Generation of Large-scale Multi-turn Dialogues from Reddit

Daniil Huryn
Computer Science
Emory University
Atlanta, GA, USA
danikhur@gmail.com

William M. Hutsell
Computer Science
Emory University
Atlanta, GA, USA
mack.hutsell@gmail.com

Jinho D. Choi
Computer Science
Emory University
Atlanta, GA, USA
jinho.choi@emory.edu

Abstract

This paper presents novel methods to automatically convert posts and their comments from discussion forums such as Reddit into multi-turn dialogues. Our methods are generalizable to any forums; thus, they allow us to generate a massive amount of dialogues for diverse topics that can be used to pretrain language models. Four methods are introduced, Greedy$_{Baseline}$, Greedy$_{Advanced}$, Beam Search and Threading, which are applied to posts from 10 subreddits and assessed. Each method makes a noticeable improvement over its predecessor such that the best method shows an improvement of 36.3% over the baseline for appropriateness. Our best method is applied to posts from those 10 subreddits for the creation of a corpus comprising 10,098 dialogues (3.3M tokens), 570 of which are compared against dialogues in three other datasets, *Blended Skill Talk*, *Daily Dialogue*, and *Topical Chat*. Our dialogues are found to be more engaging but slightly less natural than the ones in the other datasets, while it costs a fraction of human labor and money to generate our corpus compared to the others. To the best of our knowledge, it is the first work to create a large multi-turn dialogue corpus from Reddit that can advance neural dialogue systems.

1 Introduction

With the advent of encoder-decoder frameworks (Brown et al., 2020; Lewis et al., 2020; Raffel et al., 2020), neural-based open-domain dialogue models have recently gained a tremendous interest as they start sounding more human-like than ever (Adiwardana et al., 2020; Zhang et al., 2020; Roller et al., 2021). Training robust neural-based models requires a huge amount of dialogue data in numerous topics that is difficult to procure as real human-to-human conversations are expensive and time-intensive to conduct (Godfrey et al., 1992). Several studies have presented large dialogue data created by crowdsourcing (Zhang et al., 2018; Dinan et al., 2019; Rashkin et al., 2019; Gopalakrishnan et al., 2019). However, it still requires non-trivial configurations in the cloud platform and the performance of crowd workers needs to be monitored constantly while paying them and the service a good amount of fees.

Most encoder-decoder models used in dialogue systems are not pretrained on dialogues, just fine-tuned on relatively small dialogue datasets, which is a limiting factor. Few studies have utilized comment threads in discussion forums for the creation of dialogue data and enhanced the performance of dialogue systems (Al-Rfou et al., 2016; Mazaré et al., 2018). However, these comment-originated dialogues tend to be short and not as sensible due to a lack of contexts from the main posts that are unsuitable for training multi-turn dialogue models. Such data scarcity points toward a necessity for a parameterized model that generates dialogues of different forms, styles, and topics in high quantity.

In this paper, we first introduce four algorithms to automatically convert posts and their associated comments from discussion forums such as Reddit into multi-turn one-to-one dialogues (Section 3). Our approach leverages the vast and available suite of human content and interaction online, with the potential to create many diverse dialogues, where the choice of subreddits constitutes topic selection. It also adapts a sentence-level language model for estimating likelihoods among posts and comments to sequence sounding utterances, and is analyzed across 10 subreddits (Section 4). We then create a large dialogue corpus and demonstrate the efficacy of our approach through head-to-head evaluation against dialogues from three well-known datasets, which indicates that our dialogues are as engaging and natural as those from others that are manually generated (Section 5). This work will facilitate the development of dialogue models in all kinds of areas that have been hindered by the data scarcity.[1]

[1]Our resources are available through https://github.com/emorynlp/reddit-to-dialogue.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3360–3373
October 12–17, 2022.

2 Related Work

Several dialogue datasets that are created through crowdsourcing have been presented. *Persona Chat* was created by assigning specific personas to two annotators who act as characters with those personas and generate a dialogue (Zhang et al., 2018). *Wizard of Wikipedia* was created by assigning two roles to annotators, the appentice acting as a curious learner of a specific topic and the wizard informing about the topic with a retrieved Wikipedia article (Dinan et al., 2019). *Empathetic Dialogues* was created by assigning two roles to annotators, the speaker describing a situation when a specific emotion would occur and the listener reacting to such a emotional situation (Rashkin et al., 2019). *Blended Skill Talk* was created by asking annotators to combine the personal, knowledgeable, and empathetic aspects of the previous three datasets together to generate more natural dialogues (Smith et al., 2020). *Topical Chat* was created by giving Wikipedia sections, fun facts, and news articles for a specific topic to annotators and asking them to generate a dialogue (Gopalakrishnan et al., 2019).

Only a few dialogue datasets have been created automatically. *Daily Dialogue* was crawled from various websites including dialogues scripted for English learners to practice daily conversations (Li et al., 2017). Mazaré et al. (2018) scrapped Reddit comments with replies and considered them short dialogues, which would not be multi-turn.

2.1 Comparisons to BST, DD, and TC

For comparisons to our corpus (Section 5.2), *Topical Chat* was chosen because it was least restricted in creation among crowdsourced datasets, *Blended Skill Talk* was chosen because it combined those 3 important aspects in dialogue, and *Daily Dialogue* was chosen because it was not crowdsourced but scripted by English educators. Table 1 shows comparisons among popular datasets and our corpus.

Data	DIA	UTT	TOK
Empathetic Dialogues	24,850	107,104	1,627,973
Wizard of Wikipedia	22,311	201,999	3,359,456
Daily Dialogue	13,118	103,632	1,504,635
Persona Chat	12,949	195,180	-
Topical Chat	10,784	235,434	4,614,506
Blended Skill Talk	6,808	77,502	1,058,325
Our Corpus	10,098	109,916	3,317,807

Table 1: The statistics of dialogue datasets including our corpus presented in Section 5. DIA/UTT/TOK: the total number of dialogues/utterances/tokens.

3 Reddit-to-Dialogue Generation

We introduce four algorithms for the dialogue generation: greedy baseline (Section 3.1), greedy advanced (Section 3.2), beam search (Section 3.3), and threading (Section 3.4). The main objective is to generate a multi-turn dialogue using a post and its comments (and replies)[2] that flows naturally in context. All algorithms assume that the number of sentences in the input post is less than or equal to the number of comments. The generated dialogues involve two speakers where utterances of Speakers 1 and 2 are extracted from the post and comments, respectively. All algorithms are evaluated on posts from diverse subreddits (Section 4).

