on blending skills, the dataset should contain sufficient sessions grounded to specific skill in conversations for the model to learn the ability of skill grounding. Since the provenance skill of an utterance from original dataset is only available for seed utterances, we explore the continuity of skills based on the skill type of the utterance subsequent to seed utterances. Table 4 breaks down the results by provenance skill of the seed utterances. The fraction of utterances resembling a given dataset increases when the seed utterances are from that same dataset, and more than half of the utterances subsequent to the seed utterances are labeled the same skill type as the seed utterances.

Data Quality We perform ACUTE-Eval (Li et al., 2019), which is a popular metric for human evaluation, on BST and BS𝔹T datasets. Evaluation results show that BS𝔹T achieves slightly higher scores than BST, indicating the validity of BotsTalk framework for collecting multi-skill dialogues. We provide more details in Appendix E.

5.2 Evaluation on Multi-skill Benchmark

We conduct a set of experiments to test our BS𝔹T over BST benchmark. 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. The base architecture used throughout the experiments is a 256-million parameter poly-encoder (Humeau et al., 2020) pre-trained on pushshift.io Reddit dataset. We fine-tune the base architecture on individual datasets, *i.e.*, ConvAI2, WoW, ED, BST, and BS𝔹T, and consider them as our baselines.

Model	R@1	R@5	R@10	MRR
ConvAI2	75.92	94.04	97.19	83.96
WoW	67.48	89.57	94.33	77.11
ED	65.96	88.69	93.80	76.10
BST	75.92	94.76	97.83	84.14
BS𝔹T	80.12	95.53	97.88	86.94

Table 5: Evaluation on BST benchmark.

	Single-skill benchmarks			
Model	ConvAI2	WoW	ED	Avg.
ConvAI2	**88.46**	79.84	47.90	72.06
WoW	57.90	**90.79**	45.86	64.85
ED	61.31	80.05	**62.81**	68.05
BST	74.13	82.12	48.11	68.12
BS𝔹T	84.12	89.68	60.85	**78.21**

Table 6: Evaluation on single-skill benchmarks. Numbers in bold indicate the best performing model on the corresponding benchmark.

	Benchmarks			
Model	BST	ConvAI2	WoW	ED
MTL	78.97	86.43	90.41	59.76
MTL + BS𝔹T50K	80.88	86.49	90.48	59.80
MTL + BS𝔹T100K	80.94	86.71	90.63	59.92
MTL + BS𝔹T200K	82.01	86.82	91.36	61.37

Table 7: Evaluation on benchmarks of MTL models.

The results of baselines on BST benchmark are shown in Table 5. For metrics, we measure Recall@k, or R@k, where each test example has 100 possible candidates to select from, as well as mean reciprocal rank (MRR). It is observed that multi-skill models, *i.e.*, BST and BS𝔹T models, consistently outperform single-skill models, *i.e.*, ConvAI2, WoW, and ED models. This suggests that the single-skill models are able to do well on each of them in isolation, but struggle to seamlessly blend them over the course of a single conversation. Moreover, BS𝔹T model outperforms all of the baselines on all automatic evaluation. This result indicates that our dataset properly works as the training resource to learn the ability to blend skills. We also provide performance of BS𝔹T model by varying the number of dialogues on BST benchmark for scalability analysis in Appendix G.

5.3 Evaluation on Single-skill Benchmark

Table 6 summarises the results of baselines on single-skill benchmarks, *i.e.*, ConvAI2, WoW, and ED benchmarks, measured by R@1. The single-skill models each perform the best on their respective original benchmark and not as well on other benchmarks, compared to the multi-skill models, supporting our hypothesis that single-skill agents are specialized to their corresponding skills. On the other hand, the performance of all multi-skill models is more balanced than single-skill models, in the sense that none of the single-skill models does as well averaged over the three categories (except for the ConvAI2 model doing a tiny bit better than the BST model). In particular, BS𝔹T model performs noticeably better on all single-skill benchmarks than BST model. This suggests that BS𝔹T is able to not only inject the ability of blending various skills but also maintain the ability for grounding specific skill.

