Learning From Free-Text Human Feedback – Collect New Datasets Or Extend Existing Ones?

Anonymous ACL submission

Abstract

Continuous learning from free-text human feedback, such as error corrections, new knowledge, or alternative responses, is essential for today's chatbots and virtual assistants to stay up-to-date, engaging, and socially acceptable. However, for research on methods for learning from such data, annotated data is scarce. To address this, we examine the error and user response types in dialogues from six popular dialogue datasets of various types, including MultiWoZ, SGD, BABI, PersonaChat, Wizards-of-Wikipedia, and the human-bot split from the Self-Feeding Chatbot to assess their extendibility with the needed annotations. For this corpus study, we manually annotate a subset of each dataset with error and user response types using an improved version of the Integrated Error Taxonomy and a newly proposed user response type taxonomy. We provide the resulting dataset (EURTAD) to the community. Our findings provide new insights into dataset composition, including error types, user response types, and the relations between them[1].

1 Introduction

Chatbots and virtual assistants, such as OpenAI's ChatGPT[2] or Google's BARD[3], are increasingly important to our digitized society. One important reason for their success is that they are continuously improved using user interaction data (Shuster et al., 2022; Christiano et al., 2023; Ouyang et al., 2022; Ung et al., 2022; Xu et al., 2022). This is key to keep them up-to-date, engaging, and socially acceptable. In this regard, free-text is particularly important, since users tend to provide textual descriptions of what went wrong or what they would have expected rather than choosing from a list of predefined error types (See and Manning, 2021; Xu et al., 2022). Unfortunately, only a few publicly available datasets provide annotations for learning from such data, e.g., FITS (Xu et al., 2022) or SaFeRDialogues (Ung et al., 2022), which limits research in this direction. As this is a common issue, the use of large pretrained language models for automatic data annotation has recently come into focus (Kim et al., 2022; Zheng et al., 2022). Since many high-quality dialogue datasets are already publicly available, e.g., MultiWoZ (Zang et al., 2020), PersonaChat (Zhang et al., 2018), or Wizards-of-Wikipedia (Dinan et al., 2018), it might also be possible to use such approaches to extend the annotations of these datasets for learning from free-text human feedback. However, a prerequisite for this to work is that the datasets to be augmented already contain the information needed to generate the annotations. For learning from human free-text feedback, this means errors in system utterances that users respond to with corrections, new knowledge, or alternative answers. Based on our current knowledge, we can only say little about the extent to which available dialogue datasets contain this information. To investigate this is the purpose of this work.

We examine the types of errors and user responses in dialogues from six popular non-feedback-annotated datasets of different types, including MultiWoZ, SGD (Rastogi et al., 2020), BABI (Bordes et al., 2016), PersonaChat, Wizards-of-Wikipedia, and the human-bot split from the Self-Feeding Chatbot (Hancock et al., 2019) to assess their extendibility with annotations for learning from free-text human feedback. Since many of the dialogues may not contain any errors, we follow a two-step approach for this study: We first use Sentence-Transformer (Reimers and Gurevych, 2019) to identify potentially relevant dialogues, and then human annotators for annotation and subsequent in-depth analysis. Overall, our contribution is three-fold:

[1] Code and data are available on GitHub: http://test.test.

[2] https://chat.openai.com/

[3] https://bard.google.com

- Our corpus study provides new insights into the error and user response types (and their relations) included in the dialogues of the datasets examined.
- For human annotation, we improve the Integrated Error Taxonomy proposed by Higashinaka et al. (2021) to be broadly applicable across different dialogue types. Moreover, we propose a new taxonomy for the classification of user response types.
- We provide a dataset of 1,155 dialogues of different types, collected from the investigated datasets, but manually enriched with annotations for errors and user responses to support research on methods for learning from free-text human feedback, feedback detection, or feedback annotation in dialogue data. It is the result of our human annotation study. We refer to this dataset as EURTAD.

Our results show that the errors in system utterances and how users respond to them largely depend on the dialogue type and whether the dialogue is between humans or between a human and a chatbot. In particular, human-human dialogues do not provide enough error situations to be interesting for extending annotations for learning from free-text human feedback.

2 Related Work

2.1 Datasets Annotated With Free-Text Human Feedback

Datasets annotated with free-text human feedback are scarce, which is why most works that address this research direction collect data from scratch (mostly during in-production use). For evaluation of their approach for learning from implicit user feedback, Park et al. (2021) collected dialogues with annotations for user dissatisfaction and rephrases. Veron et al. (2021) proposed an approach for evaluation of continuous learning and collected dialogues annotated with new knowledge for this purpose. Both works address task-oriented dialogues. Unfortunately, their data was never made publicly available. For the Self-Feeding Chatbot, Hancock et al. (2019) collected and published 60,000 open-domain human-bot dialogues, partly annotated with alternative responses for unsatisfying system outputs. However, a more common alternative is the FITS dataset (Xu et al., 2022). It consists of 14,000 human-bot dialogues annotated with up to five different feedback types, including free-text human feedback. It targets open-domain and knowledge-grounded dialogues. SaFeRDialogues (Ung et al., 2022) is another feedback-annotated dataset. It provides 7,000 human-bot dialogues with annotations for offensive responses along with respectful alternatives.