3.1 Greedy Baseline Algorithm

Algorithm 1 depicts the baseline greedy approach that finds the most appropriate top-level comment for each sentence in a post. Given the input post $P = [p_1, .., p_n]$ where p_i is the i'th sentence in P, and the set of P's comments $\mathbb{C} = \{C_1, .., C_m\}$ s.t. $C_j = [c_{j1}, .., c_{j\ell}]$ where C_j is the j'th comment in $\mathbb{C}$ and c_{jk} is the k'th sentence in C_j, it first creates the set of comment segments T using $\mathbb{C}$ (L2), then visits every sentence $p_i \in P$ (L3), which gets appends to the output dialogue D (L4). Next, it finds the most-likely segment $\hat{t} \in T$ (L5)[3] and adds $\hat{t}$ to D (L6). Finally, T gets trimmed with $\hat{t}$ (L7) and the algorithm returns D as the output (L8).

Algo. 1: Greedy$_B$: greedy baseline

Input : P: a post, $\mathbb{C}$: a set of P's top-level comments.
Output: D: a dialogue.

1 $D \leftarrow [\,]$;
2 $T \leftarrow$ *segment*$(\mathbb{C})$;
3 **while** $\exists\, p_i \leftarrow$ *first*(P) **do**
4 $\quad D \leftarrow D \oplus [p_i]$;
5 $\quad \hat{t} \leftarrow \text{argmax}_{\forall t \in T}$ *ranker*(D, t);
6 $\quad D \leftarrow D \oplus [\hat{t}]$;
7 $\quad T \leftarrow$ *trim*$(T, \hat{t})$;
8 **return** D;

The *first* method removes and returns the first sentence in P. The *segment* method makes each comment a segment s.t. *segment*$(\mathbb{C}) = \{C'_1, \ldots, C'_m\}$, where $C'_j = c_{j1} \frown ... \frown c_{j\ell}$ ($\frown$: text concatenation).

[2] In this section, 'comments' imply the top-level comments of the post, and "replies" imply the replies to those comments.
[3] The *any* method returns any item in the input set.

The *ranker* method takes D comprising all previous utterances and p_i, then estimates the likelihood of t being the next utterance. Two models are used for this estimation, the human-like classifier (`HLC`) in DialogRPT (Gao et al., 2020) and BERT's next sentence predictor (`NSP`) (Devlin et al., 2019). At last, the *trim* method removes $\hat{t} = C'_j$ from T such that $\mathit{trim}(T, \hat{t}) = T \setminus \{C'_j\}$.

For `HLC`, p_i and t are fed into the model, which gives a score of how natural t is to follow p_i.[4] For `NSP`, since the original language model does not expect dialogue contents as input, we finetune it on the Multi-Session Chat dataset (Xu et al., 2022), the largest human-to-human chat dataset comprising ≈300K utterances. Given the finetuned model, the last two utterances in D (the last one is p_i and the second last one is a comment selected for p_{i-1}) are fed into the model with t, which gives scores for the two labels, `IsNext` and `NotNext`, s.t.

$$\mathit{score\_of}(\texttt{IsNext}) - \mathit{score\_of}(\texttt{NotNext})$$

is used for our likelihood estimation.[5]

3.2 Greedy Advanced Algorithm

Algorithm 2 shows the advanced greedy approach that makes two major updates from Algorithm 1.

Algo. 2: Greedy$_A$: greedy advanced

Input : P: post, $\mathbb{C}$: comment set, q: the max-length of comment segments.
Output: D: a dialogue.

1 $D \leftarrow [\,]$;
2 $T \leftarrow \mathit{segment}_a(\mathbb{C}, q)$;
3 **while** $\exists\, p_i \leftarrow \mathit{first}(P)$ **do**
4 $\quad D \leftarrow D \oplus [p_i]$;
5 $\quad$ **if** $\exists\, p_{i+1} \in P$ **then** $T \leftarrow T \cup \{p_{i+1}\}$;
6 $\quad \hat{t} \leftarrow \mathrm{argmax}_{\forall t \in T}\, \mathit{ranker}(D, t)$;
7 $\quad$ **if** $\hat{t} = p_{i+1}$ **then**
8 $\quad\quad P \leftarrow [\mathit{last}(D) \frown \mathit{first}(P)] \oplus P$;
9 $\quad$ **else**
10 $\quad\quad D \leftarrow D \oplus [\hat{t}]$;
11 $\quad\quad T \leftarrow \mathit{trim}_a(T, \hat{t}, q)$;
12 $\quad$ **if** $\exists\, p_{i+1} \in T$ **then** $T \leftarrow T \setminus \{p_{i+1}\}$;
13 **return** D;

First, it treats the next sentence $p_{i+1} \in P$ as a segment to rank if it exists (`L5`). If p_{i+1} is selected (`L7`), implying that it is better to have both p_i and p_{i+1} in one utterance, p_i is removed from D by $\mathit{last}(D)$, p_{i+1} is removed from P by $\mathit{first}(P)$, and their concatenation is prepended to P (`L8`). Once processed, p_{i+1} is removed from T (`L12`).

Second, the $\mathit{segment}_a$ method is updated (`L2`) such that it generates finer-grained segments using Algorithm 3. It is inspired by the fact that a single comment can (and often) address multiple aspects expressed in sentences that are not adjacent in P. In other words, one part of the comment may be appropriate for p_i while another part may be for p_j not adjacent to p_i; thus, using the whole comment as a response to either of them would be unnatural.

Algo. 3: $\mathit{segment}_a$: comment segmentation

Input : $\mathbb{C}$: a set of comments, q: the max # of sentences to join.
Output: T: a set of comment segments.

1 $T \leftarrow \varnothing$;
2 **foreach** $C_h \in \mathbb{C}$ **do**
3 $\quad$ **foreach** $i \in [1, |C_h|]$ **do**
4 $\quad\quad n \leftarrow \min(i + q - 1, |C_h|)$;
5 $\quad\quad$ **foreach** $j \in [i, n]$ **do**
6 $\quad\quad\quad T \leftarrow T \cup \{\mathit{join}(C_h, i, j)\}$;
7 **return** T;

The algorithm takes $\mathbb{C}$ and q, indicating the max-number of sentences allowed in any segment, and visits every comment $C_h = [c_{h1}, .., c_{h\ell}] \in \mathbb{C}$ (`L2`). For each sentence $c_{hi} \in C_h$ (`L3`), it joins all sentences between c_{hi} and c_{hj} using the *join* method as follows (`L4`: $i \le j \le \min(i + q - 1, |C_h|)$):

$$\mathit{join}(C_h, i, j) = \begin{cases} c_{hi} & \text{if } i = j \\ c_{hi} \frown \ldots \frown c_{hj} & \text{otherwise} \end{cases}$$

All joined segments are added to T (`L5-6`), which is returned as the output set (`L7`). For our experiments, $q = 3$ is used since the average number of sentences in Reddit comments (in our data) is < 4.

