Sources of Noise in Dialogue and How to Deal with Them

Derek Chen
Columbia University, NY
dc3761@columbia.edu

Zhou Yu
Columbia University, NY
zy2461@columbia.edu

Abstract

Training dialogue systems often entails dealing with noisy training examples and unexpected user inputs. Despite their prevalence, there currently lacks an accurate survey of dialogue noise, nor is there a clear sense of the impact of each noise type on task performance. This paper addresses this gap by first constructing a taxonomy of noise encountered by dialogue systems. In addition, we run a series of experiments to show how different models behave when subjected to varying levels of noise and types of noise. Our results reveal that models are quite robust to label errors commonly tackled by existing denoising algorithms, but that performance suffers from dialogue-specific noise. Driven by these observations, we design a data cleaning algorithm specialized for conversational settings and apply it as a proof-of-concept for targeted dialogue denoising.

Figure 1: An example of label errors within MultiWoz 2.0 which contains partially filled and missing labels. We categorize this as two types of instance-level noise.

1 Introduction

Quality labeled data is a necessity for properly training deep neural networks. More data often leads to better performance, and dialogue tasks are no exception (Qian and Yu, 2019). However, in the quest for more data, practitioners increasingly rely on crowdsourcing or forms of weak supervision to meet scaling requirements. Even when acting in good faith, crowdworkers are not trained experts which understandably leads to mistakes. This ultimately results in noisy *training inputs* for our conversational agents. Moreover, when dialogue systems are deployed into the real world, they must also deal with noisy *user inputs*. For example, a user might make an ambiguous request or mention an unknown entity. All these sources of noise eventually take their toll on model performance.

Before building noise-robust dialogue systems or denoising dialogue datasets, it would be helpful to know what types of noise exist in the first place. Then our efforts can be spent more wisely tackling the sources of noise that actually make a difference. Prior works have looked into counteracting noisy user interactions (Peng et al., 2021; Liu et al., 2021), but did not study the impact of noisy training data. Moreover, they lack analysis on how noise influences performance across different model types or conversational styles. Other works claim that dialogue agents can be easily biased by offensive language found in noisy training data (Ung et al., 2022; Dinan et al., 2020). Given such a danger, we wonder "How much toxic data actually exists in annotated dialogue data?"

To investigate these concerns, we survey a wide range of popular dialogue datasets and outline the different types of naturally occurring noise. Building on this exercise, we also study the patterns of annotation errors to determine the prevalence of each noise type and identify the most likely causes of noise. Next, we run transformer models through the gamut to find out how well they handle the different types of noise documented in the previous step. In total, we test 3 model types on 7 categories

Dataset	Abbr.	Num. Dialogs	Collection Methodology	Open Domain	Goal Oriented	Synchronous Chat	KB/ Document
Action-Based Conversations Dataset	ABCD	10,042	Expert Live Chat		X	X	X
DailyDialog	DD	13,118	Post-conv Annotation	X			
Empathetic Dialogues	ED	24,850	Live Chat	X		X	
Google Simulated Conversations	GSIM	3,008	Machine to Machine		X		
Key-Value Retrieval for In-Car	KVRET	3,031	Wizard of Oz		X		X
Machine Interaction Dialog Act Schema	MIDAS	468	Live Chat	X		X	
MultiWoz 2.3	MWOZ	10,419	Wizard of Oz		X		
Schema Guided Dialogue	SGD	42,706	Post-conv Annotation		X		
TicketTalk (TaskMaster 3)	TT	23,789	Dialogue Self-Play		X		
Wizard of Wikipedia	WOW	22,311	Wizard of Oz	X		X	X

Table 1: Breakdown of ten dialogue datasets used in constructing the noise taxonomy. The datasets were chosen to span a wide variety of annotation schemes, task specifications and conversation lengths. KB/Document refers to a dataset containing an external knowledge base or document to ground the conversation. (See Appendix A)

of noise across 10 diverse datasets spanning 5 dialogue tasks. We discover that most models are quite robust to the label errors commonly targeted by denoising algorithms (Natarajan et al., 2013; Reed et al., 2015), but perform poorly when subjected to dialogue-specific noise. Finally, to verify we have indeed identified meaningful noise types, we apply our findings to denoise a dataset containing real dialogue noise. As a result, we are able to raise joint goal accuracy on MultiWOZ 2.0 by 42.5% in relative improvement.

In total, our contributions are as follows: (a) Construct a realistic taxonomy of dialogue noise to guide future data collection efforts. (b) Measure the impact of noise on multiple tasks and neural models to aid the development of denoising algorithms. (c) Establish a strong baseline for dealing with noise by resolving dialogue specific concerns, and verify its effectiveness in practice.