As of today, many high-quality and widely used dialogue datasets are available from various types and for various use cases, e.g., MultiWOZ (Zang et al., 2020) for task-oriented dialogues or PersonaChat (Zhang et al., 2018) for persona-grounded open-domain dialogues. If it would be possible to extend them for learning from free-text human feedback, research in this direction could benefit from these advantages without the need to collect data from scratch.

2.2 Taxonomies for Error and User Response Types

Error taxonomies are usually use case or dialogue type specific. For example, the datasets discussed in Section 2.1 are all based on use case-specific taxonomies. For FITS, Xu et al. (2022) distinguish errors in search queries, results, or final responses. For SaFeRDialogues, Ung et al. (2022) distinguish between good and bad responses. For the Self-Feeding Chatbot dataset, Hancock et al. (2019) made no difference between error types.

Dybkjaer et al. (1996) proposed a dialogue type-specific error taxonomy for task-oriented dialogues that takes background knowledge into account. They distinguish four error categories, e.g., whether the user is an expert or novice. Möller et al. (2007) also addressed task-oriented dialogues but focused on practical aspects and ignored content-related errors like factually incorrect information. However, today's dialogue systems are versatile, and the difference between dialogue types rather recedes into the background, resulting in a need for generally applicable error taxonomies. In this regard, Higashinaka et al. (2021) proposed the Integrated Error Taxonomy which covers all dialogue types. It consists of 17 error types across four categories divided into two violation types (refer to Table 1). Unfortunately, it comes with some limitations (see Section 4.1.1), which we try to address by proposing an improved version (see Section 4.2.1).

Regarding user response types, See and Manning (2021) proposed a taxonomy for classifying user dissatisfaction. However, it does not clearly differentiate between errors and user response types, e.g., repetition, which is a common indicator of a bot repeating itself, is considered a type of user dissatisfaction. For this reason, and to conduct this corpus study, we propose a new taxonomy that focuses only on the different types of user responses.

3 Examined Datasets

In this corpus study, we consider six popular datasets with dialogues of various types, including task-oriented, open-domain, and knowledge-grounded dialogues. Some of them contain dialogues between humans, some of them contain dialogues between humans and chatbots. For simplicity, we use the same terminology and always refer to the partner's utterance as a system utterance.

3.1 Task-Oriented Datasets

We consider three task-oriented datasets in this work: MultiWoZ (Zang et al., 2020), SGD (Rastogi et al., 2020), and BABI (Bordes et al., 2016). MultiWoZ contains 8,438 dialogues across seven different domains. SGD provides 16,000 dialogues across 16 domains. Both datasets consist of human-human dialogues and provide extensive annotations, such as for natural language understanding or state tracking. BABI consists of 6,235 human-bot dialogues across six tasks of increasing difficulty. It is limited to a single domain, restaurant booking.

3.2 Open-Domain Datasets

For open-domain datasets, we consider PersonaChat (Zhang et al., 2018) and the human-bot split of the Self-Feeding Chatbot (Hancock et al., 2019) in our corpus study. 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 the conversation. The human-bot split of the Self-Feeding Chatbot consists of 60,000 dialogues and is partially annotated with alternative responses (we only consider the non-annotated dialogues in this work).

3.3 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.

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, respectively.

4 Methodology and Taxonomies

Our study focuses on non-feedback-annotated dialogue datasets. This means that a significant portion of the dialogues may not contain any errors. Since this would make a purely manual analysis highly inefficient and costly, we follow a two-step approach for this corpus study:

1. We use Sentence-Transformer (Reimers and Gurevych, 2019) to identify dialogues that potentially contain errors in system utterances (Section 4.1). Hereinafter, we refer to this step as *SAF* (semi-automatic filtering).

2. We manually annotate and analyse a subset of the potentially relevant dialogues with error and user response types (Section 5). For this, we use the error and user response type taxonomies presented in Section 4.2.1 and 4.2.2.

4.1 Semi-Automatic Filtering (SAF)

To identify potentially relevant dialogues, i.e., dialogues with errors in system utterances that are answered by users with, e.g., corrections or response alternatives, we filter the datasets for dialogues that contain user responses that are likely to indicate an error in the previous system utterance[4]. For this, we use Sentence-Transformer (Reimers and Gurevych, 2019) to calculate the similarity between the user responses (splitted into sentences) and *error-indicating sentences* (Section 4.1.2). An error-indicating sentence is a sentence that is known to contain an error-indicating phrase, a text fragment of arbitrary length (n-grams) that indicates user dissatisfaction[5] or an error in the previous system utterance, e.g., *that's non-sense* or *you're wrong*. We manually collect them in advance (Section 4.1.1).

[4]We use user responses to identify errors in system utterances, since preliminary studies indicated that they are easier to classify.

[5]We focus on user dissatisfaction since preliminary studies revealed that focusing on corrections, new knowledge or response alternatives is too restrictive.

4.1.1 Error-Indicating Sentences

For collecting error-indicating sentences, we first manually analyze a randomly sampled set of $1,200$ dialogues (with 200 dialogues from each of the datasets) for containing errors in system utterances, using the Integrated Error Taxonomy proposed by Higashinaka et al. (2021). The taxonomy consists of 17 error types (I1-I17) across four levels: utterance, response, context, and society. They further categorize error types into content violation, where the content of the response may cause a dialogue breakdown, and form violation, where the content is not interpretable due to massive grammatical problems. Table 1 presents a summary of the error types (see Appendix A for more details).