When $\mathit{segment}_a$ is applied to Algorithm 2 (`L2`), $\hat{t}$, which is appended to D, is a segment of a comment (`L10`). Let $\hat{t} = c_{jk}$ where c_{jk} is the k'th segment of C_j. The trim_a method then removes all segments generated for C_j from T (`L11`) such that

$$\mathit{trim}(T, c_{jk}, q) = T \setminus \mathit{segment}_a(\{C_j\}, q)$$

It is possible to keep the rest of unused segments from C_j that have no overlap with c_{jk}. However,

[4] Since `HLC` expected a sing-turn as input, we fed only p_i, although we also experimented by feeding more utterances in D, which led to worse performance.

[5] Feeding only p_i to `NSP` gave worse results whereas feeding more than two utterances in D gave very similar results, implying that BERT successfully learned to weigh more on the last two utterances. We also used only *score*(`IsNext`) as the estimator, which resulted in slightly worse performance.

such segments generally sound like "speeches out of context". Thus, we decided to remove all segments associated with C_j for the future selections.

3.3 Beam Search Algorithm

Algorithm 4 shows the beam search approach with an additional parameter k for the beam size. It creates the beam set $\mathcal{B}$ with the tuple of 6 items: (an input post P, a dialogue D, a segment set T, a sequence score θ, a sequence count ϕ), and the set of output dialogues $\mathcal{F}$ (L1-2). While there is any beam, the state set G is initialized (L3-4). Let Ω_α be $(P_\alpha, D_\alpha, T_\alpha, \theta_\alpha, \phi_\alpha)$. For every beam $\Omega_\alpha \in \mathcal{B}$, the first sentence $p_i \in P_\alpha$ is added to D_α (L5-7); p_{i+1} is added to T_α if it exists (L8-9). For each segment $t \in T_\alpha$, the copies $P'_\alpha, D'_\alpha, T'_\alpha$ are created from their correspondents in Ω_α (L11) and t gets handled the same as in Algorithm 2 (L12-16).

Algo. 4: Beam$_k$: beam search

Input : P: post, $\mathbb{C}$: comment set, q: max segment length, k: a beam size.
Output: D: a dialogue, θ: the sequence score, ϕ: the sequence count.

1 $\mathcal{B} \leftarrow \{(P, [\,], \mathit{segment}_a(\mathbb{C}, q), 0, 0)\}$;
2 $\mathcal{F} \leftarrow \varnothing$;
3 **while** $\mathcal{B} \neq \varnothing$ **do**
4 $G \leftarrow \varnothing$;
5 **foreach** $\Omega_\alpha \in \mathcal{B}$ **do**
6 $p_i \leftarrow \mathit{first}(P_\alpha)$;
7 $D_\alpha \leftarrow D_\alpha \oplus [p_i]$;
8 **if** $\exists\, p_{i+1} \in P_\alpha$ **then**
9 $T_\alpha \leftarrow T_\alpha \cup \{p_{i+1}\}$
10 **foreach** *segment* $t \in T_\alpha$ **do**
11 $(P'_\alpha, D'_\alpha, T'_\alpha) \leftarrow \mathit{copy}(\Omega_\alpha)$;
12 **if** $t = p_{i+1}$ **then**
13 $P'_\alpha \leftarrow [\mathit{last}(D'_\alpha)^\frown \mathit{first}(P'_\alpha)] \oplus P'_\alpha$;
14 **else**
15 $D'_\alpha \leftarrow D'_\alpha \oplus [t]$;
16 $T'_\alpha \leftarrow \mathit{trim}_a(T'_\alpha, t, q)$;
17 $s \leftarrow \mathit{ranker}(D'_\alpha, t)$;
18 $(\theta'_\alpha, \phi'_\alpha) \leftarrow (\theta_\alpha + s, \phi_\alpha + 1)$;
19 **if** $\exists\, p_{i+1} \in T'_\alpha$ **then**
20 $T'_\alpha \leftarrow T'_\alpha \setminus \{p_{i+1}\}$;
21 $G \leftarrow G \cup \{(\Omega'_\alpha, s)\}$;
22 **else**
23 $\mathcal{F} \leftarrow \mathcal{F} \cup \{(D'_\alpha, \theta'_\alpha, \phi'_\alpha)\}$;
24 $\mathcal{B} \leftarrow \mathit{top\text{-}k}(G, k)$;
25 **return** $\mathit{best}(\mathcal{F})$;

Given the ranking score s, the new sequence score θ' and count ϕ' are measured (L17-18). If p_{i+1} exists, it is removed from T'_α and the state (Ω'_α, s) is added to G, where $\Omega'_\alpha = (P'_\alpha, D'_\alpha, T'_\alpha, \theta'_\alpha, \phi'_\alpha)$ (L19-21). If $p_{i+1} \notin T'_\alpha$, no more sentences exist; thus, the current dialogue D'_α, its score θ'_α and the count ϕ'_α are stored in $\mathcal{F}$. Once all beams are used, $\mathcal{B}$ is reinitialized by the top-k states in G (L24) s.t.

$$G' \leftarrow [(\Omega_1, s_1), \ldots, (\Omega_{|G|}, s_{|G|})]\ (\forall i.\ s_i \geq s_{i+1})$$
$$\mathcal{B} \leftarrow \mathit{top\text{-}k}(G, k) = \mathbf{map}(\lambda_x : x[0], G')[:k]$$

Finally, $\mathit{best}(\mathcal{F}) = (D_\beta, \theta_\beta, \phi_\beta)$ is returned (L25) where $\theta_\beta \geq \theta_i : \forall_i.(D_i, \theta_i, \phi_i) \in \mathcal{F}$. Notice that although dialogues in $\mathcal{F}$ at L25 may comprise different lengths, the number of predictions made for every dialogue is the same as depicted in Figure 1. Although the first and second dialogues consist of 8 and 2 utterances respectively, the number of predictions made is 4 for both of them so that the min-length of any dialogue generated by our algorithm is 2 while the max-length is $2 \cdot n$ where $n = |P|$. Thus, it is safe to use θ_i as the sequence score, that is the sum of all prediction scores for the i'th path instead of the average score of θ_i/ϕ_i.