2 Dialogue Datasets

A data-driven taxonomy of dialogue noise was designed by manually reviewing thousands of conversations across ten diverse datasets and their accompanying annotations. The datasets were chosen from non-overlapping domains to exhaustively represent all commonly considered dialogue tasks. At a high level, they are divided into six task-oriented dialogue datasets and four open domain chit-chat datasets. The task-oriented datasets are comprised of MultiWoz 2.0 (MWOZ) (Budzianowski et al., 2018), TicketTalk (TT) (Byrne et al., 2019), Schema Guided Dialogue (SGD) (Rastogi et al., 2020), Action Based Conversations Dataset (ABCD) (Chen et al., 2021), Google Simulated Conversations (GSIM) (Shah et al., 2018), and Key-Value Retrieval for In-car Assistant (KVRET) (Eric et al., 2017). The open domain datasets include DailyDialog (DD) (Li et al., 2017), Wizard of Wikipedia (WOW) (Dinan et al., 2019b), Empathetic Dialogues (ED) (Rashkin et al., 2019), and Machine Interaction Dialog Act Schema (MIDAS) (Yu and Yu, 2021). The datasets also span a variety of data collection methodologies, such as M2M or Wizard-of-Oz, which has a close connection to the types of noise produced. We also consider whether the interlocutors engage in real-time vs. non-synchronous chat. Details of each dataset can be found in Table 1 and Appendix A.

The taxonomy creation process starts by uniformly sampling 1% of conversations from each corpus, rounding up as needed to include at least 100 dialogues per dataset. Five expert annotators then conducted two rounds of review per conversation to tally noise counts, with a third round to break ties if needed. The group also cross-referenced each other to merge duplicate categories and resolve disagreements. Notably, the final taxonomy purposely excludes sources of noise that occur less than 0.1% of the time. This active curation supports future denoising research by focusing attention on the most prominent sources of noise.

3 Sources of Noise

Through careful review of the data, we discover that dialogue systems encounter issues either from noisy training inputs during model development or from noisy user inputs during model inference.

3.1 Training Noise

Noisy training data impacts model learning, before any user interaction with the system. The sources of noise are derived from labeling errors, ontology inconsistencies or undesirable discourse attributes.

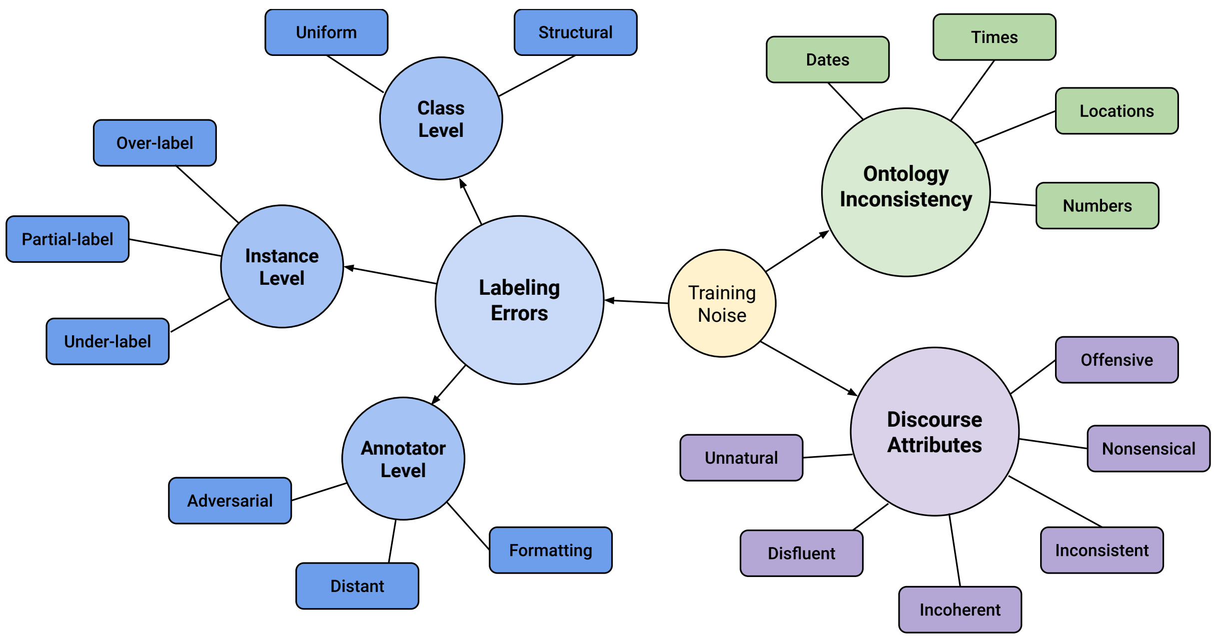

Figure 2: Diagram of the main sources of noise that affect training, based on review of the data. Our taxonomy also includes inference noise which occurs when users interact with the dialogue agent (See Fig 3).

3.1.1 Labeling Errors

For a given dataset of (X, Y) pairs, any occasion when the target label y is labeled incorrectly can be considered a labeling error.

Class Level When noise occurs due to confusion between two classes, this is considered a class-level labeling error. This can be further sub-divided into *Uniform Label Swapping* or *Structured Label Swapping*. In the former, symmetric noise implies all classes have equal likelihood to be confused with any other class, whereas in the latter certain classes are more likely to confused with other related classes. For example, "anger" as a label is more likely to be confused with "frustration" than "joy" when performing emotion detection.