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 of the Integrated Error Taxonomy by Higashinaka et al. (2021). The numbers in the brackets are the corresponding identifiers.

Once we find an error, we analyze the following user response for the error-indicating phrase and add the surrounding sentence to our list of error-indicating sentences. This way, we collected a set of 67 error-indicating sentences with an average sentence length of approximately 6.52 words (see Appendix B for all collected phrases and sentences). Each sentence contains a unique error-indicating phrase with an average length of 3.52 words. Contractions (two words that have been connected, e.g., *don't* or *it's*) are considered as one word. Table 2 shows the distribution of error-indicating sentences across datasets.

	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	10

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

We find most error-indicating sentences in open-domain and knowledge-grounded datasets, especially in SFC (Hancock et al., 2019).

4.1.2 Filtering for Relevant Dialogues

For each dataset, we decompose every dialogue into turns (alternating utterances), extract the user response, and segment it into sentences. Next, we pair these sentences with each error-indicating sentence (collected in Section 4.1.1) and use a pretrained Sentence-Transformer (Reimers and Gurevych, 2019) based on MPNet (Song et al., 2020) to calculate their similarity (see Appendix C for implementation details). We consider a dialogue to be relevant if it contains at least one user response that is likely to address an error in the previous system utterance, i.e., if its similarity to at least one error-indicating sentence is $\geq 50\%$. Table 3 presents the sizes of the filtered subsets.

	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
SAF Size	4,936 (58.5%)	5,824 (36.4%)	421 (6.76%)	974 (8.9%)	15,960 (26.6%)	1,689 (7.57%)

Table 3: Size comparison between the original and the filtered subsets. The number in brackets shows the portion of the original dataset.

MWoZ (Zang et al., 2020) contains the largest portion of relevant dialogues, i.e., 58.5%. PC (Zhang et al., 2018) and WoW (Dinan et al., 2018) have the smallest portion of identified dialogues, i.e., 8.9% and 7.57%, respectively (see Appendix D for a sentence-level analysis)[6]. Overall, only 25% of the data seems relevant, i.e., contains at least one user utterance that is similar to one of the error-indicating sentences.

4.2 Taxonomies

4.2.1 Improved Integrated Error Taxonomy

During the collection of error-indicating sentences (Section 4.1.1), we found that the Integrated Error Taxonomy (see Table 1) is not optimal for identifying errors in system utterances. First of all, six of the 17 error types are never observed in the data, e.g., *uninterpretable* (I1), which describes system responses that consist of linguistically invalid text fragments. Secondly, some error types are ambiguous or similar, e.g., *ignore request* (I6) and *ignore proposal* (I7), since the system ignores the user's

[6] We also used SAF with only the error-indicating phrases instead of the complete sentences. However, we found that they are not expressive enough due to their small length (see also Section 4.1.1).

request in either case given the original definition. Based on these insights, we improve the taxonomy for the classification of errors in system utterances. Table 4 shows the result.

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.
	Attribute Error (E4)	The system utterance suggests that the system did not get the attributes/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 4: Taxonomy for the classification of errors in system utterances.

We ignore *lack of information* (I12 in Table 1), since it is rarely observed in the original paper and we never observed it in our study. For the same reason, we ignore I1-I3. However, we also found them to be rather ambiguous. For example, the *semantic error* (I3 in Table 1) is intended to be used for invalid predicate/argument combinations, such as situations where a missing letter results in a different meaning (*raining* instead of *training*). This is similar to the *lack of common sense* error type (I17 in table 1, now E10), since the model is supposed to be aware of the concept, but not in the given context. For *wrong information* (I4 in Table 1), we introduce a new error type, *factually incorrect* (E5), that extends the original definition for also taking factually incorrect knowledge into account. Furthermore, we ignore *contradiction* (I14 in Table 1), since it is covered by *lack of common sense* and *factually incorrect* (E5) errors. We merge *ignore proposal* (I7 in Table 1) and *ignore request* (I6 in Table 1) into one error type (E2 in Table 4), since both are very similar in meaning. 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. We merge *repetition* (I15 in Table 1) and *self-contradiction* (I13 in Table 1) into a new error type, *conversationality* (E7), since we observed both very rarely and only in situations that the system had lost the thread of the conversation. We also observed instances of incorrectly conveyed attributes (slots) that are not accounted for in the original taxonomy. To address this, we introduce the *attribute error* error type (E4).

4.2.2 User Response Type Taxonomy

During the collection of error-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.
- **UR3** — The user makes the system aware of the error and provides information to address what is missing or wrong in its utterance.
- **UR4** — The user makes the system aware of the error without providing additional information.
- **UR5** — The user asks for clarification.

Among these, we find that UR2, UR3, and UR5 are likely to contain free-text human feedback, such as corrections, new knowledge, or response alternatives.

5 Findings

To investigate the error types, user response types, and their relations, we manually annotate 555 dialogues identified by SAF (100 from each dataset, if available) using the taxonomies presented in Section 4.2.1 and 4.2.2. To avoid bias from SAF, we also annotate 600 randomly selected dialogues (100 from each dataset) that were not identified by SAF (similarity <50%, see also Section 4.1.2). Overall, we manually annotate 1,155 dialogues with error and user response types. For annotation, we always consider the entire dialogue (the context).