5.4 Evaluation with Multi-task Learning

A straightforward approach of developing a multi-skill chatbot given access to multiple single-skill datasets is to multi-task on all of them during the fine-tuning step. Therefore, we consider MTL model, the poly-encoder pre-trained on pushshift.io Reddit and fine-tuned in multi-task fashion across ConvAI2, WoW, and ED. To probe the effectiveness of BS𝔹T as training resource, we further fine-tune MTL model on BS𝔹T datasets with different sizes (50K, 100K, 200K), respectively. Table 7 compares the performance of these models on all benchmarks, *i.e.*, ConvAI2, WoW, ED, and BST, reported by R@1 and MRR. As expected, MTL models fine-tuned on BS𝔹T datasets with varying scales outperform MTL model for all benchmarks, indicating that BS𝔹T is orthogonally applicable to MTL. The overall tendency also shows that the model performs better when we fine-tune the MTL model with a large scale of BS𝔹T. Such results suggest that the scalability of BS𝔹T is indeed crucial to model performance.

6 Conclusion

We present a novel machine-sourced approach BotsTalk for generating multi-skill dialogues. We further propose a large-scale multi-skill dialogue dataset BS𝔹T consisting of 200K conversations to inject a dialogue system the ability of skill blending and grounding. We demonstrate the effectiveness of our approach in comparison with several baselines by experiments on both single- and multi-skill dialogue benchmarks. Despite the inherent errors stemmed from its machine-sourced nature, our proposed data creation method can be applied when efficiently preparing datasets that cover diverse communicative skills.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP*.

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. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of EMNLP*.

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*.

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. In *arXiv preprint*.

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

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of ACL*.

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*.

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. 2020a. Open-domain conversational agents: Current progress, open problems, and future directions. In *arXiv preprint*.

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. 2020b. Recipes for building an open-domain chatbot. In *arXiv preprint*.

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. In *arXiv preprint*.

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 *arXiv preprint*.

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. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *arXiv preprint*.

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.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of AAAI*.

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 provide more details on BotsTalk framework and BS𝔹T dataset. Specifically, we lay out the details of single-skill dialogue datasets and how they are incorporated into BotsTalk framework to construct BS𝔹T dataset in Appendix B. We also provide hyperparameter setups for all component models of skill agents and the moderator agent in Appendix C. Appendix D presents conversation examples of BS𝔹T. In Appendix E, we provide human evaluation results of BS𝔹T to support the validity of BotsTalk framework. In Appendix F and G, we expand the experimental results on single-skill benchmark presented in Section 5.3 and conduct the scalability analysis of BS𝔹T.

B Single-skill Datasets into BotsTalk

We describe the details on the single-skill dialogue datasets used for BS𝔹T and 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 10, 11, 12.

To integrate dialogue setups from different single-skill datasets as a seed information, we follow the basic settings for constructing a dialogue dataset and assume a multi-turn, one-to-one conversation between two speakers. We simulate turn-taking in a conversation by switching between two different sets of skill contexts for the input skill context stx to a dialogue model f in skill agents. We present details on the single-skill datasets used to construct BS𝔹T and elaborate on how the seed information is constructed based on them.

B.1 ConvAI2

Based on PersonaChat (Zhang et al., 2018), ConvAI2 is a dataset 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 ConvAI2 setting, we provide two different profiles composed of 5 different persona sentences as the skill contexts stx_{P}. The personas are retrieved from the ConvAI2 dataset based on their relevance to the seed utterances, measured by a TF-IDF retriever. The two profiles are used as the input stx_{P} to the dialogue model f in an alternating manner.

B.2 Wizard of Wikipedia

The Wizard of Wikipedia task involves discussing a given topic in depth, where the goal is to both engage the partner as well as display expert knowledge (Dinan et al., 2019). 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, while 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 the simpler version of the task and ignore the retrieval aspect of the task. We specify the topic of the conversation as the seed information of the skill agent. The skill context stx_{K} of the Apprentice is thus defined as the given topic, while stx_{K} of the Wizard is defined as a topic and 7 relevant knowledge sources.