Instance Level Noise comes from the example itself due to the complexity of interpreting natural language, which is especially common within dialogues (Zhang et al., 2021). For example, annotators may carry over the dialogue act from the previous turn, even though it is no longer relevant, resulting in *Over Labeling*. Conversely, *Under Labeling* is when a label is missed. *Partial Labels* occur when some labels are correct, while others are not. This is common in dialogue due to the prevalence of multi-label examples, such as an utterance with two slot-values to fill. (See Figure 1)

Annotation Level Noise arises due to the labeler or data collection process. (Snow et al., 2004). Applying heuristics on a gazetteer to label named entities in NER produces *Distant Supervision* noise. Human annotators are also a source of noise either purposely from *Adversarial Actors* or inadvertently from annotators acting in good faith still leading to *Formatting Mistakes*. (See Table 2)

3.1.2 Ontology Inconsistency

Another source of noise comes from inconsistent formatting when constructing the ontology. The only entities which actually contained issues are (a) **Dates**: tomorrow, Jan 3rd, 1/3/2022, January 3 (b) **Times**: 14:15, 2:15 PM, quarter past 2, 215pm (c) **Locations**: NYC, New York, ny, the big apple (d) **Numbers**: three, 'wife daughter & I', 3, 'Me and my two buddies'. In contrast, inconsistent names (ie. Fred Miyato, Mr. Miyato, fred miyato, my father) only occurred occasionally. Lack of standardization in the ontology was so pronounced in certain datasets that classifying labels becomes untenable, leaving generation or copying as the only viable method of predicting slot-values.

3.1.3 Discourse Attributes

Dialogue agents developed for response generation often mimic the behavior found in the training examples, so one hopes they contain positive discourse attributes while avoiding negative ones. We identify six such attributes by following qualitative metrics commonly used for dialogue evaluation and through our own review of the conversations.

Dialogue	Labels

SGD – [Ontology Inconsistency > Date, Time]
User: I need a rental car in Chicago on the 3rd of this month.
System: When and for how long will you need the car?
User: I'd like it from 12:30 in the afternoon till next Wednesday.
. . .
System: So you'd like to reserve a standard car from March 3rd at 12:30 pm until March 6th from the O'Hare International Airport location?
User: Yes that'll work
	`GetCarsAvailable(pickup_city=Chicago, pickup_date=3rd of this month)`
	
	`GetCarsAvailable(pickup_time=12:30 in the afternoon, dropoff_date=Wednesday)`
	
	`ReserveCar(dropoff_date=March 6th, pickup_time=12:30 pm)`

MIDAS – [Discourse Attribute > Incoherent] User: one guy Agent: what do you think about christopher nolan's acting User: you can't get a boy	Revised dialog act: `statement` → `nonsense`

TT – [Labeling Error > Annotator Level > Formatting]
User: We would like to see the Rhythm Section. That sounds good.
Assistant: How many tickets will you need today?
User: We will need 4 tickets.
Assistant: Where would you like to see the movie?
User: We would like to see it in San Antonio at Cinemark McCreless Market.
	`(name.movie='the Rhythm Section')`
	
	`(num.tickets=4)`
	`(location='San Antonio', name.theater ='inemark McCreless Market.')`

Table 2: Selected qualitative examples of dialogue noise. Best viewed in color. Many more examples in Appendix I.

(1) **Fluent** utterances flow well, obey proper grammar, and are syntactically valid. (2) **Coherent** dialogues are semantically valid, and make sense such that they are interpretable and understandable by a general audience. (3) **Consistent** models do not contradict what was stated earlier in the conversation, or haphazardly change their stance on a subject. (4) **Sensible** models follow common sense principles and understand basic natural laws (ie. gravity). (5) **Polite** dialogue models avoid toxic language or offensive speech. They should not exhibit overt bias towards certain groups or minorities. (6) **Natural** dialogues reflect how people generally talk in real life. In addition, the speakers should not break the fourth wall by directly or indirectly referring to the data collection process.

3.2 Inference Noise

Inference noise refers to issues that occur in test time, during user interaction with the system after deployment to production. This aligns nicely with the concept of out-of-scope errors (Chen and Yu, 2021), which are made up of two categories: out-of-distribution cases and dialogue breakdowns.

3.2.1 Out-of-Distribution (OOD)

Causes of OOD (Peng et al., 2021) include:

Novel queries The user asks the model to do something it was not trained to do. Example: the customer asks about frequent flyer miles, but the agent is only capable of making or modifying flight reservations. The model fails for these requests since it was never taught to handle such queries.

Unseen entities Facing new entities or values not seen during training. Although difficult, we could still expect a model to understand a portion of such queries by generalizing from the context. For example, "I would like a flight from Miami to Puffville". Even though the model has never heard of 'Puffville', it can infer from context that this is the desired value for the destination slot.

Domain shift The dialogue system must make predictions in a new domain (taxi vs. flight). Commonly tackled in zero-shot settings, we can expect models to occasionally generalize because there may be shared slots across domains (ie. departure time is shared by both taxi and flights queries).

3.2.2 Dialogue Breakdown

In contrast to OOD issues, dialogue breakdowns are situations a model should be able to handle since the scenario is within the bounds (i.e. in-domain) of what the model was trained to understand (Higashinaka et al., 2016). However, due to noise from ambiguous or unclear user input, communication breaks down and the conversation is unable to continue. (Higashinaka et al., 2015).