5.1 Error Type Distribution

During this analysis, we identified 188 errors across all dialogues. Table 5 shows the distribution.

	Task-Oriented			Open-Domain		Know.-Grounded
Dataset	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
SAF Dialogues	8/100	3/100	2/95	6/71	92/100	19/89
Random Dialogues	2/100	0/100	5/100	2/100	46/100	3/100

Table 5: The number of errors identified in each dataset.

As expected, the SAF dialogues contain a larger number of errors (130 overall) compared to the random dialogues (58 overall), especially for open-domain and knowledge-grounded dialogues, such as SFC (Hancock et al., 2019) and WoW (Dinan et al., 2018) (+46 in case of SFC and +16 in case of WoW).

	Task-Oriented			Open-Domain		Know.-Grounded
Dataset	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Ignore Question (E1)	1 (10.0%)	-	1 (14.3%)	1 (12.5%)	67 (48.5%)	-
Topic Trans. Error (E6)	-	-	-	1 (12.5%)	62 (44.9%)	4 (18.1%)
Factually Incorrect (E5)	-	2 (66.6%)	-	1 (12.5%)	3 (2.1%)	13 (59.1%)
Ignore Expect. (E3)	2 (20.0%)	1 (33.3%)	1 (14.3%)	-	2 (1.4%)	1 (4.5%)
Ignore Request (E2)	3 (30.0%)	-	1 (14.3%)	-	-	-
Lack of Sociality (E9)	-	-	-	2 (25.0%)	3 (2.1%)	-

Table 6: The most common error types and their frequencies found in both the SAF and random dialogues. The number in brackets shows the ratio to all errors found for the respective dataset.

Table 6 shows the most frequent error types and their frequency for both SAF and random dialogues, which already accounts for 172 of all identified errors (see Appendix I.1 for an aggregated distribution of all errors and user responses). In the case of open-domain dialogues, the most frequent error types are *ignore question* (E1) and *topic transition error* (E6). This is particularly the case in the SFC dataset (Hancock et al., 2019), where we find the system utterances to be often out of context. In the case of task-oriented dialogues, *ignore request* (E2) and *ignore expectation* (E3) are the most common error types. We observe these errors when requests are only partially processed, e.g., when the user requests to book a hotel room and a train, but the system only books the hotel room. Moreover, we find that there is only little variety in language in task-oriented dialogs, regardless of the number of tasks reflected in the dataset (see Appendix E for examples). In the case of WoW (Dinan et al., 2018), the knowledge-grounded dataset, the *factually incorrect* (E5) error is the most commonly observed error type.

5.2 User Response Type Distribution

As described in Section 4.2.2, UR2 (repeat or rephrase concern), UR3 (providing a correction), and UR5 (asking for clarification) are likely to contain free-text human feedback. Table 7 shows the distribution of user response types to errors in system utterances.

In the case of SAF dialogues, we find that

	Task-Oriented						Open-Domain				Know.-Grounded	
Dataset	MWoZ (HH)		SGD (HH)		BABI (HB)		PC (HH)		SFC (HB)		WoW (HH)	
	S	R	S	R	S	R	S	R	S	R	S	R
Errors	8	2	3	0	2	5	6	2	92	46	19	3
UR1	1	**2**	2	0	1	**3**	0	**1**	4	**35**	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	**1**	9	0
UR4	1	0	0	0	0	0	2	**1**	34	**2**	0	**1**
UR5	2	0	0	0	0	0	4	0	51	**8**	10	**1**

Table 7: User response types found in the SAF (*S*) and the random (*R*) dialogues. Those from the random dialogues that are irrelvant (see Section 4.2.2) are highlighted in bold red. The relevant ones are highlighted in bold green.

UR3 and UR5 are more often observed in open-domain and knowledge-grounded dialogues, such as SFC (Hancock et al., 2019) or WoW (Dinan et al., 2018). UR2 is only rarely observed, and only in task-oriented dialogues. However, UR1 (the user ignores the error) is also frequently observed, especially in SFC. For randomly selected dialogues, this is the most common user response type (it occurs 42 times).

5.3 Relation Between Error and User Response Type

To get a better understanding of the underlying error situations, we also investigate the relations between the most common error types (the 172 errors presented in Table 6) and user response types (see Table 7) in both the SAF and random dialogues. Figure 1 illustrates the results.

We find that UR1, UR4, and UR5 are the most frequently observed user response types, particularly when the system ignores a user's question (E1) or unexpectedly changes the topic (E6), which is mostly observed in open-domain datasets (see Table 6). Along 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 sometimes occurs as a response type to E2 (*ignore request*) and E3 (*ignore expectation*), which are mostly found in task-oriented dialogues.

