Learning From Textual User Feedback — Collect New Datasets Or Extend Existing Ones?

Anonymous ACL submission

Abstract

Learning from textual user feedback, i.e., user responses that address errors in system utterances, is increasingly important to continuously improve dialogue systems, but datasets that provide the needed annotations, i.e., annotations for causing errors and user responses, are scarce. As creating new datasets involves an immense manual effort, we investigate the extendability of various existing dialogue datasets with annotations for errors and user responses. In order to facilitate the detection of dialogues that contain such data, we propose Textual Feedback Detection (TFD), a semi-automatic approach to identify textual user feedback. Furthermore, we propose two taxonomies optimized to categorize such data, a user response type taxonomy and an error type taxonomy. In our study, we annotate 1,155 dialogues from six different dialogue datasets with both errors and corresponding user responses. Our findings give insights on dataset-specific error and user response types. We show that open-domain and knowledge-grounded dialogue datasets are more appropriate to be extended with annotations for causing errors and user responses than task-oriented dialogue datasets[1].

1 Introduction

Textual user feedback is of growing importance for continuously improving dialogues systems (Hancock et al., 2019; Veron et al., 2021; Park et al., 2021) or their components, e.g., external knowledge bases (Mazumder et al., 2019). It is defined as a user response that addresses an error in a previous system utterance (e.g., factually incorrect information) by expressing user satisfaction or dissatisfaction, describing new knowledge (new concepts), providing corrections or alternative responses. However, due to a lack of datasets that provide the needed annotations, i.e., annotations for causing errors and user responses, authors of previous works collected this data on their own. To address this issue, recent works conducted resource-intensive manual collection efforts to publish large-scale curated datasets, such as FITS (Xu et al., 2022). However, they mostly focus on open-domain dialogues. For other dialogue types, such as task-oriented dialogues, the lack of publicly available datasets is still persistent. This hinders research on continuously improving dialogue systems, e.g., through lifelong learning, and it seems infeasible and impractical to collect new and appropriately annotated datasets for each case. Alternatively, existing datasets could be extended with the needed annotations. Many of the publicly available dialogue datasets are well-established and have been improved over years. If extended with the needed annotations, new learning paradigms, such as lifelong learning, could benefit from these advantages.

In this paper, we investigate the extendability of six existing dialogue datasets with annotations for textual user feedback, i.e., causing errors and user responses. We focus on datasets with task-oriented dialogues, i.e., MultiWoZ (Zang et al., 2020), BABI (Bordes et al., 2016) and SGD (Rastogi et al., 2020). However, to cover a broad variety of dialogue types, we also consider Wizards-of-Wikipedia (Dinan et al., 2018) as knowledge-grounded, and PersonaChat (Zhang et al., 2018) and the human-bot split from the Self-Feeding Chatbot (Hancock et al., 2019) as open-domain dialogue datasets. Except for the latter one (which is partly annotated with alternative responses), none of these datasets is originally intended for learning from textual user feedback. Accord-

[1] Code and annotated data is available here: http://test.test

ingly, a significant amount of dialogues may not contain such data at all. Therefore, a purely manual analysis might be very inefficient. To facilitate this by filtering for potentially relevant dialogues, we propose Textual Feedback Detection (TFD), a semi-automatic approach to identify textual user feedback, i.e., user responses that address errors in system utterances. For the subsequent analysis of extendability, we annotate subsets of the filtered dialogues with user responses and error types. For this purpose, we propose two new taxonomies, a user response type and an error type taxonomy, since we find that none of the existing taxonomies fits our needs.

2 Related Work

2.1 Datasets for Learning From Textual User Feedback

Due to the lack of annotated data, authors of previous works collected the needed data on their own. Veron et al. (2021) proposed a general methodology for evaluating continuous learning with task-oriented dialogues systems. They generated 47,000 dialogues, annotated with new knowledge, using a pattern-based approach. However, their dataset is not publicly available. Park et al. (2021) proposed a scalable approach for continuously improving models for natural language understanding by leveraging user feedback. They collected 2,000 correction-annotated dialogues from in-production use. However, they also did not publish their data. For the Self-Feeding Chatbot, Hancock et al. (2019) collected (and published) 60,000 English open-domain human-bot dialogues, partly annotated with alternatives for problematic system responses. It is one of the largest publicly available datasets with annotations for textual user feedback. Other datasets that target open-domain dialogue systems and contain textual user feedback are FITS (Xu et al., 2022) and SaFeRDialogues (Ung et al., 2022). FITS is a manually collected dataset of 14,000 human-bot conversations annotated with up to five different feedback types, including textual user feedback. SaFeRDialogues consists of 7,000 human-bot dialogues with annotations for offensive answers along with respectful alternatives.

While FITS and SaFeRDialogues are already widely adopted, it seems like the dataset from Hancock et al. (2019) was never reused. We are not aware of any publicly available dataset for learning from textual user feedback for other dialogue types than open-domain, such as task-oriented dialogues.

2.2 Taxonomies For Errors In Dialogues

In order to identify errors in system responses, the works presented in Section 2.1 only use coarse-grained taxonomies, customized for their specific use case. FITS (Xu et al., 2022) only differentiates search query-based, results-based, or final response-based errors. SaFeR-Dialogues (Ung et al., 2022) only considers safety failures. For the Self-Feeding Chatbot, Hancock et al. (2019) collected user feedback based on a measured satisfaction score. They do not even distinguish error types.

Since we do not know which error types are represented in the datasets examined in this study (if any), none of these taxonomies is applicable to our case. Fortunately, errors in human-machine interaction, especially in the context of conversations, have been studied for long, which is why there are established taxonomies available for this purpose. Dybkjaer et al. (1996) proposed an error taxonomy for task-oriented dialogues that consists of four error categories and also takes background knowledge into account, e.g., whether the user is an expert or a novice. The taxonomy proposed by Möller et al. (2007) provides six error types, i.e., goal-level, task-level, command-level, concept-level, recognition-level, or other errors. However, it focuses on practical aspects and ignores content-related errors like factually incorrect information. Recently, Higashinaka et al. (2021) proposed an integrated taxonomy, that is applicable to all types of dialogues. It consists of 17 error types, e.g., ignore question, contradiction, or lack of sociality, across four categories (levels), i.e., utterance, response, context, and society, divided in two violation types, i.e., violation of form and violation of content (see also Table 1).

Due to its wide applicability, we use the integrated taxonomy by Higashinaka et al. (2021) as the base taxonomy for errors in this work. However, none of the existing works proposed a taxonomy for classifying user responses that address errors in system utterances and we are not aware of any other work that proposes something in this regard.

3 Datasets

In our study, we consider multiple well-established datasets of different dialogue types, i.e., task-oriented, knowledge-grounded, and open-domain dialogues. However, we choose the majority of datasets from task-oriented dialogues as there are no datasets publicly available for learning from textual user feedback. Many of these datasets consist of human-human dialogues. For simplicity, we do not distinguish in the terminology between a system and a human being as interlocutor. We always refer to the interlocutor's utterance as a system utterance.

3.1 Task-Oriented Datasets

We consider three task-oriented datasets in this work, i.e., MultiWoZ (Zang et al., 2020), SGD (Rastogi et al., 2020), and BABI (Bordes et al., 2016). While MultiWoZ and SGD consist of human-human conversations, BABI only contains human-bot dialogues. The dialogues in MultiWoZ span seven different domains across 8,438 dialogues (with up to five different domains in one dialogue). SGD consists of 16,000 dialogues across 16 domains. Both datasets provide extensive annotations, such as for natural language understanding or state tracking. BABI only targets one domain, restaurant booking, and consists of 6,235 dialogues across six tasks of increasing difficulties.

3.2 Knowledge-Grounded Datasets

For knowledge-grounded datasets, we focus on Wizards-of-Wikipedia (Dinan et al., 2018). It consists of 22,311 human-human dialogues across 1,365 different topics.

3.3 Open-Domain Datasets

In case of open-domain datasets, we consider PersonaChat (Zhang et al., 2018) and the human-bot split of the Self-Feeding Chatbot (Hancock et al., 2019). PersonaChat consists of 10,907 dialogues between two partners that are randomly assigned to one of 1,155 different personalities. The task is to get to know each other during conversation. The human-bot split of the Self-Feeding Chatbot consists of $60,000$ dialogues and is partially annotated with alternative responses[2]. With this, it exceeds the size of the other datasets considered in our study.

Hereinafter, we refer to MultiWoZ (Zang et al., 2020), PersonaChat (Zhang et al., 2018), Wizards-of-Wikipedia (Dinan et al., 2018), and the human-bot split of the Self-Feeding Chatbot (Hancock et al., 2019) as *MWoZ*, *PC*, *WoW*, and *SFC*.