Ambiguous Meaning Query or statement that the model should be able to handle, but caused confusion, possibly because the model failed to take the dialogue context into consideration. For example, a co-reference issue may cause difficulty in interpreting the user intent. "Yea, let's go with that one" is unclear when viewed in isolation. To resolve this type of noise a model should look at the broader conversational context.

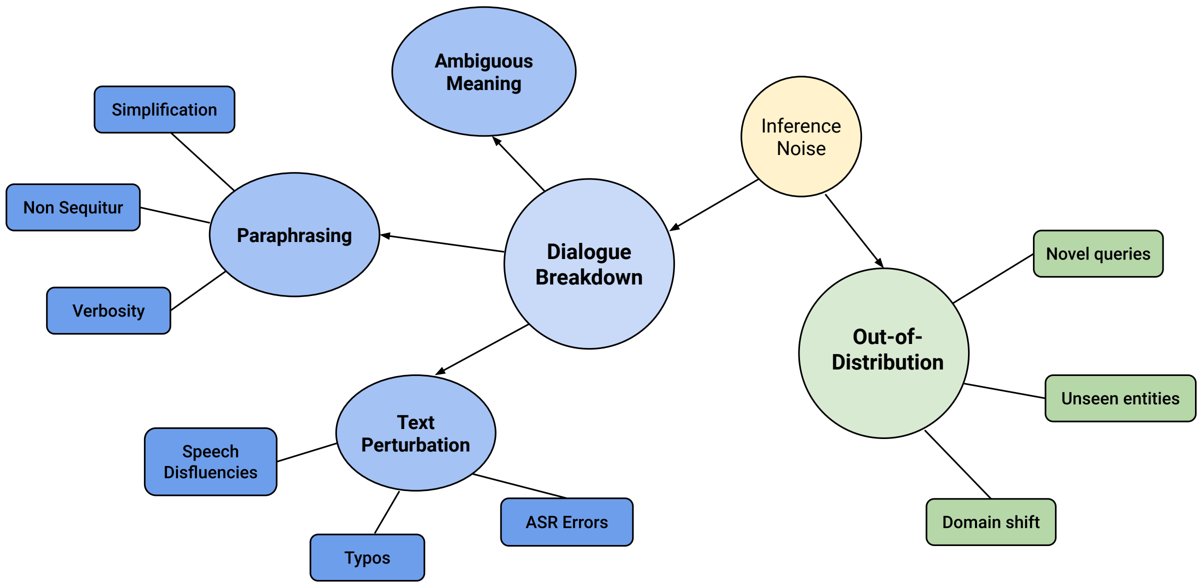

Figure 3: Diagram of the sources of noise that affect dialogue systems during inference.

Paraphrasing The text is rephrased to become: (a) *Simplification*: request may be simplified or shortened that makes it unclear what the user wants. (b) *Non Sequitur*: response is plausibly in-distribution, but does not reasonably answer the question. (c) *Verbosity*: request is so verbose that the underlying request is lost. (See Appendix D)

Text Perturbations Notable instances include (a) *ASR Errors* that fail to "wreck a nice beach" (recognize speech) (b) *Typos* and other syntax errors on the user input. This is distinct from formatting mistakes by annotators, which are errors on the target output. (c) *Speech Disfluencies* such as repeats, corrections, or adding 'umm' to start a utterance (Liu et al., 2021; Peng et al., 2021).

4 Noise Patterns

Beyond categorization, manually reviewing 10K+ utterances also provides unique insights.

How often does noise appear? The percentage of dialogues with at least one instance of noise comes out to an average of 11.2%, a median of 10.6%, with a standard deviation of 3.7%. However, given the approximate nature of sampling, the extra digits may not be significant. Instead, we assert the rate of noise in curated dialogue datasets is usually over 5%, rarely above 20% and typically around 10%. Since these rates are relatively low, denoising techniques aiming to combat extremely high levels of noise may be impractical.

What noise types are most common? While most existing denoising algorithms are designed to resolve class confusion (Sukhbaatar et al., 2015; Patrini et al., 2017; Goldberger and Ben-Reuven, 2017), our analysis reveals that instance-level noise is actually much more common, showing up in nearly 10% of cases compared to just 5% for class-level errors. Class-level noise assumes a latent noise transition matrix stochastically switches labels from one class to another. However, the prevalence of instance-level noise implies that the more likely explanation is that some examples are simply more confusing then others due to the genuinely ambiguous nature of dialogue (Pavlick and Kwiatkowski, 2019; Nie et al., 2020)

From an algorithmic perspective, the upshot is that developing denoising methods to target individual examples rather than class errors are likely to be most effective. Furthermore, we discovered that noise is clustered rather than evenly distributed, so filtering out or relabeling these particularly noisy instances should have an out-sized impact.