5.4 Discussion

The goal of this work is to get an understanding of the error and user response types in the dataset examined to investigate whether they are extendible with annotations for learning from free-text human feedback. We find that this depends on the dia-

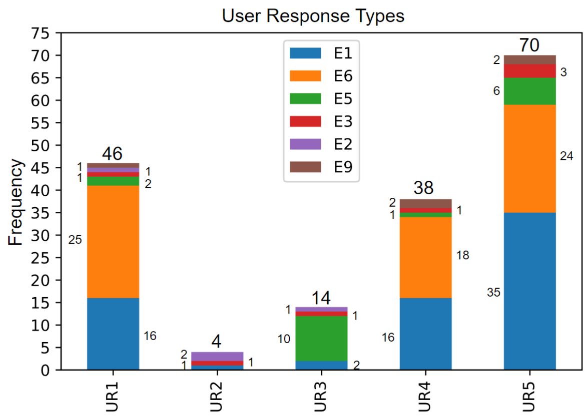

Figure 1: Illustration of the relations between frequent error (E-values, Section 4.2.1) and user response types (UR-types, Section 4.2.2) in both the SAF and random dialogues. The numbers above the bars are the total number of errors for each UR-type. The numbers to the left and right of each bar indicate the portion of the respective error type (see color coding).

logue type and whether it is between humans or between a human and a chatbot. In general, we find that open-domain and knowledge-grounded dialogues contain a larger number of errors and user responses that are likely to contain free-text human feedback, making them more suitable for this purpose (Section 5.1). This especially applies to human-bot dialogues, where we often find that humans react harshly and accusingly to errors in system utterances, resulting in more direct feedback. For task-oriented dialogues, we find that errors are few. However, this might also be due to the fact that these are predominantly human-human dialogues. We find that humans rather suggest disagreements in a very polite way instead of accusing the partner of a mistake (see Appendix F for examples). Accordingly, there is only little free-text human feedback available that could be used for learning (Section 5.2 and 5.3). Therefore, it might be hard and ineffective to extend these datasets with annotations for learning from such data.

6 Quality Assessment

6.1 Impact of Semi-Automatic Filtering

In this section, we discuss the impact of SAF (Section 4.1) on our findings. For this corpus study, we manually annotated 1,155 dialogues with error and user response types, 555 dialogues that were identified by SAF as potentially relevant, and 600 randomly selected dialogues that were missed by SAF (see also Section 5). As Table 5 shows, a total of 188 dialogues contains errors. 130 of these are SAF dialogues and 58 are randomly selected dialogues that were missed by SAF. Considering this at the level of user response types (Table 7), 46 of these 58 errors were ignored by users or did not provide any additional information (UR1 or UR4, the ones marked in bold red in the table), meaning that they are irrelevant because they do not contain free-text human feedback. For the remaining 12 missed errors (UR3 or UR5, the ones marked in bold green), we find that they are not reflected in the set of 67 error-indicating sentences used for SAF (Section 4.1.1). Although this limits the effectiveness of SAF, we find that SAF itself has no negative impact on the results of our corpus study, but rather improved annotation efficiency. An approximated recall of 0.72 supports this assumption. The recall is approximated with respect to the ratio between the size of the filtered and the original subsets (see Table 3)[7]. Moreover, we only considered the 12 missed relevant errors for this (when considering all missed errors, the recall is 0.35). We provide a more detailed analysis in Appendix D and G.

6.2 Inter-Annotator Agreement

To assess the quality of our annotations, we asked nine experts with NLP background and sound English skills to annotate smaller subsets of the SAF and randomly selected dialogues (300 from each, 50 from each of the datasets examined, 600 overall) for error and user response types (see Appendix H for more background on participating annotators and Appendix K for the annotation guidelines). Each of these subsets was assigned to two of these experts, and each dialogue was annotated three times in total (including our initial annotation). For calculating the inter-annotator agreement (IAA), we use Krippendorff's Alpha (Krippendorff, 2004)[8]. For comparison, we mapped all annotations to the Integrated Error Taxonomy (Higashinaka et al., 2021)[9]. Table 8 shows the results summarized by human-human and human-bot dia-

[7] We only considered 139 randomly sampled dialogues (25%) of the 555 annotated SAF dialogues and all 600 missed dialogues for this. To increase the significance, we repeated this process a thousand times and averaged the recall. On average, the 139 randomly sampled SAF dialogues consisted of 31 errors.

[8] We use the Python library *annotation_analysis* for this: https://github.com/ai-nikolai/annotation_analysis, last accessed on 15/01/23.

[9] For merged error types (Section 4.2.1), we asked annotators for a second assessment using the original error types.

logues (see Appendix I for a detailed analysis).

Dataset		Ours HH	Ours HB	Integrated HH	Integrated HB
Error Type	SAF	0.16	0.91	0.02	0.89
	Random	0.17	0.40	0.16	0.39
User Res. Type	SAF	0.06	0.48	-	-
	Random	0.01	0.40	-	-

Table 8: Inter-annotator agreement using the Integrated Error Taxonomy (Higashinaka et al., 2021) (*Integrated*) and our improved version (*Ours*).

In the case of human-human dialogues, the overall agreement is rather low. This is also reflected in the user response type agreement, as this depends on the error type annotation. However, this is different for human-bot dialogues. We attribute this to the different characteristics of human-human and human-bot dialogues, i.e., humans rather suggest disagreement in human-human dialogues and tend to provide direct feedback in human-bot dialogues (Section 5.4).