3.4 Threading Algorithm

Many top-level comments have threads of replies responded by the author of the post or other users. These replies are left out from our previous algorithms (Sections 3.1, 3.2, 3.3) because they do not necessarily address contents in the post. However, some of them can be used as bridging statements between comments selected by the algorithms and their following sentences from the post, which had been written before the comments were made. The challenge is how to glue a reply with the following sentence so that it does not sound disjointed.

Algo. 5: *thread*: threading

Input : P: post, D: dialogue, t: comment segment to be appended to D.
Output: D: the output dialogue including t and possibly a reply of t.

1 **if** $R \leftarrow \mathit{replies}(t)$ **then**
2 $\hat{r} \leftarrow \mathrm{argmax}_{\forall r \in R}\, \mathit{score}(P, D, t, r)$;
3 **if** $\mathit{glue}(P, D, t, \hat{r})$ **then**
4 **if** $P = \varnothing$ **then return** $D \oplus [t, \hat{r}]$;
5 **else** $P \leftarrow P \oplus [\hat{r}^\frown \mathit{first}(P)]$;
6 **return** $D \leftarrow D \oplus [t]$;

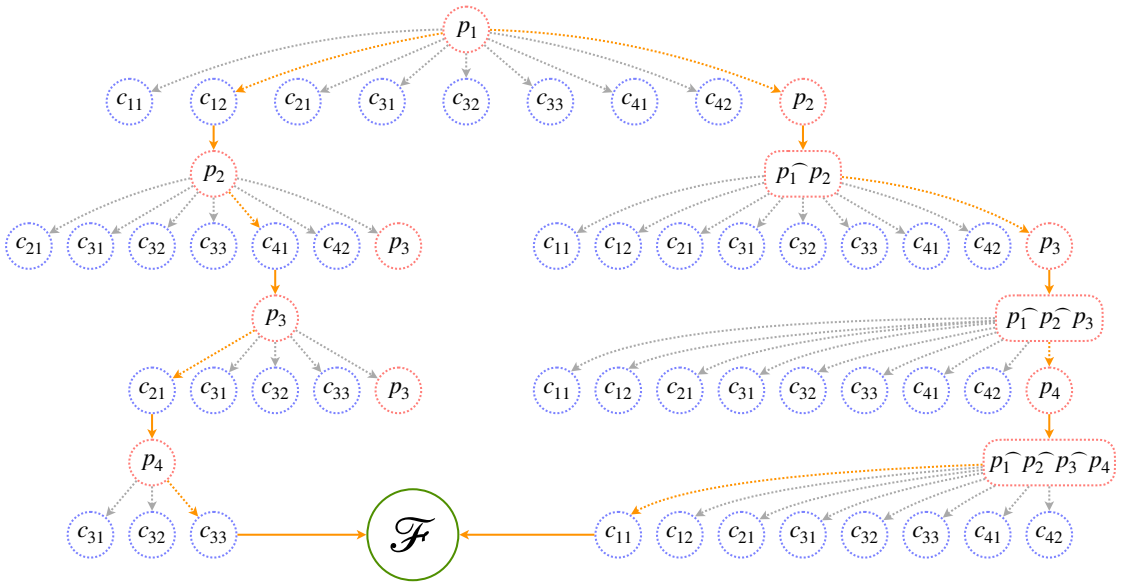

Figure 1: A demonstration of our beam search algorithm where one beam never selects p_{i+1} whereas the other beam always chooses p_{i+1} (the two extreme cases). The orange lines indicate the paths of the two beams where the solid lines are deterministically chosen while the dashed lines are predicted. The first beam results in $D_1 = [p_1, c_{12}, p_2, c_{41}, p_3, c_{21}, p_4, c_{33}]$ and the second beam results in $D_2 = [p_1 \frown p_2 \frown p_3 \frown p_4, c_{11}]$.

Algorithm 5 describes the threading algorithm that replaces `L15` in Algorithm 4 as follows:

$$D'_\alpha \leftarrow \textit{thread}(P'_\alpha, D'_\alpha, t)$$

Given the post P, the dialogue D, and the segment t from `L14` in Algorithm 4, it first retrieves the set R of all top-level replies using the *replies* method for the comment that t belongs to (`L1`). For every reply $r \in R$, given the last utterance $p \in D$ (that is p_i from `L7` in Algorithm 4) and the next sentence $n = \textit{first}(P)$, the *score* method measures how natural $\hat{r}$ would be in between t and n such that (`L2`):

$$\textit{score}(P, D, t, r) = \begin{cases} (1-\gamma)\mathbf{P}(r|p,t) + \gamma\mathbf{P}(n|t,r) & \text{if } P = \varnothing \\ \mathbf{P}(r|p,t) & \text{otherwise} \end{cases}$$

$\mathbf{P}(r|p,t)$ estimates how likely r is to follow p & t while $\mathbf{P}(n|t,r)$ estimates how likely n is to follow t & r. For this likelihood estimation, the next sentence predictor `NSP` in Section 3.1 is adapted s.t.:

$$\mathbf{P}(r|p,t) = \texttt{NSP}(p \frown t, r)$$
$$\mathbf{P}(n|t,r) = \texttt{NSP}(t \frown r, n)$$

For our experiments, $\gamma = 0.7$ is used, which gives more weight on $\mathbf{P}(n|t,r)$ than $\mathbf{P}(r|p,t)$. Then, the *glue* method takes the highest scoring reply $\hat{r}$ and adjusts the score by its length as follows (`L3`):

$$\hat{s} = \textit{score}(P, D, t, \hat{r}) \cdot (1 + 0.01(\lambda - |\hat{r}|))$$

In our case, $\lambda = 15$ so that it would boost the score if $|\hat{r}| < 15$, indicating that the number of tokens in $\hat{r}$ is less than 15, whereas it would drop the score if $|\hat{r}| > 15$. This guides the algorithm to select short replies that are found to be more useful. The *glue* method returns a boolean value as follows:[6]

$$\textit{glue}(P, D, t, \hat{r}) = \begin{cases} \textit{true} & \text{if } P = \varnothing \text{ and } 0 < \hat{s} \\ \textit{true} & \text{if } P \neq \varnothing \text{ and } \textit{wor}(p,t,n) < \hat{s} \\ \textit{false} & \text{otherwise} \end{cases}$$

$$\textit{wor}(p, t, n) = (1-\beta)\mathbf{P}(t|p) + \beta\mathbf{P}(n|t)$$

If there is no more sentence to be added and $0 < \hat{s}$, t and $\hat{r}$ are appended to D, which is returned (`L4`). If $P \neq \varnothing$ and $\hat{s}$ is greater than the weighted sum of $\mathbf{P}(t|p)$ and $\mathbf{P}(n|t)$, implying that it is more natural to include $\hat{r}$, the next sentence $n \in P$ is prepended by $\hat{r}$ (`L5`). In our case,[7] $\beta = 0.5$ so that $\mathbf{P}(t|p)$ and $\mathbf{P}(n|t)$ get equally weighed. Once R is processed, D is appended by t and returned (`L6`).[8]

[6] Note that $\hat{s} < 0$ if $\lambda < |\hat{r}| - 100$ (in our case, $115 < |\hat{r}|$) that is intended since it is better not to select such long replies.