4 Study Design And Implementation

Since most of the datasets examined in this study are not originally intended for learning from textual user feedback, many of the dialogues may not contain the needed data at all, i.e., user responses that address errors in system utterances by expressing user satisfaction or dissatisfaction, describing new knowledge (new concepts), providing corrections, or alternative responses. For this reason, a purely manual analysis would be very inefficient. Therefore, we follow a two-step semi-automatic approach: (1) Filtering the investigated datasets for dialogues that potentially contain textual user feedback, i.e., user responses that address errors in system utterances, (Section 4.1), and (2) manually analyzing the extendability of these datasets by annotating subsets of the filtered dialogues with user responses and error types. For this purpose, we propose two new optimized taxonomies, a user response type taxonomy (Section 4.2) and an error type taxonomy (Section 4.3).

4.1 Textual Feedback Detection

We propose Textual Feedback Detection (TFD) to identify potential textual user feedback (user responses that address errors in system utterances) by exploiting the semantic similarity between user responses and feedback-indicating sentences. It is a two-step process that first requires (manual) collection of feedback-indicating sentences, to then filter for relevant dialogues (automatically). A feedback-indicating sentence is a sentence that is known to contain a feedback-indicating phrase, a text fragment of arbitrary length (n-grams) that in-

[2] We only consider the non-annotated dialogues in our study.

dicates user dissatisfaction or an error in the previous system response.

4.1.1 Collection of Feedback-Indicating Sentences

To collect feedback-indicating sentences, we first identify errors in system utterances based on the error taxonomy from Higashinaka et al. (2021), and then collect the feedback-indicating sentence from the following user responses. For this step, we manually analyse a randomly sampled set of 1,200 dialogues (200 from each of the six investigated dataset).

The taxonomy from Higashinaka et al. (2021) consists of 17 error types (I1-I17) across four levels, i.e., utterance, response, context, and society. They further differentiate between content violation, i.e., the content of the response might cause a dialogue breakdown, and form violation, i.e., the content is not interpretable due to massive grammatical problems. Table 1 shows the error types[3].

Level	Form Violation	Content Violation
Utterance	Uninterpretable (I1)	Semantic error (I3)
	Grammatical error (I2)	Wrong information (I4)
Response	Ignore question (I5)	Ignore expectation (I9)
	Ignore request (I6)	
	Ignore proposal (I7)	
	Ignore greeting (I8)	
Context	Unclear intention (I10)	Self-contradiction (I13)
	Topic transition error (I11)	Contradiction (I14)
	Lack of information (I12)	Repetition (I15)
Society	Lack of sociality (I16)	Lack of common sense (I17)

Table 1: Error Types defined by Higashinaka et al. (2021). The numbers in the brackets are the corresponding identifiers.

Overall, we collect a set of 68 feedback-indicating sentences ($\sim$ 6.52 words per sentence), each with a unique feedback-indicating phrase ($\sim$ 3.52 words per phrase)[4]. Table 2 shows the distribution of feedback-indicating sentences across datasets.

We find most feedback-indicating sentences in open-domain and knowledge-grounded datasets, especially in SFC (Hancock et al., 2019), a human-bot dataset.

[3]See Appendix A for details on error types.

[4]See Appendix B for all collected phrases and sentences. Contractions (two words that have been connected, e.g., *don't* or *it's*) are considered as one word.

	Task-Oriented			**Open-Domain**		**Know.-Grounded**
Dataset	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
#Sentences	7	0	5	9	36	36

Table 2: Distribution of feedback-indicating sentences across datasets. *HH* denotes human-human dialogues and *HB* denotes human-bot dialogues.

4.1.2 Filtering For Potential Textual User Feedback

For each dataset, we decompose every dialogue into turns (pairs of user and system utterance), extract the user response, and decompose it into sentences. Next, we pair each of these sentences with every feedback-indicating sentence (collected in Section 4.1.1), and use a pretrained Sentence-Transformer (Reimers and Gurevych, 2019) to calculate the semantic similarity of each pair. We assume a dialogue to contain textual user feedback (user responses that address errors in system utterances) if it contains at least one user response with a sentence that has a similarity $\geq$ 50% to at least one of the feedback-indicating sentences.

For implementation, we use PyTorch (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020). As pretrained Sentence-Transformer, we use *all-mpnet-base-v2* [5]. It is based on MPNet (Song et al., 2020) but finetuned on a large corpus of sentence pairs from multiple tasks and domains, e.g., Yahoo Answers (Zhang et al., 2015) and Reddit Comments (Henderson et al., 2019), using a contrastive objective. It is a 12-layer Transformer model with a vocabulary size of 30,527 words that calculates the cosine similarity between two sentences in a 768-dimensional dense vector space.

4.2 User Response Type Taxonomy

While collecting feedback-indicating sentences (Section 4.1.1), we observed five different types of user responses that follow errors in system utterances:

- **UR1** — The user ignores the error and continues the conversation.
- **UR2** — The user repeats or rephrases his/her concern.

[5]The model is available here: `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`, last accessed 11/10/2022.

- **UR3** — The user makes the system aware of the error and provides a correction.
- **UR4** — The user makes the system aware without providing a correction.
- **UR5** — The user asks for clarification.

Among these, we find that UR2, UR3, and UR5 are likely to contain textual user feedback, i.e., user responses that express satisfaction, dissatisfaction or provide corrections. However, we have never observed other kinds of user responses in this context, e.g., responses that provide alternatives or new concepts (new knowledge), which is why we do not include them in our taxonomy.

4.3 Error Taxonomy

During the collection of feedback-indicating sentences (Section 4.1.1), we found that the integrated taxonomy as proposed by Higashinaka et al. (2021) is not optimal for identifying errors in system responses. We observe (1) that six of the 17 error types are never observed in the data – e.g., *uninterpretable* (I1), which describes system responses that just consist of linguistically invalid text fragments –, and (2) three ambiguous error types – e.g., *ignore expectation* (I9) and *ignore request* (I6) are very similar, as in both cases the system does not provide the expected output. For this reason, we propose a condensed error taxonomy that is optimized for the classification of errors in system utterances. Table 3 shows this new taxonomy.

Level	**Error Type**	**Description**
Response	Ignore Question (E1)	The system utterance ignores the user's question.
	Ignore Request (E2)	The system utterance ignores the user's request to do something.
	Ignore Expectation (E3)	The system utterance does not fulfill the user's expectation.
	Slot Error (E4)	The system utterance suggests that the system did not get the slots right.
	Factually Incorrect (E5)	The system utterance contains information that is factually incorrect.
Context	Topic Transition Error (E6)	The system utterance transitions to another / a previous topic without reasonable explanation.
	Conversationality (E7)	The system utterance indicates that the system lost track, e.g., it repeats previous responses (without asking for missing information) or contradicts itself.
	Unclear Intention (E8)	The system utterance suggests that the user's intent was not successfully conveyed.
Society	Lack of Sociality (E9)	The system utterance lacks consideration of social standards, e.g., greetings, is toxic or disrespectful.
	Lack of Common Sense (E10)	The information in the system utterance opposses the opinion of the majority.

Table 3: Taxonomy for the classification of errors in system utterances.

We ignore the utterance-level error from the original taxonomy as we never observe them. For the same reason, we ignore *lack of information* (I12 in Table 1). This does not mean that these error types are in general irrelevant. We just do not observe them in any of the system utterances. Furthermore, we ignore *contradiction* (I14 in Table 1) as it is only rarely observed – and only as a result of *lack of common sense* (I17 in Table 1, now E10) or *factually incorrect* (E5) errors. We merge *ignore proposal* (I7 in Table 1), a response-level error type, and *Ignore Request* (I6 in Table 1), as both are very similar (now E2 in Table 3). Next, we merge *ignore greeting* (I8 in Table 1) with *lack of sociality* (I16 in Table 1, now E9), as the latter implies the first one. E5 is a new error type that replaces *wrong information* (I4 in Table 1) by extending its original definition for taking also factually incorrect knowledge into account. We also merge *repetition* (I15 in Table 1) and *self-contradiction* (I13 in Table 1) to what we call *conversationality* (E7), as we observe both error types rarely, and if, only in situations that suggest that the system has lost the thread. We also observe cases of (obviously) incorrectly conveyed attributes in task-oriented dialogues that were not covered by the original taxonomy. For such cases, we introduce *slot error* (E7).

5 Findings

We apply TFD on the datasets investigated in this work to filter them for dialogues that potentially contain textual user feedback, i.e., user responses that address errors in system utterances (see Section 4.1)[6]. Table 4 shows the results[7].