Overall, using our improved error taxonomy improves IAA over the original Integrated Error Taxonomy (Higashinaka et al., 2021) in all cases. This is most obvious in the case of the human-human SAF dialogues, where it improves IAA by 0.14 points. A detailed analysis revealed that this is mainly due to (1) the condensed number of abstract error types, e.g., we merged ambiguous error types such as *ignore proposal* and *ignore request*, and (2) the newly added error types, such as *factually incorrect*, which were not covered in the original taxonomy (see Section 4.2.1 for our taxonomy modifications and Appendix I for a detailed analysis, including edge cases).

7 EURTAD

To support research into methods for learning from free-text human feedback, feedback detection, or feedback annotation, we provide the 1,155 manually annotated dialogues from our corpus study as EURTAD (*Error and User Response Type Annotated Dataset*) to the community. In comparison to existing feedback-annotated datasets, such as FITS (Xu et al., 2022), which is rather focused on functional errors, or SaferDialogues (Ung et al., 2022), which only focuses on safety errors, it provides annotations for a broad range of error types. Moreover, it provides annotations for user response types. Especially because only 16% of EURTAD contain annotations for errors (with a bias towards the SFC (Hancock et al., 2019) dataset), we consider the annotations for user response types as the more interesting part of the dataset. From our point of view, distinguishing between user response types could be an interesting alternative to binary signals, such as user satisfaction (Hancock et al., 2019) or thumbs-down (Shuster et al., 2022), as an indicator of an error in a system utterance. Future research may pick up this question. To support this, EURTAD provides enough annotations of the relevant user response types to be used in few-shot scenarios. In order to be easily accessible, we use a unified json format (inspired by FITS) that also provides the original annotations (see Appendix J for the json structure).

8 Conclusion

Continuous learning from free-text human feedback is essential for today's chatbots and virtual assistants to stay up-to-date, engaging, and socially acceptable. Unfortunately, appropriately annotated datasets are scarce, which limits research on methods for using such data. In this work, we have examined the dialogues of six popular datasets from various types, including MultiWoZ, SGD, BABI, PersonaChat, Wizards-of-Wikipedia, and the human-bot split from the Self-Feeding Chatbot for errors in system utterances and how users respond to them, to assess their extendibility with annotations for learning from free-text human feedback. We found that this largely depends on the dialogue type and whether it's between humans or between a human and a chatbot. Human-human dialogues rarely provide free-text human feedback (especially in the case of task-oriented dialogues). Therefore, it might be ineffective to make these datasets available for learning from such data. This is different in open-domain and knowledge-grounded dialogues. For this reason, it might be possible to extend these datasets with the needed annotations to support research into methods for learning from free-text human feedback, e.g., using pretrained language models for data augmentation, instead of collecting new datasets from scratch. For our corpus study, we improved the Integrated Error Taxonomy and proposed a new taxonomy for classifying user response types. We provide the resulting set of manually annotated dialogues (EURTAD) to the community.

9 Limitations

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. It might be possible that our results would have been clearer when we would have considered more dialogues for the collection of error-indicating sentences. However, our analysis shows that errors found in the randomly selected dialogues are mostly ignored by the user, i.e., the user does not provide free-text human feedback that could be used for learning. Thus, as far as we are concerned, this does not limit the meaningfulness of our results.

Regarding dataset selection, our corpus study (and its results) have only limited expressiveness for knowledge-grounded dialogue datasets, since 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, e.g., FITS (Xu et al., 2022), and we considered a representative number of datasets for other dialogue types for which there is a lack of publicly available feedback-annotated datasets, such as task-oriented dialogues.

The taxonomies used in this work are also subject to limitations. In the case of the improved error type taxonomy, our results show that it improves agreement across different dialogue types. However, its abstract error types might limit application for specific use cases, e.g., for a more fine-grained consideration of different types of social errors. The same applies to the user response type taxonomy. Its focus on abstract types might limit application if it is necessary to distinguish between different kinds of free-text human feedback.

References

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

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

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.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization.

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

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*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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.

Abigail See and Christopher Manning. 2021. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.

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.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models.

A The Integrated Error Taxonomy – Details

In this section, we describe the Integrated Error Taxonomy as proposed by Higashinaka et al. (2021). In principle, they differentiate 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., inconsistent or redundant utterances, or other things that might cause a dialogue breakdown. Content violation is hereinafter abbreviated as *CV* (form violation as *FV*). Furthermore, they generally refer to *utterances*, while we refer to *system utterance*, as this is evident from their examples and, from our perspective, 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, 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.

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 Implementation of Semi-Automatic Filtering

To implement SAF (see Section 4.1) we use PyTorch (Paszke et al., 2019), the Transformers library (Wolf et al., 2020), and the pretrained *all-mpnet-base-v2* Sentence-Transformer. The model is available here: `https://huggingface.co/sentence-transformers/all-mpnet-base-v2` (last accessed 11/02/2023). It is based on MPNet (Song et al., 2020) and 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.

Our compute infrastructure consists of one Tesla V100-SXM3 GPU (with 32 GB memory) and it takes an average of 76 mins to run SAF on one dataset.

D Error Distribution – Sentence-Level Analysis

As described in Section 4.1.2, we use sentence-level for semi-automatic filtering (SAF). Table 9 shows the impact on SAF 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 SAF.