[7] The hyperparameters β, γ, and λ are optimized by analyzing their performance on the dataset in Table 2.

[8] Since t and $r \frown n$ are introduced as separate utterances in D, estimating $\mathbf{P}(r, n|t) = \texttt{NSP}(t, r \frown n)$ instead of $\mathbf{P}(n|t,r) = \texttt{NSP}(t \frown r, n)$ seems to make more sense. However, we tried many different combinations of n-gram estimations including $\mathbf{P}(n|t,r)$, and the ones presented in this section yielded the best results overall.

3.5 Title Repetition Handling

Often on Reddit, the earlier part of a post assumes the context in the title so that it makes more sense to consider the title the first utterance of Speaker 1. On the other hand, the title and the first sentence in a post can be nearly or exactly the same such that including such a redundant title with the first sentence would lower the naturalness of our generated dialogue. Thus, any title that has a string match of 70% or higher with the first sentence in the post is excluded from the generation.

4 Algorithm Analysis

This section provides an in-depth analysis among the following five methods:

- G_B: Greedy Baseline (Section 3.1)
- G_A: Greedy Advanced (Section 3.2)
- B_2: Beam Search, $k = 2$ (Section 3.3)
- B_4: Beam Search, $k = 4$ (Section 3.3)
- T_2: Threading, $k = 2$ (Section 3.4)

To evaluate the quality of dialogues generated by these methods, diverse posts are collected from the following 10 subreddits:

- `ADV`: Advice
- `BKS`: Books
- `COL`: College
- `CaC`: Casual_Conversation
- `FIT`: Fitness
- `LTM`: LetsTalkMusic
- `MOV`: Movies
- `GAM`: TrueGaming
- `WRT`: Writing
- `TFR`: TalesFromRetail

Table 2 shows the statistics of our analysis set consisting of 50 posts uniformly distributed across the subreddits. On average, posts have 11.4 sentences (200.1 tokens) with 107.8 top-level comments that comprises 3.3 sentences (46.2 tokens), where a comment has 0.7 top-level replies (23.1 tokens).[9] The number of top-level comments varies quite a bit by the popularity of each subreddit.

All dialogues created by the five methods above are assessed by two undergraduates trained for this task. For this assessment, every utterance in these dialogues is double-annotated for whether it is appropriate for the context; more explicitly: "Is this a normal response or continuation of the previous statement?". This metric is chosen to evaluate the quality of dialogues as closely aligned with human dialogue-related intuition as possible.

[9]Tokens are split by whitespace, not linguistically.

Subreddit	Posts	Comments	Replies
`ADV`	5 (16.9)	36.3 (5.3)	0.2 (38.0)
`BKS`	5 (7.9)	591.4 (2.8)	0.6 (14.9)
`CaC`	5 (8.4)	142.4 (3.0)	0.7 (27.3)
`COL`	5 (7.1)	30.8 (3.9)	0.5 (39.3)
`FIT`	5 (3.9)	155.5 (3.5)	0.8 (27.7)
`GAM`	5 (19.0)	32.9 (6.5)	0.9 (52.0)
`LTM`	5 (13.6)	23.2 (5.5)	0.7 (49.0)
`MOV`	5 (8.8)	22.7 (2.5)	0.6 (21.5)
`RET`	5 (22.4)	23.5 (3.0)	1.0 (35.5)
`WRT`	5 (5.1)	35.0 (4.8)	0.3 (40.0)
Total	**50** (11.4)	**107.8** (3.3)	**0.7** (23.1)

Table 2: The analysis set used to evaluate our generation methods. Comments/Replies: the average number of top-level comments/replies per post/comment, (*) in Posts & Comments: the average number of sentences, (*) in Replies: the average number of tokens.

Table 3 shows the analysis results. Note that only the first 10 sentences in every post are used to create dialogues for this analysis to fairly score ones from distinct subreddits, although posts from certain subreddits (e.g., `FIT`, `WRT`) do not have 10 sentences on average, yielding shorter dialogues.

	κ	σ	UT	TK_1	TK_2	TK_a
G_B	55.9	39.1	13.9	20.2	35.4	27.8
G_A	27.1	51.1	11.4	21.2	35.1	28.1
B_2	24.0	67.2	9.8	22.9	36.9	29.9
B_4	45.8	58.9	9.8	23.0	38.1	30.5
T_2	**41.8**	**75.4**	**12.2**	**20.6**	**36.1**	**28.2**

Table 3: The analysis results from the five methods. κ: Cohen's kappa, σ: the avg-score of dialogues, UT: the avg-number of utterances, $TK_{1|2|a}$: the avg-number of tokens in every utterance from Speaker 1, Speaker 2, and all speakers, respectively.

For the inter-annotator agreement, Cohen's Kappa is used, which shows moderate agreements for G_B and T_2, where most utterances are found to be inappropriate and appropriate respectively such that they are easier to assess. The score of the dialogue is measured by macro-averaging the scores of the two annotators as follows ($n = 50$: the number of dialogues, $m = 2$: the number of annotators):

$$\frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{\#_i \text{ of appropriate utterances}}{\#\text{ of utterances}}\right)$$

G_A shows a good improvement of 12% over G_B, implying that it is often natural to include multiple sentences to compose utterances for Speaker 1. B_2 shows even a larger improvement over G_A, depicting the effectiveness of beam search, although no

ID	Utterance
S1	How do I get over the loss of a pet? Yes I have loss Pets before and I was sad.. But in this case I had a dog named Oscar he was in my life ever since toddlerhood.
S2	Losing a pet is like losing a family member. I think you might be taking this death harder because he didn't die on his own accord.
S1	I was even the one to name him.
S2	Unfortunately from my personal experience you don't really get over it but as days go on it gets easier. Just allow yourself to feel the emotions and get them out.
S1	Everyday I walked to my great granny's house to see him. He was my best friend seeing him became harder after my great grandmas death.
S2	Give yourself time. Accept your feelings and know that grieving is a process.
S1	He moved to Texas and I was only able to see him every summer.
S2	This will just make your grief and depression deeper and could spark an unending cycle of sadness. Join a support group. Speak with others who are also grieving.
S1	He was in my life srom toddler hood to now 17 years old.
S2	Unfortunately there is no real guideline to how long after the person is deceased, it is normal to be depressed, it will be different for everyone.
S1	He was in Texas and got hit by something being old and scared he wouldn't let anyone work on him.