	Task-Oriented			Open-Domain		Know.-Grounded
Dataset	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Original Size	8,438	16,000	6,235	10,907	60,000	22,311
TFD-Filtered Size	4,936 (58.5%)	5,824 (36.4%)	421 (6.76%)	974 (8.9%)	15,960 (26.6%)	1,689 (7.57%)

Table 4: Size comparison between the original datasets and the TFD-filtered datasets.

With 58.5%, most potentially relevant dia-

[6]Our compute infrastructure consists of one Tesla V100-SXM3 GPU (with 32 GB memory) and it takes 76 mins on avg. to run TFD with one dataset.

[7]See Appendix C for a sentence-level analysis. We also used TFD with just the feedback-indicating phrases (not the complete sentences) but found that they are not expressive enough due to their small length (see Section 4.1.1).

logues were identified in MWoZ (Zang et al., 2020). Only 26.6% of dialogues in SFC (Hancock et al., 2019) were identified as potentially containing textual user feedback. In case of PC (Zhang et al., 2018) and WoW (Dinan et al., 2018), TFD only identifies 8.9%, and 7.57% as containing such data, respectively.

To answer whether the investigated datasets are extendable with annotations for learning from textual user feedback, we annotate $80-100$ of the TFD-filtered dialogues for each investigated dataset (depending on availability)[8] for error types and user responses using our proposed error type and user response taxonomies (Section 4.3 and 4.2). We refer to these as TFD-filtered subsets hereinafter. Overall, they consist of 555 dialogues. To ensure that TFD does not bias our findings, we additionally analyse a second set of 600 randomly selected dialogues that were not identified by TFD (similarity $< 50\%$; 100 dialogues from each of the original datasets) for potentially containing textual user feedback in the same way. We refer to these as random subsets hereinafter. Overall, we annotate $1,155$ dialogues for error types and user responses.

5.1 Error Type Analysis

Table 5 shows the result of our error type annotation for both the TFD-filtered and the random subsets (in relation to the number of considered dialogues)[8].

	Task-Oriented			Open-Domain		Know.-Grounded
Dataset	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
TFD-Filtered Subsets	8/100	3/100	2/95	6/71	92/100	19/89
Random Subsets	2/100	0/100	5/100	2/100	43/100	3/100

Table 5: The number of errors identified in the TFD-filtered and random subsets in relation to the data considered for each dataset.

Overall, we find that the number of annotated errors in the TFD-filtered subsets is considerably higher than in the random subsets, especially in case of open-domain and knowledge-grounded dialogues, such as SFC (Hancock et al., 2019) and WoW (Dinan et al., 2018) (+49 in case of SFC and +17 in case of WoW).

[8] See Appendix F for details on sampling for the TFD-filtered subsets and a more detailed error type analysis.

Table 6 combines the shares of the most common error types across both the TFD-filtered and the random subsets.

	Task-Oriented			Open-Domain		Know.-Grounded
Dataset	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Ignore Question (E1)	0.13	0.0	**0.50**	0.16	**0.54**	0.0
Topic Trans. Error (E6)	0.0	0.0	0.0	0.16	0.37	0.23
Factually Incorrect (E5)	0.0	0.0	0.0	0.0	0.02	0.46
Ignore Expect. (E3)	0.25	**0.33**	**0.50**	0.0	0.02	0.05
Ignore Request (E2)	**0.38**	0.0	0.0	0.0	0.02	0.0
Lack of Sociality (E9)	0.0	0.0	0.0	**0.33**	0.02	0.0

Table 6: The share of the most common error types for both the TFD-filtered and the random subsets (combined).

In case of open-domain dialogues, *ignore question* (E1) and *topic transition error* (E6) are the most frequent error types. Especially in case of SFC (Hancock et al., 2019), we often find system utterances to be out-of-context. For PC (Zhang et al., 2018), we also often observe a *lack of sociality* (E9) in system utterances. In case of task-oriented dialogues, *ignore request* (E2) and *ignore expectation* (E3) are common error types. We often observe these errors when requests are only partially processed. We also find little variety in language and flow in these dialogues, regardless of the number of tasks reflected in the dataset[9]. In case of WoW (Dinan et al., 2018), the knowledge-grounded dataset, we mostly observe *factually incorrect* (E5) errors in system utterances.

5.2 User Response Type Analysis

Table 7 shows the annotation results for user responses to errors in system utterances (Section 5.1). T refers to the corresponding TFD-filtered subset and R to the respective random one.

As described in Section 4.2, UR2 (repeat or rephrase concern), UR3 (providing a correction), and UR5 (asking for clarification) are user responses likely to contain textual feedback. In case of the TFD-filtered subsets, we find that UR3 and UR5 are more often observed in open-domain and knowledge-grounded dialogues, such as WoW (Dinan et al., 2018) or

[9] See Appendix D for examples.

	Task-Oriented						Open-Domain				Know.-Grounded	
Dataset	MWoZ (HH)		SGD (HH)		BABI (HB)		PC (HH)		SFC (HB)		WoW (HH)	
	T	R	T	R	T	R	T	R	T	R	T	R
Errors	8	2	3	0	2	5	6	2	92	43	19	3
UR1	1	2	2	0	1	3	0	1	4	36	0	1
UR2	2	0	1	0	1	0	0	0	0	0	0	0
UR3	2	0	0	0	0	2	0	0	3	0	9	0
UR4	1	0	0	0	0	0	2	1	34	1	0	1
UR5	2	0	0	0	0	0	4	0	51	6	10	1

Table 7: User response types observed in the TFD-filtered (T) and the random (R) subsets.

SFC (Hancock et al., 2019). UR2 is only rarely observed, and only in task-oriented dialogues. Other user responses that are less likely to contain textual user feedback, i.e., UR1 and UR4 are also frequently observed, especially in case of SFC. In case of the random subsets, we find that errors are more often ignored by users (UR1), or are indicated by feedback phrases that are not represented in our set of feedback-indicating-sentences (see Section 4.1.1).

5.3 Analysis Of Cause And Effect

Figure 1 illustrates the relation between frequent errors (see Table 6) and user responses, i.e., which error type causes which user response.

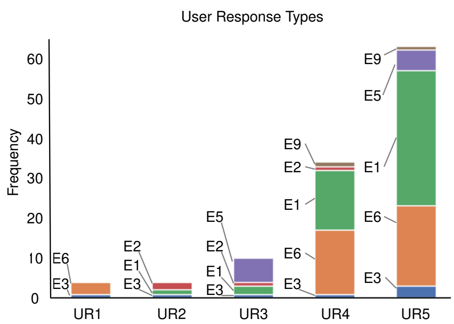

Figure 1: Illustration of the relation between frequent errors and user responses. The E-values represent the error types (see Section 4.3) and UR1-UR5 represent the user response types (see Section 4.2).

We observe UR5 as the most frequent user response type, e.g., when ignoring a user's question (E1) or unexpectedly changing the topic (E6). However, according to Table 6, those error types mostly occur in open-domain datasets, especially in SFC (Hancock et al., 2019) with a share of 0.54 and 0.37, respectively. Together with UR3, UR5 is also a frequent response type to E5 (factually incorrect), which is mostly observed in WoW (Dinan et al., 2018). UR2 is only rarely observed. It is sometimes used as response type to E2 (ignore request), and E3 (ignore expectation), which are mostly observed in task-oriented dialogues. However, the share of UR4 (which does not provide textual feedback, like UR1) is also high for most of the frequently observed error types. This mostly affects SFC (Hancock et al., 2019).

5.4 Collect New Datasets Or Extend Existing Ones?

Based on the insights into error types (Section 5.1), user responses (Section 5.2), and relations between them (Section 5.3), we find that every dialogue type has different error and user response types. In case of task-oriented dialogues, errors are few. Accordingly, there is only little textual user feedback available that could be used for learning. For this reason, it might be hard and ineffective to make these datasets available for learning from textual user feedback. In our view, they are not suitable for this purpose. In contrast, open-domain and knowledge-grounded dialogues contain a higher number of errors and user responses that are likely to contain textual user feedback. For this reason, it might be possible (and valuable) to extend these datasets with the needed annotations to make them available for learning from textual user feedback.

5.5 On The Effectiveness of Textual Feedback Detection

Combining the insights on (1) the impact of TFD on dataset sizes (Table 4), (2) the number of identified errors in TFD-filtered and random subsets (Table 5), and (3) the correspondingly identified user response types (Table 7), we find that the dialogues that are selected by TFD contain more textual user feedback (user responses that address errors in system utterances) compared to those selected randomly. Furthermore, the comparison with user responses found in the random subsets (Table 7) shows that TFD only rarely misses relevant user responses, i.e., user responses that are likely to contain textual user feedback, such as UR2, UR3 or UR5 (Section 4.2). If TFD misses such user responses, we find that the corresponding feedback-indicating phrases are not represented in our set of feedback-indicating sentences (Section 5.2).