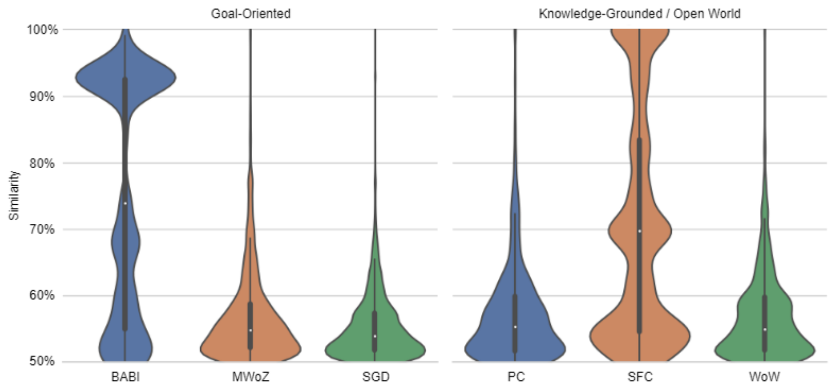

Figure 2: Illustration of the distribution of dialogues with potential real-world error scenarios across datasets.

	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 SAF (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 datasets which might be the reason for the clusters in the lower ranges.

E 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 show 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 do not 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).

F Human-Human Dialogues – Examples

We find that it is hard to identify errors in human-human dialogues, as the tone is in principle mostly polite, even if one of the partners 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, one partner claims that photography is science. However, it's commonly considered visual art. Instead of directly correcting him, the other person 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, one partner makes a joke that the other does not understand. The response is quite polite. He informs the partner that his statement does not make much sense to him.

G Effectiveness Of Semi-Automatic Filtering – A Detailed Analysis

The semi-automatic-filtered subsets (SAF) considered for manual annotation consist of 20 dialogues from each of the datasets examined and 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), SAF consists only of 555 dialogues (instead of 600 like the randomly selected dialogues). Table 10 shows the results of our error type analysis with respect to the similarity ranges identified by SAF (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 dialogues randomly sampled from the respective similarity range, and *Error* (E) represents the number of dialogues identified in our manual analysis to contain an error in a system utterance.

		Task-Oriented						Open-Domain				Know.-Grounded	
Dataset		MWoZ (HH)		SGD (HH)		BABI (HB)		PC (HH)		SFC (HB)		WoW (HH)	
Overall / Error		O	E	O	E	O	E	O	E	O	E	O	E
SAF Dialogues	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 Dialogues		100	2	100	0	100	5	100	2	100	46	100	3

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

Overall, only 58 dialogues of the randomly selected ones (9.6%) contain errors. In the case of

SAF, we observe 130 of such cases. Therefore, SAF shows to facilitate the process of identifying free-text human feedback. 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, $50\% - 60\%$ (see also Figure 2).

H Human Annotators

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

I Inter-Annotator Agreement – Detailed Analysis

This section gives more insights on the inter-annotator agreement. Table 11 shows the inter-annotator agreement for each dataset using our improved error type taxonomy.

Dataset		Task-Oriented MWoZ (HH)	Task-Oriented SGD (HH)	Task-Oriented BABI (HB)	Open-Domain PC (HH)	Open-Domain SFC (HB)	Know.-Grounded WoW (HH)
Error Type	SAF	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	SAF	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 11: Inter-annotator agreement for each dataset.

In the case of human-human dialogues, the overall agreement is rather low (except for PersonaChat (Zhang et al., 2018)). We find that errors are hard to identify in these dialogues, as humans rather suggest disagreements instead of accusing the partner of a mistake. This is also reflected in the user response type agreement since it depends on the error type annotation. However, PersonaChat seems to be different (according to Table 6). We attribute this to the dialogue type, which is open-domain, where we find that humans react harshly and accusing to errors in system utterances, resulting in more direct feedback that is easier to identify.

Table 12 shows the inter-annotator agreement for each dataset using the original error type taxonomy as proposed by Higashinaka et al. (2021). Using this taxonomy deteriorates the inter-annotator agreement. This is most obvious in case of MWoZ (Zang et al., 2020) and PC (Zhang et al., 2018), which are both human-human datasets. A detailed analysis revealed that this is mostly due to over-specialized error types which were merged in our improved taxonomy, such as *ignore expectation* and *ignore request*, I9 and I6 in the original taxonomy (Table 1). Another reason are the newly added error types, such as *factually incorrect*, E5 (Table 4), which were not covered in the original taxonomy, but occur in the dialogues. Overall, the results show that our revised taxonomy improves the general applicability and acceptance of the original error type taxonomy proposed by Higashinaka et al. (2021).

Dataset		Task-Oriented MWoZ (HH)	Task-Oriented SGD (HH)	Task-Oriented BABI (HB)	Open-Domain PC (HH)	Open-Domain SFC (HB)	Know.-Grounded WoW (HH)
Error Type	SAF	-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 12: Inter-annotator-agreement for the Higashinaka et al. (2021) taxonomy.

I.1 Edge Cases

To get a better understanding of the disagreement between annotators, Table 13 shows the aggregated error type distribution (error type annotation from both the SAF and random subsets).