Table 4: An example dialogue generated by T_2 using a post from the `Advice` subreddit. The original post can be found `https://www.reddit.com/r/Advice/comments/ub7k62`. S1/2: Speaker 1/2.

improvement is made when the bigger beam size is used for B_4. T_2 gives an additional improvement of 8.2% over B_2 and achieves the average score of 75.4%, implying that over ¾ of utterances in these dialogues are found to be appropriate.

For the dialogue lengths, it is not surprising that G_B yields longer dialogues as each sentence in the post is considered a separate utterance. Notice that B_* yield shorter dialogues than the others, indicating that the beam search prefers to combine more sentences from the post to compose utterances for Speaker 1. It makes sense because the `NSP` scores between consecutive sentences from the post are likely higher than ones with comments. However, the length is resolved with T_2 when the replies are concatenated to those sentences so that their `NSP` scores become more comparable to the `NSP` scores with the comments. Utterances for Speaker 2 tend to be longer than Speaker 1's ones. Table 4 shows a dialogue example generated by T_2, where most utterances are found to be appropriate.

5 Multi-turn Dialogue Corpus

With our best method, Threading (Section 3.4), we create a corpus consisting of 10,098 dialogues that can be used to pretrain language models for multi-turn dialogue systems (Section 5.1). Among those, 570 of them are compared to dialogues from three other datasets for quality assurance (Sec. 5.2).

5.1 Corpus Creation

A total of 28,686 posts and their comments/replies are collected from the 10 subreddits in Section 4 using the Python Reddit API Wrapper (PRAW).[10] The corpus creation follows 3 stages, pre-filtering, dialogue generation, and post-filtering. During the pre-filtering, posts that meet any of the following criteria are discarded:

- Include many non-standard characters (e.g., unicode characters, ones not in English)
- Include reddit-specific markers (e.g., `REPLY`, `DELETE`, `EDIT`, `OP`, `TL;DR`)
- Include many URLs, lists, or numbers
- Reference posts or comments in other posts
- The title includes the word 'thread' or the first sentence includes the word 'title'.
- The title is redundant to other posts.

The Threading algorithm is run on pre-filtered posts to automatically generate dialogues. During the post-filtering, any dialogue with the average sequence score (θ/ϕ in Algorithm 4) less than 6.0 gets discarded. Finally, dialogues are assessed by GRADE (Huang et al., 2020), an automatic coherence metric for open-domain dialogue, discarding dialogues with utterances with scores less than 0.21. Table 5 describes the statistics of each stage. 48.5% and 31.7% of the posts are discarded after the pre- and post-filtering, respectively. Our corpus consists of 109,916 utterances and 3,317,807 tokens, which makes it one of the largest dialogue datasets. More importantly, new large corpora can be created for a variety of topics with our method.

[10] PRAW: `https://praw.readthedocs.io`

It is especially useful for those who already have a small dialogue dataset and need a large corpus to pretrain language models for performance boost.

SR	ORG	PRE	POST	UTT	TOK
ADV	6,339	3,078	1,527	10.7	30.1
BKS	2,476	1,419	1,077	10.6	29.7
CaC	3,386	1,959	1,441	9.7	26.7
COL	4,008	1,637	1,117	8.3	30.0
FIT	1,964	422	342	10.5	29.2
GAM	1,873	1,057	632	15.9	37.4
LTM	1,882	1,049	819	12.7	36.1
MOV	2,341	1,417	1,071	8.5	26.0
RET	1,997	1,035	780	18.3	27.7
WRT	2,897	1,701	1,292	8.9	30.4
Total	28,686	14,774	10,098	10.9	30.2

Table 5: The statistics of our dialogue corpus. ORG: the number of collected posts, PRE/POST: the number of posts retained after pre/post-filtering, UTT/TOK: the avg-number or utterances/tokens.

5.2 Dialogue Evaluation

To evaluate the quality of our corpus, 570 of them are selectively sampled by their sizes and sources. Three size groups are formed, *small*, *medium*, and *large*, comprising dialogues with [6, 10], [11, 14], and [15, 17] utterances, respectively. Dialogues in this evaluation set are uniformly distributed across the 10 subreddits for fair comparisons.

Each of our dialogues is displayed with another dialogue with a similar size (and topic if possible) from one of the three datasets, Blended Skill Talk (BST), Daily Dialogue (DD), and Topical Chat (TC) such that a total of 1,710 (570×3) pairs are generated for comparisons.[11] Each pair is then compared by two annotators for *engagingness* and *naturalness* as follows:

Engagingness:
Which dialogue is more engaging or interesting?

Naturalness:
Which sounds more natural or human-like?

- 2: A is significantly more engaging/natural than B.
- 1: A is more engaging/natural than B.
- 0: A is as engaging/natural as B.
- -1: A is less engaging/natural than B.
- -2: A is significantly less engaging/natural than B.

[11] Section 2.1 explains why BST/DD/TC are chosen and gives detailed comparisons between their and our datasets.

The order of Dialogue A and B in each pair is randomly shuffled so that the annotators would not be able to tell their sources. For the annotation, we hired a professional team through SurgeHQ[12] and paid $0.5/task, costing a total of $1,710. Table 6 shows the evaluation results.