B.3 Empathetic Dialogues

The Empathetic Dialogues (Rashkin et al., 2019) dataset includes 50K utterances of crowdworker conversations grounded in an emotional situation. In a conversation, one speaker describes a personal situation based on an emotion label and the other speaker, named the listener, displays empathy in their response. Specifically, a pair of workers 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 retrieve the situation description and its corresponding emotion label from the Empathetic Dialogue dataset. They are used to define the skill context stx_{E} of the speaker in an ED setting. Note that we do not define 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

Our implementation is based on the ParlAI toolkit,[3] which is specialized in training and evaluating dialogue systems. We will release our agents and dataset for public use.

C.1 Skill Agent

In our 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. For generative models, at inference time, one must choose a decoding method to generate a response to the dialogue context. In this work, we use nucleus sampling as a decoding strategy.

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). Poly-encoders encode 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 per-

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

Model	R@1	R@5	R@10	MRR
Evaluation on ConvAI2 benchmark				
ConvAI2 (Zhang et al., 2018)	88.46	98.92	99.71	93.03
WoW (Dinan et al., 2019)	57.90	86.85	95.80	70.59
ED (Rashkin et al., 2019)	61.31	89.44	96.69	73.53
BST (Smith et al., 2020)	74.13	95.64	98.80	83.37
BS$\mathbb{B}$T (Ours)	84.12	97.30	99.22	89.91
Evaluation on WoW benchmark				
ConvAI2 (Zhang et al., 2018)	79.84	96.97	98.84	87.62
WoW (Dinan et al., 2019)	90.79	99.28	99.66	94.67
ED (Rashkin et al., 2019)	80.05	96.25	98.37	87.34
BST (Smith et al., 2020)	82.12	97.57	98.99	89.11
BS$\mathbb{B}$T (Ours)	89.68	99.38	99.77	94.09
Evaluation on ED benchmark				
ConvAI2 (Zhang et al., 2018)	47.90	76.14	85.87	60.60
WoW (Dinan et al., 2019)	45.86	74.79	85.15	58.94
ED (Rashkin et al., 2019)	62.81	88.91	94.58	74.18
BST (Smith et al., 2020)	48.11	77.09	86.96	61.04
BS$\mathbb{B}$T (Ours)	60.85	87.74	94.09	72.67

Table 8: Evaluation on single-skill benchmarks, *i.e.*, ConvAI2, WoW, and ED benchmarks.

	BST (Win %)	BS$\mathbb{B}$T (Win %)
Engagingness	43	57
Interestingness	47	53
Humanness	44	56

Table 9: Human evaluation for pairwise comparison between BST and BS$\mathbb{B}$T datasets.

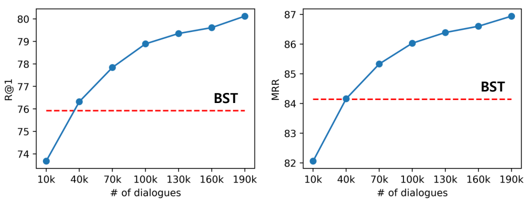


Figure 5: The effect on performance by varying the number of dialogues, reported by R@1 (left) and MRR (right).

formance 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. The poly-encoder is based on large pre-trained transformer models with the same architecture and dimension as BERT-base (Devlin et al., 2019), 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.

C.2 Moderator Agent

In our 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 implementation[4] 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 hypothesis sentence and each skill context $stx_k \in \tilde{stx}$ as a premise sentence, then conduct unidirectional NLI 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 skill k as labels. The model was trained

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

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.

D Dataset Examples

We present a number of dialogue examples randomly sampled from BS$\mathbb{B}$T in Table 13, 14, 15, 16. In each dialogue episode in 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.