To improve the effectiveness of TFD in future work, we suggest to investigate the impact of considering more context, e.g., the dialogue history or at least the previous system utterance. In this regard, it might also be interesting to increase the number and variety of feedback-indicating sentences. Another valuable approach might be to finetune TFD on a small set of annotated dialogues, i.e., in a few-shot setup, and to add a classification layer on top. This way, TFD could also be enabled to directly annotate system utterances for containing errors and user utterances for potentially containing textual feedback.

5.6 Inter-Annotator Agreement

To get a better understanding of our results, we asked nine experts (all non-native speakers, but with sound English skills and NLP background) to annotate smaller subsets of the TFD-filtered and the random subsets (300 from each, 50 from each of the investigated datasets, 600 overall) for error types and user responses[10]. Each of these smaller subsets was assigned to two of these experts (including our annotation, each of these dialogues was annotated three times). For calculating the Inter-Annotator Agreement (IAA), we use Krippendorff's Alpha (Krippendorff, 2004)[11]. Table 8 shows the result[12].

		Task-Oriented			Open-Domain		Know.-Grounded
Dataset		MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Error Type	TFD-Filtered	0.01	0.0	1.0	0.51	0.81	0.12
	Random	0.55	0.01	-0.01	0.09	0.80	0.02
User Res. Type	TFD-Filtered	0.04	0.0	0.23	0.16	0.72	0.04
	Random	0.05	0.0	0.0	0.01	0.79	-0.02

Table 8: Inter-Annotator Agreement calculated using Krippendorff's Alpha.

While the agreement on error types is comparatively high in case of open-domain dialogues, it is rather low for task-oriented dialogues with the exception of BABI (Bordes et al., 2016). In contrast to BABI, MWoZ (Zang et al., 2020) and SGD (Rastogi et al., 2020) consist of human-human dialogues. We find that errors are hard to identify in these dialogues, as humans rather suggest disagreements in a very polite way instead of accusing the partner of a mistake[13]. In case of human-bot dialogues, we rather observe humans react harshly and accusing to errors in system utterances, resulting in more direct feedback. This is reflected by the annotator's feedback and in the inter-annotator agreement for those datasets. Since the annotations of the error type have an impact on the annotations of the user response types, i.e., whether an error was identified or not, the agreements here are also lower for most datasets[14].

6 Conclusion

In this work, we investigated the extendability of existing dialogue datasets from various types with annotations for learning from textual user feedback, i.e., user responses that address errors in system utterances. We focused on task-oriented datasets as there are no appropriately annotated datasets available. With Textual Feedback Detection (TFD), we propose a semi-automatic approach to filter dialogues for potentially containing textual user feedback. Furthermore, we propose two taxonomies optimized to categorize such data, a user response type taxonomy and an error type taxonomy. In our study, we annotate 1,155 dialogues from six different dialogue datasets with both errors and corresponding user responses. In case of task-oriented dialogues, we find that errors are few. Accordingly, there is only little textual user feedback available that could be used for learning. Therefore, they are hardly extendable with the needed annotations and not suitable for this purpose. In contrast, the investigated open-domain and knowledge-grounded dialogue datasets contain a higher number of errors and user responses that are likely to contain textual user feedback. Therefore, they are more appropriate to be extended for learning from textual user feedback.

7 Limitations

With Textual User Feedback (TFD), we propose an approach to identify potential textual

[10]See Appendix G for more background on participating annotators and Appendix J for the annotation guidelines.

[11]We use the Python library *annotation_ analysis* for this: `https://github.com/ai-nikolai/annotation_analysis`, last accessed on 11/22/22.

[12]See Appendix H for a detailed analysis including class distributions.

[13]See Appendix E for examples.

[14]We also calculated the inter-annotator agreement using the taxonomy of Higashinaka et al. (2021) (see Appendix I). It further deteriorates the agreement.

user feedback (user responses that address errors in system utterances) by exploiting the semantic similarity between user responses and feedback-indicating sentences. Even when our analysis shows that it does not miss a significant amount of textual user feedback, taking more context into account, e.g., the dialogue history or at least the previous system utterance, might improve the hit rate or result in more complex textual user feedback, such as corrections that targets errors from multiple turns ago.

Regarding dataset selection, our study (and result) has only limited expressiveness for knowledge-grounded dialogue datasets. Due to limited availability, we only consider one of such datasets in our study, Wizards-of-Wikipedia (Dinan et al., 2018). However, this does not affect the relevance of our work, as there are already feedback-annotated datasets available for similar dialogue types, e.g., FITS (Xu et al., 2022).

The majority of our evaluation was done manually. Therefore, with respect to the original dataset sizes, we only consider a small fraction of the data in our study. This might have a negative impact on the identified feedback-indicating sentences. Our results might have been clearer when we would have considered more dialogues for feedback-indicating sentences collection. This way, it might also have been possible to identify other (or more) user response types that possibly contain textual user feedback (and causing errors) as a result.

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents.

Laila Dybkjaer, Niels Ole Bernsen, and Hans Dybkjaer. 1996. Grice incorporated: Cooperativity in spoken dialogue. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Reliability in content analysis. *Human Communication Research*, 30(3):411–433.

Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. 2019. Lifelong and interactive learning of factual knowledge in dialogues. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 21–31, Stockholm, Sweden. Association for Computational Linguistics.

Sebastian Möller, Klaus-Peter Engelbrecht, and Antti Oulasvirta. 2007. Analysis of communication failures for spoken dialogue systems. In *Eighth Annual Conference of the International Speech Communication Association*.

Sunghyun Park, Han Li, Ameen Patel, Sidharth Mudgal, Sungjin Lee, Young-Bum Kim, Spyros Matsoukas, and Ruhi Sarikaya. 2021. A scalable framework for learning from implicit user feedback to improve natural language understanding in large-scale conversational AI systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset.

Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. SaFeRDialogues: Taking feedback gracefully after conversational safety failures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6462–6481, Dublin, Ireland. Association for Computational Linguistics.

Mathilde Veron, Sophie Rosset, Olivier Galibert, and Guillaume Bernard. 2021. Evaluate on-the-job learning dialogue systems and a case study for natural language understanding.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2022. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. 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.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A The Integrated Error Taxonomy – Details

In this section, we provide descriptions on the integrated error taxonomy proposed by Higashinaka et al. (2021). In principle, this taxonomy differentiates between *form violation* and *content violation*. The form violation usually represents errors that oppose some kind of meta criteria, e.g., the form of language or the ignorance of social norms. In contrast, content violations refer to, e.g., inconsistency or redundant utterances, or other things that might cause a dialogue breakdown. Content violation is hereinafter abbreviated by *CV*. Form Violation by *FV*. Furthermore, while the original work always refers to *utterances* in general, we refer to *system utterance*, since this is evident from the examples in the original paper and in our view simplifies understanding.

A.1 Utterance-Level

Utterance-level Errors typically expose language-generation deficiencies of the system.

- **Uninterpretable (I1), FV** – The system's utterance does not contain recognizable words or is just a text fragment, e.g., the utterance just consists of the words *of yours*.
- **Grammatical Error (I2), FV** – The system's utterance is grammatical incorrect, e.g., *he will attend yesterday*.
- **Semantic Error (I3), CV** – The combination of predicates and arguments does not result in any meaning, e.g., *I feel painting*.
- **Wrong Information (I4), CV** – The given information is clearly wrong, e.g., *Meat Loaf was the songwriter of Jim Steinman* (it was vice versa).

A.2 Response-Level

Response-level errors describe errors in system utterances, e.g., in responses to user questions or requests.

- **Ignore Question (I5), FV** – The system utterance ignores the user's question, e.g., *U: What will the weather be like tomorrow? S: Look at these beautiful flowers...*

- **Ignore Request (I6), FV** – The system ignores the user's request to do something, e.g., *U: Please book me a room in the west part of town. S: I like sports.*

- **Ignore Proposal (I7), FV** – The system ignores the user's proposal to do something, e.g., *U: Let's talk about movies. S: I like pizza.*

- **Ignore Greeting (I8), FV** – The system ignores the user's greeting, e.g., *U: Hello, my friend! S: What do you want from me?*

- **Ignore Expectation (I9), CV** – The system utterance does not have the expected content but the user's intention has been successfully conveyed, e.g., *U: Do you have a favorite football team? S: Yes.* (the system should tell the user the name of his favorite football team).

A.3 Context-Level

Context-level errors refer not only to the local context, i.e., adjacent pairs of user utterance and system utterance/response, but to a broader (sometimes global) context.