Why is X source of noise missing? The expected influence of some sources of noise are greatly exaggerated. Building out the taxonomy not only shows the most likely sources of noise, but equally notable is uncovering the *least* likely noise types. Concretely, the threat of adversarial actors is largely overblown (Dinan et al., 2019a), as spam-like activity appears less than 2% of the time. Offensive speech is the subject of numerous dialogue studies (Khatri et al., 2018; Xu et al., 2021; Sun et al., 2022), but is practically non-existent in reality (<0.5% of cases). While hate speech may be a problem when training on raw web text (Schmidt and Wiegand, 2017), our empirical review reveals that toxic language is exceedingly rare in curated datasets. Instead, unnatural utterances generated by crowdworkers role-playing as real users occurs much more often. (Full breakdown in Appendix E)

Other types of noise occur so infrequently that they are missing from the taxonomy completely! Noteworthy options include inconsistent *names* or *titles* within the ontology (See Appendix C), as well as *improper reference* texts for dialogue generation tasks. While these noise types are possible, they did not occur in practice. We intentionally exclude all such candidates from the taxonomy since the aim is not to be comprehensive, but rather to highlight where researchers should spend their efforts.

Where does noise come from? Our survey found that each data collection method had a propensity to produce certain kinds of noise. This suggests noise arises as a result of how examples are annotated, rather than other factors such as conversation length (number of utterances) or dialogue style (open-domain vs. task-oriented). For example, positive discourse attributes are most common with Post-conversation Annotation and Live Chat, which involve two human speakers engaging in real dialogue. Wizard-of-Oz datasets are less time-consuming to produce, but contain more label noise. In contrast, dialogues from Machine-to-Machine or Dialog Self-play (ie. starting with the labels to generate the dialogue) contain fewer label errors, but also sound less natural. Separately, annotator and ontology issues can be mitigated with well-written agent guidelines and proactive crowdworker screening. Thus, practitioners should consider these noise trade-offs when collecting dialogue data.

5 Experiments and Results

This section explores to what degree various models and dialogue tasks are impacted by each of the seven different categories of noise outlined in Section 3. To study this, a model is trained on a clean version of the dataset and on a corrupted version with either natural or injected noise. The level of corruption for all trials is held constant at 10% to allow for comparison across noise types. Datasets for each noise type are selected to maximize the overall variety, while always keeping one instance of MultiWOZ 2.3 to aid comparison. Intuitively, sources of noise that induce a larger gap in models trained on cleaned versus corrupted data are more significant, and consequently deserve more attention as targets to denoise.

5.1 Task Setup

All trials are conducted with GPT2-medium as a base model (Brown et al., 2020). The chosen tasks are: (1) *Conversation Level Classification* (CLC) – Choose from a finite list of labels for each conversation. (2) *Turn Level Classification* (TLC) – Make a prediction for each turn that contains a label. (3) *Dialogue State Tracking* (DST) – Predict the overall dialogue state, which may contain multiple slot-values or no new slot-values at all. Individual values come from an enumerable or open-ended ontology. (4) *Response Generation* (RG) – Produce the agent response given the dialogue context so far. (5) *Information Retrieval* (IR) – Find and rank the appropriate information from an external data source, such as a knowledge base (KB) or separate document. Metrics were chosen to adhere to the evaluation procedure introduced with the original dataset or from related follow-up work.

5.2 Noise Injection

For each noise category, we start by independently sampling 10% of the data, adding the corresponding noise and training a model to convergence. For example, consider instance-level label errors applied to MultiWOZ. This dataset contains 113,556 total utterances so 11,356 of them are selected for corruption. Next, one of the three sub-categories of instance noise are chosen uniformly at random. Over-labeling occurs when a label that has recently appeared in previous turns is no longer valid. To match this behavior, we keep a running tally of recent slot-labels and occasionally insert an extra one from this pool into the current training example. Partial-labeling is achieved by replacing a slot-label with a randomly selected one from the recent pool, and under-labeling is achieved by simply dropping a slot-label from the example. Finally, a model is trained with the noisy data applying the same hyper-parameters as the ones used for training the standard, original model. This process is repeated for each other noise type, with details for each source of noise found in Appendix F.

5.3 Main Results

Denoising methods targeting class-level noise may have minimal impact since it turns out such label er-

Noise Source	**MultiWoz**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Average**
Label Noise by Class	84.1 (0.13%)	75.8 (0.37%)^{DD}	58.1 (1.15%)^{ED}	78.8 (1.92%)^{MIDAS}	0.89%
Label by Instance	59.1 (4.88%)	82.4 (3.03%)^{SGD}	72.9 (0.96%)^{TT}	98.9 (0.12%)^{GSIM}	2.25%
Label by Annotator	58.2 (18.1%)	73.6 (3.36%)^{DD}	90.2 (1.43%)^{TT}	44.7 (15.9%)^{WOW}	9.68%
Discourse Attributes	62.9 (9.31%)	36.8 (8.42%)^{WOW}	25.6 (5.08%)^{ABCD}	39.2 (10.7%)^{KVRET}	8.38%
Ontology Inconsistency	61.9 (3.41%)	98.7 (0.40%)^{GSIM}	58.7 (26.8%)^{ED}	84.9 (0.94%)^{SGD}	7.89%
Out-of-Distribution	48.1 (28.9%)	83.2 (2.04%)^{SGD}	83.3 (10.5%)^{ABCD}	74.6 (23.6%)^{SGD}	**16.3%**
Dialogue Breakdown	61.8 (11.3%)	49.8 (4.02%)^{WOW}	4.07 (4.44%)^{ED}	72.1 (2.08%)^{TT}	5.45%

Table 3: Performance across various datasets when injected with 10% noise. Scores in parentheses are the percent degradation when compared to the clean version of the data. Datasets 2-4 contain a superscript representing the dataset name as described in Table 1. Please see Appendix 5 for the exact task and dataset mapping for each item.