E Human Evaluation

We perform human evaluation on BS$\mathbb{B}$T dataset to validate our BotsTalk framework. To this end, we employ ACUTE-Eval (Li et al., 2019), which is a popular metric for multi-turn dialogue evaluation (Dinan et al., 2020; Li et al., 2020). We randomly sample 100 dialogues from the BST and BS$\mathbb{B}$T datasets, respectively. We then ask human evaluators to compare each pair of dialogues from BST and BS$\mathbb{B}$T datasets over three axes: engagingness, interestingness and humanness. We provide the evaluators with three questions to assess the quality of the dialogues:

- **Engaging:** Who would you prefer to talk to? Which version is more likely to hold your attention and make you want to hear more?

- **Interesting:** Who would you say is more interesting? Which version arouses your curiosity or tells you something new or useful?

- **Humanlike:** Who would you say sounds more human? Which version is more natural and personable?

The evaluation results are shown in Table 9. For all ACUTE-Eval metrics, BS$\mathbb{B}$T dataset achieves slightly higher win percentages over BST dataset, although the difference between datasets is not statistically significant. Thus, our machine-sourced approach BotsTalk can be an useful alternative to crowdsourcing when collecting multi-skill dialogues.

F Additional Evaluation on Single-skill Benchmarks

We provide the experimental results on single-skill benchmarks, *i.e.*, ConvAI2, WoW, and ED benchmarks. We report R@1, R@5, R@10, and MRR for metrics in Table 8, which elaborates the results in Table 6.

G Scalability Analysis

To gain more insights into the scalability of our approach, we construct datasets at varying scales (10K, 40K, 70K, 100K, 130K, 160K, 190K) and fine-tune the base architecture on each of them. We evaluate these models on BST benchmark to investigate the impact of BS$\mathbb{B}$T size to the model performance. Figure 5 illustrates the performance of BS$\mathbb{B}$T model in terms of R@1 when the training size of BS$\mathbb{B}$T dataset gradually increases. The model fine-tuned on BS$\mathbb{B}$T190K dataset achieves the best performance. The significant performance boost from the 10K to 190K models reaffirms the importance of large-scale training. Considering our data curation process does not involve human intervention (*i.e.*, no manual annotation and verification), this is a promising result showing the potential of our dataset for large-scale supervised learning: one can obtain datasets of arbitrary sizes and develop multi-skill dialogue systems by leveraging high portion of multi-skill conversations provided in datasets.

<table>
<tr><td colspan="2">Example 1</td></tr>
<tr><td>Persona 1</td><td>Persona 2</td></tr>
<tr><td>I like to ski</td><td>I am an artist</td></tr>
<tr><td>My wife does not like me anymore</td><td>I have four children</td></tr>
<tr><td>I have went to Mexico 4 times this year</td><td>I recently got a cat</td></tr>
<tr><td>I hate Mexican food</td><td>I enjoy walking for exercise</td></tr>
<tr><td>I like to eat cheetos</td><td>I love watching Game of Thrones</td></tr>
<tr><td colspan="2">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.</td></tr>
<tr><td colspan="2">Example 2</td></tr>
<tr><td>Persona 1</td><td>Persona 2</td></tr>
<tr><td>I enjoy going to museums</td><td>I like to party</td></tr>
<tr><td>I am married</td><td>My major is business</td></tr>
<tr><td>I live in New york city</td><td>I am in college</td></tr>
<tr><td>I have ten nieces and nephews</td><td>I love the beach</td></tr>
<tr><td>I am a writer</td><td>i work part time at a pizza restaurant</td></tr>
<tr><td colspan="2">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.</td></tr>
<tr><td colspan="2">Example 3</td></tr>
<tr><td>Persona 1</td><td>Persona 2</td></tr>
<tr><td>I work at McDonald's</td><td>I want to be successful</td></tr>
<tr><td>My parents did not want me</td><td>I have got two more years in college</td></tr>
<tr><td>I was raised by my uncle and aunt</td><td>I have no siblings</td></tr>
<tr><td>I am considering going to college</td><td>I am a student</td></tr>
<tr><td>I am looking for something different this year</td><td>I study law</td></tr>
<tr><td colspan="2">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.</td></tr>
</table>

Table 10: 3 sample conversations from ConvAI2 dataset. In dialogue episodes, each speaker is given five persona sentences respectively.