- **Unclear Intention (I10), FV** – The system utterance is on the current topic, but the intention is unclear, e.g., *U: What is your favorite movie? S: I like Daddy's Home 2. U: Oh, yes! Mel Gibson is so good in this role... S: The Evil Dead is so bloody...*

- **Topic Transition Error (I11), FV** – The system suddenly switches to another topic without explanation, e.g., *U: I really enjoyed Feynman's physics lecture. So enlightening! S: Well, what is your favorite food?*

- **Lack of Information (I12), FV** – The system utterance misses important information, such as subject, object, or modifier, e.g., *U: It's a beautiful day, isn't it? S: I hate.* (object is missing).

- **Self-Contradiction (I13), CV** – The system utterance contradicts what it said before, e.g., *S: I'm from Boston, Massachusetts. U: Oh, wow! I'm from Bartlesville, Oklahoma! S: Really? I'm from Seattle, Washington!*

- **Contradiction (I14), CV** – The system utterance contradicts what has been said by the user, e.g., *U: I like strawberries. S: Oh, why don't you like strawberries?*

- **Repetition (I15), CV** – The system repeats something it has already said before (this includes repeated questioning), e.g., *S: I look forward to the concert. U: Oh, which concert? S: I look forward to the concert.*

A.4 Society-Level

Society-level errors refer to errors that indicate that the system lacks etiquette or common sense.

- **Lack of Sociality (I16), FV** – The system utterance lacks consideration towards the user and contains hate speech or is in general disrespectful, e.g., *U: Yesterday I went to the federal employment agency. S: Okay. Are you incapable of finding a job on your own?*

- **Lack of Common Sense (I17), CV** – The system utterance contains content that differs from the opinion of the majority, e.g., *S: I like to eat chicken raw. U: Oh, dear... You have to cook it first!*

B Feedback-Indicating Sentences And Phrases

In this section, we present the collected feedback-indicating sentences along with phrases (without any grammar or spelling corrections).

1. Not really like fandoms, haha Just anything online that people make. (**Phrase**: *not really like*)

2. It is not saturday. (**Phrase**: *it is not*)

3. That doesn't make sense. (**Phrase**: *doesn't make sense*)

4. That makes no sense. (**Phrase**: *makes no sense*)
5. You should put some more things together." (**Phrase**: *you should*)
6. You shouldn't be! (**Phrase**: *you shouldn't*)
7. What do you mean by that?" (**Phrase**: *what do you mean*)
8. What are you talking about? (**Phrase**: *what are you talking about*)
9. It's so important for young people to have diverse interest and develop a wide range of skills, don't you think? (**Phrase**: *don't you think*)
10. I don't know what you're talking about. (**Phrase**: *don't know*)
11. What does that have to do with computer games? (**Phrase**: *what does that have to do with*)
12. Sorry I meant to say for the cat litter. (**Phrase**: *sorry i meant to say*)
13. That didn't have anything to do with school. (**Phrase**: *didn't have anything to do with*)
14. You do not make sense with your response. (**Phrase**: *your response*)
15. That's not what I asked you. (**Phrase**: *not what i asked*)
16. I dont understand. (**Phrase**: *don't understand*)
17. How do you mean? (**Phrase**: *how do you mean*)
18. I don't care about price. (**Phrase**: *i don't care*)
19. You're not answering the questions. (**Phrase**: *you're not answering*)
20. Like I said before I'm not one to read an actual newspaper but I do like reading opinion and political articles. (**Phrase**: *like i said before*)
21. You're not very helpful Help Desk. (**Phrase**: *not very helpful*)
22. Are you sure that there are no hotels on the west side of town? (**Phrase**: *are you sure*)
23. I didn't say anything was scary. (**Phrase**: *i didn't say*)
24. I wouldn't know this. (**Phrase**: *i wouldn't know this*)
25. That sounds too low. (**Phrase**: *too low*)
26. I'm great, but thats off topic. (**Phrase**: *that's off topic*)
27. No, I think when people shape their beards in different ways is really interesting as well! (**Phrase**: *no, I think*)
28. Your doing it wrong my friend. (**Phrase**: *you're doing it wrong*)
29. What are you saying? (**Phrase**: *what are you saying*)
30. At least you have that then. (**Phrase**: *at least you have*)
31. That doesn't answer my question. (**Phrase**: *that doesn't answer*)
32. I am too old to hike I am in my seventies. (**Phrase**: *i am too old*)
33. You aren't staying on topic at all. (**Phrase**: *not staying on topic*)
34. Off the subject, I am thinking of cutting my hair. (**Phrase**: *off the subject*)
35. I'm not ready to book just yet. (**Phrase**: *i'm not ready*)
36. That's not what I asked you. (**Phrase**: *i asked you*)
37. Dude not cool. (**Phrase**: *dude not cool*)
38. I'd really like a 4 star. (**Phrase**: *i'd really like*)
39. Thats nonsense." (**Phrase**: *thats nonsense*)

40. Actually, I apologize no need to book, I was just gathering information." (**Phrase**: *i apologize*)

41. I never said I needed one. (**Phrase**: *i never said i*)

42. No I dont think so. (**Phrase**: *no i dont think*)

43. I didn't mention anything about clowns. (**Phrase**: *i didn't mention*)

44. That is odd for alaska. (**Phrase**: *that is odd*)

45. Not sure what that means? (**Phrase**: *not sure what that means*)

46. It can be what? (**Phrase**: *it can be what*)

47. You should learn! (**Phrase**: *you should learn*)

48. Umm, what? (**Phrase**: *umm, what*)

49. You think so? (**Phrase**: *you think so*)

50. No a park is a place and not a person, (**Phrase**: *and not*)

51. Why do you say that? (**Phrase**: *why do you say that*)

52. I guess I should have asked that first. (**Phrase**: *i should have asked*)

53. I said lets talk about sports. (**Phrase**: *i said lets talk about*)

54. You're being annoying is whats happening. (**Phrase**: *you're being annoying*)

55. You could have stated the goods. (**Phrase**: *you could have stated*)

56. Who was talking about color? (**Phrase**: *who was talking about*)

57. That doesn't really matter. (**Phrase**: *doesn't really matter*)

58. It's actually a 1939 movie that was adapted from a novel written earlier. (**Phrase**: *it's actually*)

59. I don't believe a piano is a stringed instrument. (**Phrase**: *i don't believe*)

60. That's not relevant. (**Phrase**: *that's not relevant*)

61. Check again. (**Phrase**: *check again*)

62. You're wrong. (**Phrase**: *you're wrong*)

63. That doesn't have to do with track. (**Phrase**: *that doesn't have to do with*)

64. Instead could it be in Madrid? (**Phrase**: *instead could it*)

65. I would prefer in Bombay. (**Phrase**: *i would prefer*)

66. No, I don't like that. (**Phrase**: *i don't like that*)

67. No, this does not work for me. (**Phrase**: *this does not work*)

C Error Distribution – Sentence-Level Analysis

As described in Section 4.1.2, TFD works on sentence-level. We decompose every dialogue into turns, extract the user utterances, and then decompose every user utterance into sentences. Then, we pair each of these sentences with each of the feedback-indicating sentences for calculating the semantic similarity. Table 9 shows the impact on TFD on dataset sizes on sentence-level, i.e., the number of sentences from all collected user utterances before (*Sentences (Before)*), and the number of sentences after (*Sentences (After)*) applying TFD.

	Task-Oriented			**Open-Domain**		**Know.-Grounded**
Dataset	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
#Sentences (Before)	103,029	296,808	192,485	251,985	190,691	165,750
#Sentences (After)	16,545 (16.0%)	19,033 (6.4%)	26,939 (14.0%)	1,474 (0.0%)	25,711 (13.5%)	2,358 (1.4%)

Table 9: Size comparison of datasets before and after applying TFD (on sentence-level).