	RoBERTa	**GPT2**	**BART**
Original	45.7	61.9	62.3
Noised	39.4	59.1	61.4

(a) Performance on MultiWOZ for each model

	CLC	**TLC**	**DST**	**RG**	**IR**
Median	3.4%	0.9%	4.0%	10.3%	8.4%
Average	6.5%	4.6%	8.4%	10.5%	8.1%

(b) Change in performance for each task due to noise.

Table 4: Breakdown by dialogue task and model type

rors are not all that damaging with just 0.89% drop in performance. On the other hand, annotator noise is quite powerful causing a 9.7% disturbance and should be mitigated whenever possible. Luckily, our manual review showed that spamming behavior occurs infrequently in reality simply by following some best practices[1]. Negative discourse attributes can also cause major harm leading to a 8.4% gap.

Moving onto inference noise, ontology issues are not only quite common, but also have meaningful impact on performance, causing a 7.9% drop. Dataset creators can ameliorate this by deciding on an ontology upfront, rather than creating one after the fact. Dialogue breakdowns also cause noticeable degradation, but the impact of OOD is most prominent among all noise types. Neural networks are powerful enough to learn from any training signal, even complete random noise (Zhang et al., 2017). However, OOD cases are by definition areas the network has not seen, leading to poor performance. Data augmentation and other robustness methods may serve as a strong tool to cover the unknown space by maximizing the diversity of the examples (Ng et al., 2020; Chen and Yin, 2022).

[1]For example, gold checks insert questions with known labels; timers ensure adequate time is spent on each task.

5.3.1 Task Breakdown

In order to study tasks across noise types, we look at the percentage change between models, rather than absolute difference. Furthermore, to minimize the influence of outliers, we emphasize the median of change, rather than the average. The results in Table 4b show that RG and IR observe the largest drops when noise is added. Somewhat surprisingly, CLC has larger performance shift than TLC despite being an easier task. We hypothesize this is because CLC examples only occur once for each conversation, whereas TLC examples occur at every turn, leading to an order of magnitude less data. Training with the existence of noisy data depends on both the rate of noisy data as well as on a minimum number of clean examples.

5.3.2 Model Robustness

Prior work has suggested that models behave differently when faced with distinct types of noise (Belinkov and Bisk, 2018). In addition to GPT2-medium (345M parameters), we also consider a masked language model in RoBERTa-Large (355M parameters) (Liu et al., 2019) and a sequence-to-sequence model with BART-large (406M parameters) (Lewis et al., 2020). These are selected due to having a comparable number of training parameters. Based on the results in Table 4a, RoBERTa is the weakest performer of the group. We hypothesize this is because many dialogue tasks are generation based, whereas BERT-based models typically perform well on classification. Conversely, BART deals quite well with noise, suggesting encoder-decoder models as reasonable starting points for future dialogue projects.

5.4 Amount of Noise

We simulate increasing levels of noise by adding instance-level label errors and incoherent discourse

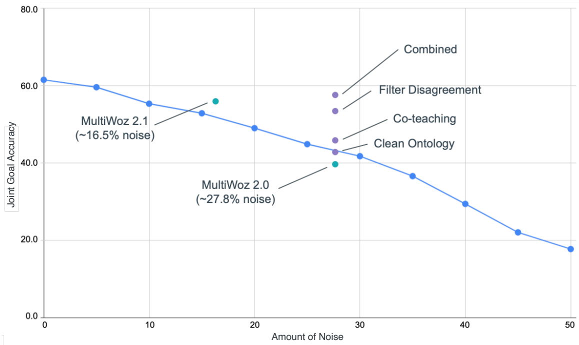

Figure 4: Impact of injecting different amounts of label and discourse noise to MultiWOZ dataset. MultiWOZ 2.3 is defined to be 0% noise. MultiWOZ 2.0 and 2.1 have estimated noise levels based on the ratio of labels that are changed compared to MultiWOZ 2.3 data.

attributes to the MultiWOZ 2.3 dataset (Han et al., 2021), which we define to be noise-free. We additionally plot the performance of models trained on MultiWOZ 2.0 (Budzianowski et al., 2018) and 2.1 (Eric et al., 2020), where all models are evaluated on the MultiWOZ 2.4 test set (Ye et al., 2021). Looking at Fig 4, we first note that scores on naturally noisy data from MWOZ 2.0 and MWOZ 2.1 fall close to the plotted trajectory, lending credence to the overall trend. Furthermore, we notice that as we vary the amount of noise, model performance decreases logarithmically, but surprisingly does not have a tipping point at which it fails to converge.