Example 1	
Topic:	Ice cream
Wizard:	I just love ice cream. I love the types with fruits and flavours. Do you like ice cream?
Apprentice:	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:	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!
Apprentice:	Surprisingly bacon ice cream doesn't surprise me. That doesn't sound appealing to me, but perhaps it could be delicious ...
Example 2	
Topic:	Armadillo
Wizard:	I love animals and think armadillos are awesome with their leathery shell.
Apprentice:	I don't think I've ever seen an armadillo in real life!
Wizard:	I've seen them at the zoo. Armadillo means little armored one in Spanish.
Apprentice:	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 3	
Topic:	Lifeguard
Apprentice:	So I am a lifeguard. Know anything about saving lives in water?
Wizard:	I'm impressed! It's a big responsibility to supervise other people's safety in the water! Tell me more.
Apprentice:	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:	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?
Apprentice:	I have! I feel like you know much about this! What brings you to know so much?
Wizard:	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 11: 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 12: 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 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? B: Hi! no, i shop in my spare time. do you? A: Yes, I do. Not the only shopper here. B: Nice! Do you like to swim? I am all about that exercise. A: I love swimming. I wish I could swim more though. B: I wish I could swim too. I am also much older. A: I'm not as young as I used to be. I lost weight then. I am wondering how old you are. B: I am 50 years old. Are you in school? A: Yep. I think its time to start. You don't like to wait long. B: Definitely. That is the life stage for me. I hope I can get back to age.	
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 look ...; 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? 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! A: Cool. I would have thought the same, but then again we have so many great dishes to eat. Did you enjoy any of them? B: I enjoyed them all! We stayed for more than one night, so I know how to do some great things! A: I am excited because I'm going on a cruise. B: I hope you have a great time! Cruise is such a great destination! A: There are so many different places to go on a cruise. Did you visit any other restaurants near where you were at? B: There were a lot! I saw a restaurant on my cruise tour that I had been to before. A: Was it expensive? Did you have a special restaurant? B: No, it wasn't expensive at all. We had hot dogs and rice, which I enjoyed most of the time.	

Table 13: Dialogue examples randomly sampled from BS$\mathbb{B}$T dataset.

Example 3	
Skill context from ConvAI2	**Skill context from ConvAI2**
I go to high school.; I play ice hockey; ...	I have a wife and two kids; I am a vegetarian; ...
Skill context from WoW	**Skill context from WoW**
Laziness	Laziness; Laziness is disinclination to activity or ...
Skill context from ED	**Skill context from ED**
I participated in my high school runners' marathon ...; proud	None
Dialogue	
A: One job would be better then my three.	
B: What do you like to eat? I'm vegetarian.	
A: I love beef broth, grilled cheese and eggs.	
B: Yum! My wife and 2 kids do the same.	
A: Family is also my life style. Maybe we should have a child!	
B: But what if you didn't want a child?	
A: If I didn't want a child I probably wouldn't have them.	
B: That's true! Our kids would just be us.	
A: That's what I thought, but then realized I didn't want a kid either.	
B: I understand that, it is hard to conceive.	
Example 4	
Skill context from ConvAI2	**Skill context from ConvAI2**
I like cupcakes; I like to read; I work at a grocery store; ...	I have 3 children; I am a karate black belt; ...
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 really like this girl at my job, but I am sort ... ; Apprehensive	None
Dialogue	
A: Do you have much experience using the different types of cleats? Which do you like best?	
B: I have a little. I also know about air jordans, a brand of footwear also popular with athletes.	
A: I enjoy air jordans as well. I like to play soccer, and it's extremely easy to get good footwear with or without shoes.	
B: I agree. Air jordans are good for the environment and for calves, and it's really easy to pull off.	
A: Air jordans are generally made with the most material so it makes sense they'd be easy to use.	
B: And now shoes can go together with clothes as well, like any other type of material.	
A: Do you know much about aeros then? They have excellent fit and beauty.	
B: I have a fan and a hard time believing it but they are great shoes.	
A: I hope you can find a fan that is compatible with your shoes.	
B: I like my new balance flying fan. Do you have any?	