Figure 2 illustrates the distribution of *Sentences (After)* across datasets with regard to similarity ranges, i.e., $50\% - 60\%$, $60\% - 70\%$, $70\% - 80\%$,$80\% - 90\%$, $90\% - 100\%$. It reflects the share in identified phrases from each of the datasets (see Table 2). Most of the phrases were identified in SFC (Hancock et al., 2019). Only a small amount of phrases came from the other

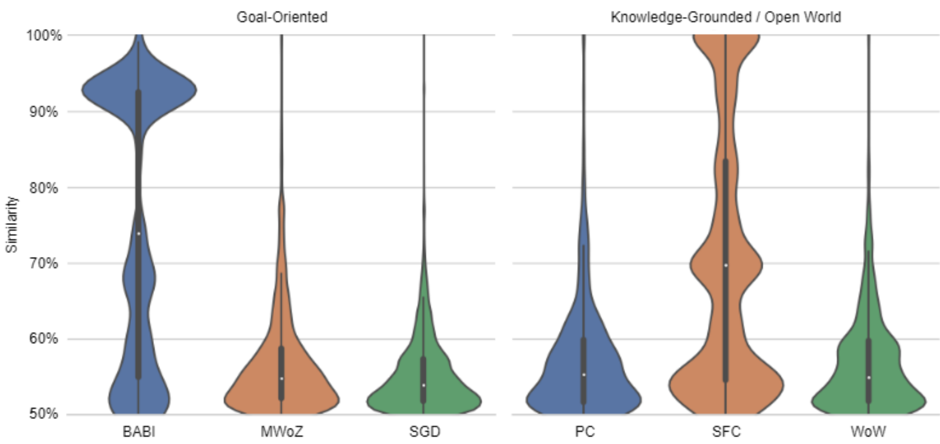

Figure 2: Illustration of the distribution of dialogues with potential real-world error scenarios across datasets.

datasets which might be the reason for the clusters in the lower ranges. However, according to our analysis in Section 5.2, we don't miss a significant amount of textual user feedback (User responses of type UR2, UR3 or UR5, see also Section 4.2). Nevertheless, BABI (Bordes et al., 2016) is a clear outlier in this case. We observe only little variety in language or flow in this dataset, what we suspect as reason for this.

D Task-Oriented Dialogues – Examples

In Section 5.1, we state that the dialogues of task-oriented datasets considered in our work are of little variety in language and flow. This is most obvious in case of BABI (Bordes et al., 2016) which we want to explain using the following example:

1. **USER:** good morning
1. **SYSTEM:** hello what can i help you with today
2. **USER:** may i have a table with british cuisine in a cheap price range
2. **SYSTEM:** i'm on it. where should it be
3. **USER:** madrid please
3. **SYSTEM:** how many people would be in your party
4. **USER:** for eight please
4. **SYSTEM:** ok let me look into some options for you
5. **USER:** actually i would prefer in paris
5. **SYSTEM:** sure is there anything else to update
6. **USER:** instead could it be for four people
6. **SYSTEM:** sure is there anything else to update
7. **USER:** instead could it be in a moderate price range
7. **SYSTEM:** sure is there anything else to update
8. **USER:** no
8. **SYSTEM:** ok let me look into some options for you

This dialogue is prototypical for BABI. The first few turns (the first four in this case) contain the original user request. Then, the user asks for changes (turn five to eight). Overall, the dialogues are very scripted. Users are mostly clear and distinctive in their responses. They contain only little variety in language and flow, which makes them seem very unrealistic. We don't observe a change in this regard across tasks. The following dialogue is from SGD (Rastogi et al., 2020):

1. **USER:** Our family is planning for a vacation trip. Please can you find a good rental car for a pleasant trip?

1. **SYSTEM:** At what time do you need the car?

2. **USER:** I can pick up the car at 15:00

2. **SYSTEM:** From when onwards you are looking for the rental car?

3. **USER:** I need the car from 11th of March.

3. **SYSTEM:** So, how long you need the car?

4. **USER:** I might need it till the 13th of this month.

4. **SYSTEM:** Okay! So, where would be your pick up location?

5. **USER:** I can pick up the car in Paris on 9th of this month.

5. **SYSTEM:** Okay! There is a nice car, Chevrolet Cruze, a Standard car is available for pick up on March 9th at Charles de Gaulle International Airport. Do you wish to rent this car?

6. **USER:** Do you have any other cars for I need the car till 10th of this month.

6. **SYSTEM:** Yup! We've 3 more cars. Do you wish to rent Honda Accord, which is a Standard car and its ready for pickup on March 9th from Charles de Gaulle International Airport?

7. **USER:** Okay. I go for the Honda Accord.

The dialogue looks much more natural, which might be due to its type. While BABI consists of human-bot dialogues, SGD consists of human-human dialogues. However, the flow is very similar. Up until turn five, the user's request is refined. Then, the system comes up with a suggestion which the user denies (for a couple of turns) by asking for something different. At some point, user's agree or disagree on the system's suggestions. Then the dialogue ends. The following dialogue is from MWoZ (Zang et al., 2020):

1. **USER:** Hi, I could use help with my planning. I am trying to find a place to stay.

1. **SYSTEM:** Ok, to help us find the right place for you lets talk about what you would like. Do you have a price range you prefer? Or an area of town?

2. **USER:** I would like it to be in the south and have a rating of 3 stars.

2. **SYSTEM:** We have a moderately priced guesthouse and an expensive hotel that match your requirements.

3. **USER:** Does either offer free parking?

3. **SYSTEM:** They both offer free parking and internet.

4. **USER:** Okay can you please book it for 7 people for 4 nights starting on Tuesday?

4. **SYSTEM:** Which one? The guesthouse or the hotel?

5. **USER:** The hotel, please.

5. **SYSTEM:** Reference number is : BS0ODSGJ. Is there anything else?

6. **USER:** Yes, I also need a train from stevenage to cambridge. I want to leave after 10:15 on tuesday.

6. **SYSTEM:** I have train TR0517 that leaves at 11:54. Would that be something that would work for you?

7. **USER:** That's perfect. Could I get the train ID and travel time?

7. **SYSTEM:** The travel time is 49 minutes.

8. **USER:** Great, that's all I need, thanks so much for your help! Have a great day!

8. **SYSTEM:** You're very welcome! Thank you! Bye!

Like in case of SGD, the dialogue looks very natural. However, in contrast to BABI and SGD, the dialogues consist less of refinements, but more of multiple tasks as shown in this example. First, the user asks for a hotel in a city, then he asks for transportation to that city. Sometimes, he also asks for locations with entertainment. While the first request is usually completed (the booking of a room in this case), the second request is usually about gathering information (the user just asks for the train number, departure and travel time, but not for booking a seat).

E Human-Human Dialogues – Examples

We find that it is hard to identify errors in human-human dialogues, as the tone is in principal mostly polite, even if one of the participants is wrong. The following dialogue is an example from WoW (Dinan et al., 2018):

1. **USER:** I adore photography even though I'm bad at it. It was never a good idea to give me access to a DSLR camera...

1. **SYSTEM:** Well, photography is a science. A little practice and you'd get it down.

2. **USER:** You think so? I think most people, myself included, would consider in an art. But I suppose it's true that there is a lot of technical detail to it.

2. **SYSTEM:** I personally do it as a hobby but many do in businesses, film ect. In my opinion it's always art. i love it because it can capture what your heart see's. If you angle it just right

In this case, the system claims that photography is science. However, it's commonly considered as one kind of visual art. Instead of directly correcting the system, the user rather suggests that the system might be wrong. The following example is from PC (Zhang et al., 2018):

1. **USER:** Hi, how it is going today?

1. **SYSTEM:** It is going great today; everyday is great in my ford truck.

2. **USER:** Oh. I am eating a piece of fried chicken.

2. **SYSTEM:** Right on. I do not allow fried chicken in my ford truck.

3. **USER:** What do you do for a living?

3. **SYSTEM:** I work on ford trucks. Do you have a ford truck?

4. **USER:** No , i drive a toyota.

4. **SYSTEM:** I am so sorry. You should get yourself a ford truck. Mine is black .

5. **USER:** I am not. I love my toyota .

5. **SYSTEM:** Well okay. How about a ford car? They are just as good.

6. **USER:** Nope, hardcore toyota family. My wife drives a 4runner.

6. **SYSTEM:** Why do they call it a 4runner?

7. **USER:** Because it will run forever.

7. **SYSTEM:** Oh, I thought it was because 4 wheeler was taken.

8. **USER:** Oh. That does not make much sense.

8. **SYSTEM:** Are you kidding me? Do you know any jokes?

In the second to last turn, the system makes a job that the user does not understand. The user's response is quite polite. He informs the system that his statement does not make much sense to him.