6 Dialogue Denoising

Informed by our understanding of the sources of dialogue noise, we now design a preliminary denoising algorithm for learning in the presence of noisy labels. We select MultiWOZ to serve as our testbed not only because it is one of the most popular dialogue datasets, but also because it is representative of how noise affects most datasets in general (see Figure 6). While our method produces promising results, our aim is not to declare the noise issue solved, but rather to establish a baseline others can further improve. (More details in Appendix G.)

6.1 Algorithm

Based on analysis in Section 3, MultiWOZ 2.0 is most plagued by three types of errors: ontology inconsistencies, instance label errors and out-of-distribution issues. We now devise three solutions to resolve each source of noise accordingly.

(1) To clean up the *ontology*, we drop values that do not conform to the correct format, and remove the associated examples from training. For example, if `time_of_day` slot expects the HH:MM format, then we remove all values referencing day formats (e.g. Friday). (2) To deal with label errors, we filter out individual *instances* where the predicted label from a pre-trained GPT2-medium model disagrees with the annotator label (Cuendet et al., 2007; Jiang et al., 2018; Chen et al., 2019). We calibrate the model with temperature scaling to prevent it from being over-confident in its predictions (Guo et al., 2017). (3) To counteract issues caused by *OOD*, we augment our training data by pseudo-labeling the examples stripped out in the first two steps. When the model used for filtering is also used for pseudo-labeling, biases may propagate across each iteration. As a result, inspired by co-teaching (Han et al., 2018), we instead use a different BART-base model for pseudo-labeling to force divergence of model parameters and avoid errors from accumulating.

6.2 Denoising Results

We once again evaluate with MultiWOZ 2.4 since this is the cleanest version of test data. As seen in Figure 4, we are able to outperform MultiWOZ 2.0 (39.8) by 16.9% absolute accuracy and 42.5% relative accuracy. Ontology Clean (43.2), Filter Disagree (53.7) and Co-teaching (46.7) all show marked improvement over the original baseline, but Combined (58.6) does the best overall, reaching a score that even surpasses MultiWOZ 2.1 (56.5). These initial efforts show our ability to successfully identify and counteract sources of noise within MultiWOZ, which we encourage others to build upon.

7 Related Works

Our work is related to efforts to categorize noise within speech and dialog. Clark (1996) proposed a theory of miscommunication consisting of channel, signal, intention and conversation where each of the four levels serves as a potential vector for noise. Others have also studied noise in spoken dialogue systems, where they found that the main culprit stems from errors in speech transcription (Paek, 2003; Bohus, 2007). Rather than a high-level framework of general communication, our hierarchical taxonomy focuses on understanding the multiple layers of noise found in written text.

More recent works on dialogue noise discuss robustness to noisy user inputs, whereas we expand this view to also analyze noisy training inputs. Peng et al. (2021) introduce RADDLE as a platform which covers OOD due to paraphrasing, verbosity, simplification, and unseen entities, as well as general typos and speech errors. Liu et al. (2021) create a robustness benchmark which considers paraphrasing through word perturbations as well as speech disfluencies. Lastly, Krone et al. (2021) considers noise from abbreviations, casing, misspellings, paraphrasing, and synonyms.

7.1 Survey of Denoising Methods

Most prior works exploring learning with noisy labels were originally developed for the computer vision domain (Smyth et al., 1994; Mnih and Hinton, 2012; Sukhbaatar et al., 2015). Some methods model the noise within a dataset in order to remove it, often through the use of a noise transition matrix (Dawid and Skene, 1979; Goldberger and Ben-Reuven, 2017). Others have designed noise-insensitive training schemes by modifying the loss function (van Rooyen et al., 2015; Ghosh et al., 2017; Patrini et al., 2017), while a final set of options manipulate noisy examples by either reweighting or relabeling them. (Reed et al., 2015; Jiang et al., 2018; Li et al., 2020). While denoising work certainly exists for NLP (Snow et al., 2008; Raykar et al., 2009; Wang et al., 2019), none of them specifically touch upon the dialogue scenario.

7.2 Denoising by Source of Noise

To support the effort of designing improved algorithms for combating dialogue-specific noise, we highlight potential methods that can be adapted to deal with the noise categories identified by our taxonomy in Section 3. To start, a common technique for dealing with *class-level* errors is to learn a noise adaptation layer to recognize label noise (Goldberger and Ben-Reuven, 2017). For *instance-level* noise, besides filtering by disagreement, core-set selection (Mirzasoleiman et al., 2020) or the Shapley algorithm (Liang et al., 2021) can be used to identify important datapoints and thereby remove the noisy ones. Modeling the likelihood of *annotator-level* error in order to reverse its impact is also worth considering (Welinder et al., 2010; Hovy et al., 2013; Guan et al., 2018). Next, a model trained on NLI data can be used to screen out inconsistent *discourse* examples (Welleck et al., 2019). A model trained on Prosocial Dialogue data can learn to reduce toxicity (Kim et al., 2022). In terms of *discourse fluency*, one can train a student model to reweight its logits during inference based on a large language model (Brown et al., 2020) to improve the fluency of the student. Another method is to create an *ontology* upfront which defines the allowed entities before data collection and enforcing this by having checks upon label submission. *Out-of-Domain* issues can be handled with the use of more examples to increase the coverage and diversity of the solution space to limit OOD errors. This can be tackled by performing data augmentation on the in-domain (Feng et al., 2021) or out-of-domain examples (Chen and Yu, 2021). Lastly, *dialogue breakdown* can be mitigated by screening for annotators through minimum acceptance rates, language filters, and pre-qualifications quizzes (ie. quals).