	CT	$\mathbf{B}_e$	$\mathbf{B}_n$	$\mathbf{D}_e$	$\mathbf{D}_n$	$\mathbf{T}_e$	$\mathbf{T}_n$
SM	190	0.39	-0.19	0.77	0.01	-0.51	-0.22
MD	190	0.52	-0.14	0.82	-0.03	-0.23	-0.16
LG	190	0.56	-0.28	0.71	-0.07	-0.04	-0.23
ADV	60	0.57	-0.23	0.76	0.05	-0.19	-0.07
BKS	60	0.53	-0.06	0.64	-0.16	-0.2	-0.23
CaC	60	0.38	-0.29	0.68	0.02	-0.39	-0.38
COL	43	0.28	-0.20	0.8	0.07	-0.43	-0.15
FIT	57	0.26	-0.34	0.56	-0.09	-0.53	-0.3
GAM	50	0.67	-0.12	1.05	0.18	0.13	-0.01
LTM	60	0.61	-0.11	1.02	0.16	-0.14	-0.20
MOV	60	0.66	-0.16	0.92	0.17	-0.12	0.00
RET	60	0.41	-0.50	0.43	-0.55	-0.40	-0.45
WRT	60	0.49	-0.04	0.87	-0.10	-0.33	-0.18
All	**570**	**0.49**	**-0.21**	**0.77**	**-0.03**	**-0.26**	**-0.20**

Table 6: The evaluation results. CT: the number of dialogues, B/D/T: BlendedSkillTalk, DailyDialogue, TopicalChat, $*_{e/n}$: the engagingness/naturalness score, SM/MD/LG: the small/medium/large size set.

For each annotation, our dialogue gets the score of 2/-2 if it is significantly better/worse, 1/-1 if it is better/worse, and 0 if it is as good as the other dialogue. The overall score is estimated by averaging all individual scores. In general, our dialogues are more engaging than BST and DD (0.49 and 0.77) but slightly less engaging than TC (-0.26) although longer dialogues are competitive to ones in TC (-0.04). On the other hand, our dialogues are less natural than the others although the differences are marginal (< 0.21). Our dialogues are found to be more natural for 6 out of 10 subreddits compared to DD. Considering how many human labors are involved for the creation of BST and TC while it costs no labor to create our corpus, these results are very promising. Example dialogues from this evaluation can be found in Appendix A.

It is worth mentioning that we first tried crowdsourcing our annotation through Mechanical Turk, which yielded random results as most turkers kept marking only Dialogue A without reading both of them carefully. When we switched the annotation tasks to SurgeHQ, a remarkable improvement was observed although the inter-annotator agreements were still low, 30.8% and 24.4% for the *engagingness* and *naturalness* tests, respectively. Such low

[12] SurgeHQ: https://www.surgehq.ai

agreements have been observed by previous works created the other datasets as well because this task is highly subjective. We will explore a more robust way of evaluating dialogues in the future.

6 Conclusion

We present four algorithms for the automatic conversion of posts and their comments/replies from Reddit discussion forums to multi-turn dialogues. Each algorithm is carefully designed and analyzed for high-quality generation. Our best method can generate dialogues with the 75% appropriateness level. Using this method, a large corpus is created consisting of 10,098 dialogues from 10 subreddits. The quality of our dialogues is tested by comparing them to dialogues from three popular datasets, BlendedSkillTalk, DailyDialog, and TopicalChat. Our dialogues are more engaging, but slightly less natural than those from the other datasets overall.

For future work, we will improve our methods, apply them to broader subreddits, and adapt them to other discussion forums as our methodology is not limited to Reddit. We will also train dialogue models on our corpus for a more in-depth extrinsic evaluation against other dialogue datasets.

Acknowledgements

We gratefully acknowledge the support of the Amazon Alexa AI grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Amazon.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *CoRR*, abs/1606.00372.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations*, ICLR.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue Response Ranking Training with Large-Scale Human Feedback Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

John J. Godfrey, Edward Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 1891–1895.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A man-

ually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason E. Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

A Appendix

A.1 Example Dialogues from Each Evaluation Grading Category

This section give example dialogues rated as "Less Natural, Less Engaging" (Table 7), "More Natural, More Engaging" (Table 8), "Significantly More Natural, Significantly More Engaging" (Table 9), and "Significantly Less Natural, Significantly Less Engaging" (Table 10). Note that these ratings are given in relation to a conversation, omitted here, from one of our comparison datasets (Section 5.2).

The negatively-rated dialogues exhibit several of our methodology's weaknesses, namely lack of interaction from Speaker 1, Speaker 2 responses to content which has not yet been introduced by Speaker 1, and references to Reddit-specific terminology such as threads which wouldn't be discussed often in regular dialogue.

The positively-rated dialogues, on the other hand, demonstrate our methodology's strengths: both successfully use threading (e.g., Utterance 5 in Table 8 and Utterance 3 in Table 9), there is consistently natural grouping of Speaker 1's content, and Speaker 2 responds for the most part at natural times to the content Speaker 1 introduces while also adding relevant content to the conversations.

ID	Utterance
S1	Bittersweet musical legacies: an examination. The popularity of the recent Big Star thread got me thinking about how bittersweet it is for music fans to deeply love a band that were either struck down by tradegy, failed to launch due to bad decisions, or never received their day in the sun beyond a cult following.
S2	I'd say Badfinger certainly qualifies for this thread. They had the support of the Beatlesĺabel, they looked to be on the road to a successful career with many big hits like ""Day After Day"", ""No Matter What"", and ""Come and Get It"" (and their song ""Without You"" was made a big hit by Nilsson)
S1	One of the saddest legacies for me is Emitt Rhodes, a guy who at the tender age of 20 in 1970 put out an album with musical chops worthy one Sir Paul McCartney.
S2	I would argue The Kinks belong in this thread. Yeah they have some songs that poked through the fog (You Really Got Me, Lola, Sunny Afternoon), but they should really be considered right in the mix in the Beatles/Stones conversation. I feel like The Kinks are not given that same respect but they deserve it.
S1	Having that level of widespread public acknowledgement is quite rare. A cynic may say he cribbed macca's sound, and in all honesty a song could've appeared on his first solo album, but at the age of 20 he had many years ahead of him to transcend his influences....except that he didn't.
S2	...and then they got royally screwed over financially, had a lot of issues within the band (I believe the royalties to "Without You" were among the reasons they fought), and because of all the fallout of the financial ruin and such, in the end, two of the band members ultimately committed suicide in eerily similar fashion. Such a sad, sad story.
S1	The fact he had a kid on the way when he hanged himself is tragic. A handful of albums into his career he was beset by lawsuits and record company entanglements that saw him walk away from releasing music in 1973 and not appear again with a new album until 2016(!).. If you look at the photos of Emitt and then jump to the album Rainbow Ends, it's like a joyful/sad pair of bookends for a man who in interviews talked about his regrets with music and his lack of success.
S2	I also think of The Verve and how the copyright shenanigans robbed them of their hit Bittersweet Symphony. And then that was kind of it for them.
S1	I loved Urban Hymms, listened to that record so much over my life. This is made even more sad by the fact he died in his sleep last year of heart failure at age 70.
S2	These are examples of artists who missed out on their full commercial potential in life rather than good artists that we like to think could've been great artists because the circumstances weren't right. I'm not familiar enough with Emmitt Rhodes to say whether he fits in either camp, just my two cents on the issue.
S1	I'd add Townes Van Zandt to that list. His talent as a songwriter was undeniable.
S2	Sadly the music industry is littered with tragic stories and acts that should have been much larger than they were, but for whatever reason, weren't. One of my favorite ""one hit wonders"" was a dude named Jonathan Edwards who had a hit with in 1971 about the craziness of the music industry and that was pretty much it despite making great music up into the 80s.
S1	The press called him a one man Beatles due to the fact he played and recorded all his own instruments at his home studio.
S2	his story is particularly sad, simply because he faced so much tragedy and hardship, and no one was really there for him. the one album he put out is a reflection of his life and mental health issues and is as equally as beautiful as it is heartbreaking. tim buckley, on the other hand, released 9 albums over the span of a 9 year career and could've have become a musical legend similar to that of a lot of psychedelic folk/folk rock artists of the time.