Table 14: Dialogue examples randomly sampled from BS𝔹T dataset.

<table>
<tr><td colspan="2">Example 5</td></tr>
<tr><td>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</td><td>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</td></tr>
<tr><td colspan="2">Dialogue
A: Ok, that’s nice. How modern is your capital city?
B: It’s been around a minute, capital and most populous city is Algiers, it’s located in the far north of the country.
A: Very cool. How do you feel about it? Do you like your city or your country?
B: I love it. and love the people. What do you do for a living?
A: I currently have a ba in law. Do you have any pets?
B: I have 3 little ones, and 2 girls.
A: Nice! Do you like them? Will you move?
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.
A: It is expensive. Do you ever vacation? I like to visit the mountains.
B: Yes. I visit a lot of places there.</td></tr>
<tr><td colspan="2">Example 6</td></tr>
<tr><td>Skill context from ConvAI2
My favorite band is bon over; I am from Seattle; ...
Skill context from WoW
Marduk (band)
Skill context from ED
My best friend’s husband cheated on her with ...; Disgusted</td><td>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</td></tr>
<tr><td colspan="2">Dialogue
A: I have never heard of a chupacabra. Would you like to live forever?
B: No. However I would like to listen to Oasis’ hit "Live Forever" since you mention the idea.
A: That is true, I was hoping I could be of some help.
B: Anything that helps would be great . thank you very much
A: You’re welcome. There is also a band called ctcr that plays a bit of rap as well.
B: I’ll have to give them a listen.
A: There is an album you should check out. It’s called the on the court, you should really listen to it.
B: I’ll have to take a listen to it.
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.
B: Yes, I will do that as soon as possible.</td></tr>
</table>

Table 15: Dialogue examples randomly sampled from BS$\mathbb{B}$T dataset.

<table>
<tr><th colspan="2">Example 7</th></tr>
<tr><td>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</td><td>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</td></tr>
<tr><td colspan="2">Dialogue
A: That is funny, what does jaja mean?
B: Same as ll but in spanish!
A: well, that is interesting, I guess. Although I don't use spanish much.
B: Well it's used for many cultures, spanish comes after. What we get from it is also used in most western countries.
A: Okay, that is cool. Thank you again, I am learning spanish also!
B: I'm learning english as well! I'm so happy for you!
A: Thanks, that is interesting. Its like learning a foreign language.
B: That's a really cool skill to have. I bet you get to be learning all sorts of different languages.
A: I sure do, because it's an important skill.
B: Did you take it in high school? I know a lot of native spanish speakers.</td></tr>
<tr><th colspan="2">Example 8</th></tr>
<tr><td>Skill context from ConvAI2
I like to make cookies; I like bagels with everything on; ...
Skill context from WoW
Obesity in the United States
Skill context from ED
My girlfriend told me she's pregnant; Surprised</td><td>Skill context from ConvAI2
I love the smell of beans; ...
Skill context from WoW
Obesity in the United States; Obesity in the United ...
Skill context from ED
None</td></tr>
<tr><td colspan="2">Dialogue
A: Oh wow what kinds of videos do you have?
B: I do mainly make up tutorials. Do you have any side hobbies?
A: I like making cookies and eating cookie dough, I love bagels too.
B: I love bagels as well! What else do you like to do?
A: Yum i love bagels too! And I also love playing sports, but I'm too busy at this moment.
B: You should try joining a fitness team! It will make you feel better about you being busy.
A: lol I'll, but I'd feel like i wouldn't be ready.
B: You could always try it out! It's always fun to try out new things!
A: I would if I could I really appreciate new things and learn new things from people like you.
B: Sounds like you have an awesome hobby! Thanks for chatting.</td></tr>
</table>

Table 16: Dialogue examples randomly sampled from BS$\mathbb{B}$T dataset.