	Task-Oriented									Open-Domain						Knowledge-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	2	2	-	-	1	1	2	1	1	2	5	67	64	66	-	1	3
Ignore Request (E2)	3	3	3	-	-	-	1	-	-	-	-	-	1	9	6	-	-	-
Ignore Expect. (E3)	2	3	3	1	-	1	1	1	1	-	-	-	2	1	-	2	-	-
Attribute Error (E4)	3	1	-	-	-	-	4	-	3	-	2	3	-	3	-	1	5	1
Factually Incorrect (E5)	-	2	-	2	-	-	-	4	-	1	2	-	3	1	-	13	1	1
Topic Trans. Error (E6)	-	-	1	-	-	-	-	-	-	2	2	10	62	58	58	4	-	1
Convers. (E7)	1	-	2	-	-	1	-	-	-	1	1	1	-	-	2	1	-	3
Unclear Intention (E8)	-	12	-	-	-	-	-	-	-	-	-	1	-	2	2	-	13	-
Lack of Sociality (E9)	-	-	-	-	-	-	-	-	-	2	1	4	2	2	1	-	-	-
Lack of Com. Sense (E10)	-	-	1	-	-	-	-	-	-	1	2	2	-	-	-	1	-	1

Table 13: Error types in both the SAF and randomly selected dialogues.

Overall, the distribution is very broadly spread. However, in most cases, it seems like at least two annotators agree. There are only a few outliers where there is a large deviation, i.e., unclear intention (E8) in case of MWoZ (Zang et al., 2020) and

WoW (Dinan et al., 2018), topic transition error (E6) in case of PC (Zhang et al., 2018), factually incorrect (E5) and attribute error (E4) in case of WoW. For example, attribute error is defined as an error type that rather addresses task-oriented dialogues, but annotator two found it five times in the WoW dataset. During our analysis, we found that factually incorrect would have described these cases more accurately. In the case of unclear intention in WoW and MWoZ, we found that annotator two marked some cases as errors that are actually not necessarily errors. The same applies to the factually incorrect errors in BABI (which consists of task-oriented dialogues). In the case of PC, we found that topic transition error is in most cases the most obvious error type, and in our opinion, annotator three was right in most of the cases.

In summary, we find that deviations are primarily the result of (1) how the annotators interpret the descriptions of the error types (based on their experience), and (2) biases in the data. The former could probably be addressed by more examples in the annotation guidelines (see Section K). The latter is a bit more difficult. In these cases, a multi-step annotation process could be useful, where annotators mark errors they are not sure about to be discussed before they are finally annotated.

	Task-Oriented									**Open-Domain**						**Knowledge-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	3	8	5	2	-	3	4	2	3	1	2	23	39	40	36	1	6	2
UR2	2	9	3	1	-	-	1	3	1	-	-	1	-	1	-	-	-	-
UR3	2	3	3	-	-	-	2	1	-	-	-	-	4	3	-	9	7	5
UR4	1	3	1	-	-	-	-	1	1	3	5	2	36	37	35	1	-	-
UR5	2	-	-	-	-	-	-	-	-	4	5	-	59	59	64	11	7	3

Table 14: User response types in both the SAF and randomly selected dialogues.

Table 14 shows the aggregated distribution of user response types. What stands out here are the differences in the situations where users ignore the error situations (UR1) in the case of PC (annotator three) and WoW (annotator two). However, we found no pattern in the underlying error situations. We suspect that this is due to the nature of human-human dialogues (PC and WoW consist of human-human dialogues). As we have pointed out before, subjects behave more cautiously and politely in human-human dialogues, while in contrast they clearly point out errors when communicating with chatbots (see Appendix F for examples).

J EURTAD – Dialogue Structure

To support research into methods for learning from free-text human feedback, feedback detection, or feedback annotation, we publish the EURTAD dataset. It consists of 1,155 dialogues from various domains. In order to maintain reusability, we provide the dialogues in a unified json format that extends the original annotations with error and user response type annotations. The following listing shows the dialogue structure:

```
{
   "unique_id": "unique id in the
       context of EURTAD, e.g., PMUL0121.
       json_multiwoz_train",
   "id": "dataset-specific id, e.g.,
       PMUL0121.json",
   "turns": [
      {
         "utterance": {
            "text": "text of the
                utterance",
            "original_annotations": [
               ...
            ],
            "error_type": "the error
                type, e.g., E5",
            "error_type_comment": "the
                annotators comment for
                why this is an error"
         },
         "response": {
            "text": "text of the
                response",
            "original_annotations": [
               ...
            ],
            "response_type": "The user
                response type, e.g., UR5
                "
         }
      },
      ...
   ]
}
```

K Annotation Guidelines

K.1 Annotation Task

Learning from textual user feedback, i.e., free-form text feedback that expresses user 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 two different taxonomies, an error type taxonomy, and a user response 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.

K.2 Dialogue Format

To avoid bias, we provide the dialogues in a unified format without the original task-specific annotations. Here's an example of 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 a unique id, an id, and its turns. 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.

K.3 Taxonomies

K.3.1 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": "Attribute Error",
      "description": "The response
           suggests that the
          system did not get the
          attributes/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 opposes the
                opinion of the majority.
                ",
        "example": "User: 'You have to
            work hard for success.',
            System: 'Hmm... No. What is to
             be, will be.'"
        }
    ]
```


K.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
                of the error and
                provides information to
                address what is missing
                or wrong in its
                utterance. ",
            "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
                without providing
                additional information",
            "example": "'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?'"
        }
    ]
```