Table 7: An example dialogue generated by T_2 using a post from the `LetsTalkMusic` subreddit that was graded as less natural and less engaging. The original post can be found at https://www.reddit.com/r/LetsTalkMusic/comments/pnwg25/bittersweet_musical_legacies_an_examination/. S1/2: Speaker 1/2.

ID	Utterance
S1	Do you use a bench/machine if someone says "I'm still using it" when they've left nothing there to show? I see the benchpress spot free and no sign of being used.
S2	so I can use it. Then walk away and use any other machine.
S1	No hoodie, no towel, no water bottle nothing.. So I put my hoodie down on the ground while I go get a paper towel to clean some sweat on the bench.. Then, some guy says ""Hey I'm still using that""
S2	If he's on another machine he's not on the one you want to use. Use it.
S1	Idek how to sue lol. He's about 25 feet away just standing and talking to someone.
S2	Just ignore him entirely and bench lol If the guy wants to say something, he can have some manners and come over instead of shouting over.
S1	His tone was kinda rude too, as if it was obvious that the damn thing's still in use.
S2	If it becomes an issue, just keep an eye out for whatever he's doing and as soon as you see him go to any piece of equipment, yell Hey, I'm using that. . . even if you are in the middle of your set, in another machine, across the room. On the other hand, maybe the guy is not a complete jerk and he really was using it and it's no big deal.
S1	Stuff gets annoying
S2	Screw that guy. Getting knocked out isn't that bad lol. Hell you may even get a bunch of money out of it.
S1	I swear.
S2	Nothing by on equipment? Take it, get your sets, and remain oblivious to obnoxious behavior.

Table 8: An example dialogue generated by T_2 using a post from the `fitness` subreddit that was graded as more natural and more engaging. The original post can be found at https://www.reddit.com/r/Fitness/comments/tth3in/do_you_use_a_benchmachine_if_someone_says_im/. S1/2: Speaker 1/2.

ID	Utterance
S1	Smartest Theft Plan That Almost Worked. When I was 17, I had a part-time job as a cashier for Rona.. During my time there I met some very interesting people at the till.
S2	Best concealment I ever saw on the job as a cashier is some guy tried to fill a beanbag chair with vitamin bottles...
S1	Thats some clever thinking! Once, a man walked in wearing what looked like a Halloween construction worker costume and headed for what looked like the power tools section.. Although, I later found out he went to the cleaning aisle first where we sell commercial yellow mops and buckets.. He opened one box and put the mop and bucket on the sales floor to make it look like it belonged to the store.. Then he used the empty box and filled it with expensive power tool kits and batteries.
S2	He assembled them in his cart and loaded them up with all the toys he wanted. Then tried to just pay for two boxes that come flat packed and obviously had to be opened. Surprise surprise they were heavy.....
S1	11 times. He had the box in a shopping cart and went up to my till.
S2	When I worked at a big box hardware store we one day had a guy come in and grab some of the $5 heavy duty moving boxes.
S1	This week. This was an elaborate plan up till now, unfortunately, he placed the box the wrong way in the cart where I couldn't just use my scanner to get to the barcode.. He tried to turn the box but since he put all these heavy tools and batteries in it, he couldn't do it.
S2	Hilarious. Pretty smart. Wonder if you would have caught him if he had put the box in the right direction.

Table 9: An example dialogue generated by T_2 using a post from the `TalesFromRetail` subreddit that was graded as significantly more natural and significantly more engaging. The original post can be found at https://www.reddit.com/r/TalesFromRetail/comments/ppsa29/smartest_theft_plan_that_almost_worked/. S1/2: Speaker 1/2.

ID	Utterance
S1	Your money is gone! Please calm down! So, I recently got a job at a store that sells cheap stuff for around a dollar or more.
S2	The grocery store I go to has a sign at each till saying don't put money on the belt as it could cause damage. Any money lost is not the responsibility of the store. Yet, idiots still do it, but their clearly posted sign covers them and they just shrug their shoulders.
S1	Most customers are usually polite and very pleasant to talk to.
S2	In Finland some stores have signs that explicitly says no money or recycling slips on the belt, probably because alot of people used to put their recycling slips in with the groceries.
S1	Im not a fin (sorry) what is a recycling slip? Some however, make me seriously regret accepting this job.. Today I had one of those customers... A woman came into my line today with one thing and, she did the dumb thing of putting her cash on the belt.. Which you should never do because the belt acts like a vending machine when you put paper money on it.. Henceforth the belt took her $5 bill and she instantly started flipping out like crazy, and I mean she was hysterical.
S2	My brother had something similar happen at a grocery store but the customer said it was a $100 bill and my brother didn't witness it.
S1	A hundred is much worse than a $5. She stated how she wasn't leaving until she gets her five dollars back.
S2	The worst thing I've seen one of those belts try to eat was a 2L of coke. The bottle fell on its side on the belt and got caught and poked a hole in the side. Then the bottle kept rolling at the end of the belt spraying soda all over the place.
S1	There is no service desk. Then she starts making a list of demands such as... - wanting to talk to the owner of the store.
S2	Everyone just kinda froze until the cashier stopped the belt and someone grabbed the bottle and put a fingerbover the hole until they could throw it out. It was horrible and hilarious at the same time.
S1	The store I work in is very small.

Table 10: An example dialogue generated by T_2 using a post from the `TalesFromRetail` subreddit that was graded as significantly less natural and significantly less engaging. The original post can be found at https://www.reddit.com/r/TalesFromRetail/comments/tx1fg0/your_money_is_gone_please_calm_down/. S1/2: Speaker 1/2.