F Effectiveness Of Textual Feedback Detection – A Detailed Analysis

The TFD-filtered subsets consist of 20 dialogues from the feedback-filtered subsets of the investigated datasets from each similarity range, i.e., $50\% - 60\%$, $60\% - 70\%$, $70\% - 80\%$, $80\% - 90\%$, $90\% - 100\%$ (if available). As the data in the upper ranges ($80\% - 100\%$) is scarce in case of WoW (Dinan et al., 2018), PC (Zhang et al., 2018), and BABI (Bordes et al., 2016), the TFD-filtered dataset consists only of 555 dialogues (instead of 600 with 100 dialogues from each feedback-filtered subset which is the case for the cold dataset). Table 10 shows the results of our error type analysis with respect to similarity ranges identified by TFD, i.e., $50\% - 60\%$, $60\% - 70\%$, $70\% - 80\%$, $80\% - 90\%$, $90\% - 100\%$, meaning that each dialogue contains at least one utterance with a sentence identified to be similar to at least one error-indicating sentence in this similarity range. *Overall* (O) represents the number of randomly sampled dialogues from the respective range, and *Error* (E) represents the number of dialogues identified

Dataset		MWoZ (HH)		SGD (HH)		BABI (HB)		PC (HH)		SFC (HB)		WoW (HH)	
		Task-Oriented						Open-Domain				Know.-Grounded	
Overall / Error		O	E	O	E	O	E	O	E	O	E	O	E
TFD-Filtered Subsets	90% - 100%	20	2	20	2	17	0	6	2	20	20	9	4
	80% - 90%	20	2	20	1	18	0	5	2	20	20	15	9
	70% - 80%	20	1	20	0	20	0	20	0	20	19	20	4
	60% - 70%	20	1	20	0	20	2	20	1	20	18	20	2
	50% - 60%	20	2	20	0	20	0	20	1	20	15	20	0
	Overall	100	8	100	3	95	2	71	6	100	92	89	19
Random Subsets		100	2	100	0	100	5	100	2	100	43	100	3

Table 10: Identified errors in all datasets across similarity ranges.

in our manual analysis to contain an error in an system utterance.

Overall, only 55 dialogues of the random subsets (9.2%) contain errors. In case of TFD, we observe 130 of such cases. Therefore, TFD shows to facilitate the process of textual user feedback identification. Even if the number of identified errors is overall low, most errors are identified in the range of $60\%-100\%$, excluding the densest section in case of MWoZ (Zang et al., 2020), SGD (Rastogi et al., 2020), PC and WoW, $50\% - 60\%$.

G Human Annotators

All additional annotators that participated in this study were experts from our lab. We did not select them according to specific criteria; they participated voluntarily. Accordingly, they were not paid extra for this, since they did the annotations during their working hours. All were non-native speakers, but with good English skills and NLP background. For annotation, we did not use any specific tool. We provided the annotators with dialogues in json format and asked them to do their annotations directly in the respective files.

H Inter-Annotator Agreement – Detailed Analysis

This section gives more insights on the inter-annotator agreement by presenting the error type and user response type class distributions. Table 11 shows the error type distribution across the TFD-filtered subsets.

In case of SFC (Hancock et al., 2019), the open-domain human-bot dataset, the distribution is comparatively dense, i.e., annotators mostly agree on error types. This is also the case with BABI (Bordes et al., 2016), the task-oriented human-bot dataset. In contrast, in case of human-human dataset, the distribu-

	Task-Oriented									Open-Domain						Know.-Grounded		
	MWoZ (HH)			SGD (HH)			BABI (HB)			PC (HH)			SFC (HB)			WoW (HH)		
Annotator	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
Ignore Question (E1)	-	1	1	-	-	1	1	1	1	1	1	2	23	23	23	-	-	1
Ignore Request (E2)	1	3	3	-	-	-	-	-	-	-	-	-	4	6	4	-	-	-
Ignore Expect. (E3)	1	3	3	1	-	1	1	1	1	-	-	-	2	1	-	-	-	-
Slot Error (E4)	1	-	-	-	-	-	-	-	1	-	1	2	1	1	-	-	4	1
Factually Incorrect (E5)	-	1	-	-	-	-	-	-	-	-	2	-	-	-	-	4	-	1
Topic Trans. Error (E6)	-	-	-	-	-	-	-	-	-	2	2	4	20	18	21	3	-	-
Convers. (E7)	-	-	1	-	-	1	-	-	-	1	-	-	-	-	-	-	-	2
Unclear Intention (E8)	-	-	-	-	-	-	-	-	-	-	-	1	-	2	2	-	1	-
Lack of Sociality (E9)	-	-	-	-	-	-	-	-	-	1	-	2	1	-	1	-	-	-
Lack of Com. Sense (E10)	-	-	-	1	-	-	-	-	-	1	1	1	-	-	-	3	-	-

Table 11: Distribution of error types in the TFD-filtered subsets.

tions are widely spread. We suspect that this is because errors in these datasets are more difficult to identify, as humans rather suggest disagreements than directly emphasizing errors (see Section 5.6. This might be the reason for the low agreement and the high disagreement in these datasets (see Table 8).

	Task-Oriented									Open-Domain						Know.-Grounded		
	MWoZ (HH)			SGD (HH)			BABI (HB)			PC (HH)			SFC (HB)			WoW (HH)		
Annotator	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
Ignore Question (E1)	-	1	1	-	-	-	-	1	-	-	1	3	5	2	5	-	1	2
Ignore Request (E2)	-	-	-	-	-	-	-	-	-	-	-	-	2	3	2	-	-	-
Ignore Expect. (E3)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Slot Error (E4)	-	1	-	-	-	-	-	-	-	-	1	1	-	2	-	-	1	-
Factually Incorrect (E5)	-	1	-	-	-	-	-	4	-	-	-	-	-	1	-	-	1	-
Topic Trans. Error (E6)	-	-	1	-	-	-	-	-	-	-	-	6	16	14	11	-	-	1
Convers. (E7)	-	-	1	-	-	-	-	-	-	-	1	1	-	-	2	1	-	1
Unclear Intention (E8)	-	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	-
Lack of Sociality (E9)	-	-	-	-	-	-	-	-	-	-	1	2	-	2	-	-	-	-
Lack of Com. Sense (E10)	-	-	1	-	-	-	-	-	-	1	1	1	-	-	-	1	-	1

Table 12: Distribution of error types in the random subsets.

Table 12 shows the error type distribution in the random subsets. However, the situation is rather similar to the TFD-filtered subsets.

	Task-Oriented									Open-Domain						Know.-Grounded		
	MWoZ (HH)			SGD (HH)			BABI (HB)			PC (HH)			SFC (HB)			WoW (HH)		
Annotator	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
UR1	-	2	4	1	-	3	1	-	1	1	2	10	6	8	6	-	5	1
UR2	-	2	3	-	-	-	1	1	1	-	-	-	-	-	-	-	-	-
UR3	1	3	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1
UR4	2	1	-	-	-	-	-	1	1	2	1	2	18	16	13	5	-	-
UR5	-	-	-	-	-	-	-	-	-	3	4	-	27	27	32	6	-	3

Table 13: Distribution of user response types in the TFD-filtered subsets.

Table 13 shows the distribution of user response types in the TFD-filtered subsets. It basically reflects the findings for the error types. The same applies to the distribution of user response types in the random subsets (see Table 14).

	Task-Oriented									Open-Domain						Know.-Grounded		
	MWoZ (HH)			SGD (HH)			BABI (HB)			PC (HH)			SFC (HB)			WoW (HH)		
Annotator	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
UR1		6	1					2				13	16	15	13	1	10	3
UR2		7	-					2				1		1				
UR3			2					1						1			1	2
UR4		2	1							1	4		2	3	2	1		
UR5											1		5	4	5		4	

Table 14: Distribution of user response types in the random subsets.

I Inter-Annotator Agreement Using The Integrated Error Taxonomy

To make sure that our error taxonomy is no source of error, we mapped the annotations from all annotators to the integrated taxonomy of Higashinaka et al. (2021). Table 15 shows the results.

		Task-Oriented			Open-Domain		Know.-Grounded
Dataset		MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Error Type	TFD-Filtered	-0.10 (-0.11)	0.0 (-0.0)	1.0 (-0.0)	0.26 (-0.25)	0.80 (-0.01)	-0.09 (-0.21)
	Random	0.55 (-0.0)	0.01 (-0.0)	-0.01 (-0.0)	0.09 (-0.0)	0.80 (-0.01)	0.0 (-0.02)

Table 15: Inter-Annotator-Agreement when using the error taxonomy as proposed by Higashinaka et al. (2021). It deteriorates the agreement on the TFD-filtered subsets (numbers in brackets).

Using this taxonomy deteriorates the inter-annotator agreement. This is most obvious in case of MWoZ (Zang et al., 2020), PC (Zhang et al., 2018), and SFC (Hancock et al., 2019). It has only little or no effect on the other datasets (the changed error types are never or only rarely observed in these datasets). This also shows the effectiveness of our proposed taxonomy for identifying errors in system utterances.

J Annotation Guidelines

J.1 Annotation Task

Learning from textual user feedback, i.e., free-form text feedback that expresses user satisfaction/dissatisfaction, describes new knowledge (new concepts), provides corrections or alternative responses, is increasingly important to continuously improve dialogue systems. As of today, only few datasets provide annotation for such data. However, there are various dialogue datasets publicly available – from a variety of domains and for all kinds of tasks. As an annotator, you are provided with three different taxonomies, an error taxonomy, an error type taxonomy, and a user reaction type taxonomy, and two different types of dialogues, *conspicuous dialogues* and *cold dialogues*. Your task is to annotate the dialogues using the given taxonomies. The conspicuous dialogues are identified to contain conspicuous sentences in user utterance that express user dissatisfaction as a result of an error in the previous system response. You have to decide (1) whether or not this conspicuous sentence correctly emphasizes an error in the previous system response, (2) which error type is represented in the previous system response, and (3) how the user reacted. The cold dialogues do not contain such information. In this case, your task is to identify error situations and user reactions in the dialogues just given the taxonomies.

J.2 Dialogue Format

Here's an example for a conspicuous dialogue:


```
{
    "unique_id": "PMUL0121.json_multiwoz_train",
    "id": "PMUL0121.json",
    "turns": [
        {
            "utterance": "Hi there! Can you give me some info on Cityroomz?",
            "response": "Cityroomz is located at Sleeperz Hotel, Station Road
                in the Centre area. It is a moderatley priced hotel with a 0
                star rating. Would you like to book a room?"
        },
        {
            "utterance": "Yes please. I need it for 7 people for 3 nights
                starting on tuesday.",
            "response": "How many days would you like to book it for?"
        },
        {
            "utterance": "I would like to book it for 3 nights beginning on
                Tuesday.",
            "response": "Ok, Your rooms have been booked. Your reference number
                is CYJDAR7R . Is there anything else I can help you with today?"
        },
        {
            "utterance": "Yeah, I need a restaurant in the west and with
                expensive pricing.",
            "response": "There are 9 expensive restaurants in the west area. Do
                you prefer a certain type of food?"
        },
        {
            "utterance": "No, I don't have a preference, but I need the phone
                number and postcode.",
            "response": "I recommend Graffiti, the number is 01223277977 and
```

```
                the postcode is cb30lx. Would you like me to book a table for you?"
        },
        {
            "utterance": "No, right now, I just need the infor. Thanks!",
            "response": "Can I help you with anything else today?"
        },
        {
            "utterance": "That is all. thanks",
            "response": "Thank you for using our system."
        }
    ],
    "conspicuous": [
        {
            "phrase": [
                "actually, i apologize no need to book, i was just gathering
                    information.",
                " i would like to book it for 3 nights beginning on tuesday."
            ],
            "confidence": 0.593,
            "turn": 2
        }
    ],
    "annotations": [
        {
            "turn": 2,
            "annotation": {
                "error_type": "E2",
                "comment": "the system misses intent/slots. the user already
                    said that he need it for three nights",
                "error": "C1",
                "user_reaction": "B3"
            }
        }
    ]
}
```

Each dialogue consists of an *unique id* , an *id*, and its *turns*. *utterance* is the user input and *response* is what we refer to as *system response*. *conspicuous* is an array. The first value is an *error-indicating phrase*, a phrase that was identified to express user dissatisfaction in the utterance of the corresponding *turn*. The second value is the value from an utterance of this dialogue that was identified to be similar to this error-indicating sentence. *confidence* represents the similarity. Dialogues with multiple *conspicuous* values are possible. The *annotations* list has an entry for each conspicuous phrase. Please add your annotations here. In *comment*, you can share your thoughts with us.

Here's an example for an cold dialogue:

```
[
    {
        "dialogue": "__p2__ cats are like cartoons. __p1__ that's cool ,
            whats your favorite food ? __p2__ pizza. __p1__ ni hao . as
            my father says . you must have great plans ahead ? __p2__
            yes, i plan to be a success.",
```

```
        "error": "C2",
        "error_type": "",
        "user_reaction": "",
        "comment": "",
        "turn": "",
        "phrase": "",
    },
    ...
]
```

The structure is a bit different. All cold dialogues are provided in one large json file, and the dialogues themselves maintain the structure of the original dataset. In this case, it is an dialogue from the human-bot split of the Self-Feeding Chatbot (*p2* represents the system, *p1* represents the user). There are two additional fields here: *turn* and *phrase*. If you by chance find a phrase that indicates dissatisfaction in the user's response to a system's error, please add phrase and turn to these fields.

J.3 Taxonomies

J.3.1 Error Taxonomy

This is the taxonomy for the field *error*. There are two values:

- C1 – Error
- C2 – No Error

In case of conspicuous dialogues, set *C1* if you find that *phrase* correctly emphasizes the previous system response as error-prone. In case of the cold dialogues, set *C1* if you find that the dialogue contains an error in any system response. Set *error* to C2 if you find that there is no error.

J.3.2 User Reaction Taxonomy

This is the taxonomy for the field *user_reaction*.

```
[
    {
        "id": "UR1",
        "short": "The user ignores the error and continues the conversation.",
        "description": "The user simply continues and does not draw the
            system's attention to the error.",
        "example": "-"
    },
    {
        "id": "UR2",
        "short": "The user repeats or rephrases his/her concern.",
        "description": "The user repeats or rephrases his originally concern.",
        "example": "'Can you book a restaurant for two for tonight?' vs.
            'Can you book a table for two for tonight?'"
    },
    {
        "id": "UR3",
        "short": "The user makes the system aware of the error and provides
            a correction.",
        "description": "The user makes the system aware and intervenes,
            e.g. with a correction.",
        "example": "'No, I didn't want you to book a table. I just wanted
```

```
        the address!'"
    },
    {
        "id": "UR4",
        "short": "The user makes the system aware without providing a
            correction.",
        "description": "The user makes the system aware but does not provide
            a correction.",
        "exmaple": "'No. You're wrong.'"
    },
    {
        "id": "UR5",
        "short": "The user asks for clarification.",
        "description": "The user is puzzled and asks for clarification,
            e.g. the system suddenly switches to another topic or mixed
            concepts up.",
        "example": "'What do you mean?'"
    }
]
```

J.3.3 Error Type Taxonomy

This is the taxonomy for the field *error_type*.

```
[
    {
        "id": "E1",
        "name": "Ignore Question",
        "description": "The response ignores a user’s question.",
        "example": "User: 'How is the weather in Berlin?', System:'Have you
            ever heard of Bobby Jean?'"

    },
    {
        "id": "E2",
        "name": "Ignore Request",
        "description": "The response ignores a user's request to do something.",
        "example": "User: 'Can you book me a room for two please?',
            System: 'Look at these flowers... Aren't they beautiful?'"
    },
    {
        "id": "E3",
        "name": "Ignore Expectation",
        "description": "The response does not fulfill the user's expectation.",
        "example": "User: 'Can you book me a room at The Inn for two nights
            starting tomorrow? Can you also provide me with the address?',
            System:'Absolutely!'"
    },
    {
        "id": "E4",
        "name": "Slot Error",
        "description": "The response suggests that the system did not get the
            slots right.",
```

```
    "example": "User: 'Can you book me an expensive restaurant in the west
        of town?', System:'Sure. I would recommend The Breakfast Club.
        It is a very cheap and nice location."
},
{
    "id": "E5",
    "name": "Factually Incorrect",
    "description": "The response contains information that is factually
        incorrect.",
    "example": "User: 'Who is the current president of the United States
        of America?', System:'Abraham Lincoln.'"
},
{
    "id": "E6",
    "name": "Topic Transition Error",
    "description": "The response transitions to another / a previous topic
        without reasonable explanation.",
    "example": "User: 'I think I have a cold.', System:'I really love pears.
        So juicy!'"
},
{
    "id": "E7",
    "name": "Conversationality",
    "description": "The response indicates that the system lost track, e.g.,
        it repeats previous responses (without asking for missing information)
        or contradicts itself.",
    "example": "System: 'Man is shaped by his environment.', User:'Yes. But
        does that justify this tone?', System: 'Man is shaped by his environment.'"
},
{
    "id": "E8",
    "name": "Unclear Intention",
    "description": "The response suggests that the user's intent was not
        successfully conveyed.",
    "example": "User: 'Can you book me a flight to London?', System: 'Yes. What
        do you want for dinner?'"
},
{
    "id": "E9",
    "name": "Lack of Sociality",
    "description": "The response lacks consideration of social standards,
        e.g., greetings, is toxic or disrespectful.",
    "examples": "User: 'Hi! My name is Jochen Schweitzer!', System: 'Sure. How
        can I help you?'"
},
{
    "id": "E10",
    "name": "Lack of Common Sense",
    "description": "The information in the responses opposses the opinion of
        the majority.",
    "example": "User: 'You have to work hard for success.', System: 'Hmm...
```

```
            No. What is to be, will be.'"
    }
]
```