8 Conclusion

This paper categorizes the different sources of noise found in dialogue data and studies how models react to them. We find that dialogue noise is divided into issues that occur during training and during inference. We also find that conversations pose unique challenges not found in other NLP corpora, such as discourse naturalness and dialogue breakdowns. Our study further reveals that the most common sources of noise are actually based on the ambiguity of individual instances, rather than systematic noise across classes or adversarial annotators actively harming data collection efforts.

Despite being surprisingly resilient, dialogue models nonetheless experience a notable drop in performance when exposed to high levels of noise. To combat this, we design a proof-of-concept denoising algorithm to serve as a strong foundation for others to compare against. We apply this algorithm successfully to the MultiWOZ 2.0 dataset, raising the accuracy by 42.5% over the original baseline. We hope our survey informs the collection of cleaner dialogue datasets and the development of advanced denoising algorithms targeting the true sources of dialogue noise.

9 Limitations

The main limitation of the taxonomy is only considering natural language text within dialogue. It could be useful to conduct a detailed breakdown of speech noise or multi-modal noise that occurs in conversations grounded by images. Our survey also does not include all theoretically possible sources of noise and instead is limited to actual sources of noise saw occuring in the data. We argue this type of taxonomy serves a more practical purpose.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Dan Bohus. 2007. *Error awareness and recovery in conversational spoken language interfaces*. Ph.D. thesis, Carnegie Mellon University.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4515–4524. Association for Computational Linguistics.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3002–3017. Association for Computational Linguistics.

Derek Chen and Claire Yin. 2022. Data augmentation for intent classification. *CoRR*, abs/2206.05790.

Derek Chen and Zhou Yu. 2021. GOLD: improving out-of-scope detection in dialogues using data augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 429–442. Association for Computational Linguistics.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. 2019. Understanding and utilizing deep neural networks trained with noisy labels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1062–1070. PMLR.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Sébastien Cuendet, Dilek Hakkani-Tür, and Elizabeth Shriberg. 2007. Automatic labeling inconsistencies detection and correction for sentence unit segmentation in conversational speech. In *Machine Learning for Multimodal Interaction , 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers*, volume 4892 of *Lecture Notes in Computer Science*, pages 144–155. Springer.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 422–428. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, François Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 37–49. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 968–988. Association for Computational Linguistics.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press.

Jacob Goldberger and Ehud Ben-Reuven. 2017. Training deep neural-networks using a noise adaptation layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3109–3118. AAAI Press.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In *Natural Language Processing and Chinese Computing - 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13-17, 2021, Proceedings, Part II*, volume 13029 of *Lecture Notes in Computer Science*, pages 206–218. Springer.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 87–95. The Association for Computer Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1120–1130. The Association for Computational Linguistics.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2309–2318. PMLR.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584.

Chandra Khatri, Behnam Hedayatnia, Rahul Goel, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Detecting offensive content in open-domain conversations using two stage semi-supervision. *NeurIPS Workshop on ConvAI*, abs/1811.12900.

Ashish Khetan, Zachary C. Lipton, and Animashree Anandkumar. 2018. Learning from noisy singly-labeled data. In *6th International Conference on*

Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *EMNLP*.

Jason Krone, Sailik Sengupta, and Saab Mansoor. 2021. On the robustness of goal oriented dialogue systems to real-world noise. In *Robust ML Workshop at ICLR 2021*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Weixin Liang, Kaihui Liang, and Zhou Yu. 2021. HERALD: an annotation efficient method to detect user disengagement in social conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3652–3665. Association for Computational Linguistics.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness testing of language understanding in task-oriented dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2467–2480. Association for Computational Linguistics.

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

Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. 2020. Coresets for robust training of deep neural networks against noisy labels. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Volodymyr Mnih and Geoffrey E. Hinton. 2012. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1196–1204.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9131–9143. Association for Computational Linguistics.

Tim Paek. 2003. Toward a taxonomy of communication errors. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, page 53–58.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2233–2241. IEEE Computer Society.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021. RADDLE: an evaluation benchmark and analysis platform for robust task-oriented dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4418–4429. Association for Computational Linguistics.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2639–2649. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8689–8696. AAAI Press.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna K. Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 889–896. ACM.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2015. Training deep neural networks on noisy labels with bootstrapping. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gökhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers)*, pages 41–51. Association for Computational Linguistics.

Padhraic Smyth, Usama M. Fayyad, Michael C. Burl, Pietro Perona, and Pierre Baldi. 1994. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 1085–1092. MIT Press.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1297–1304.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263. ACL.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. Training convolutional neural networks with noisy labels. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 843–852. IEEE Computer Society.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

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.

Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. *CoRR*, abs/1505.07634.

Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. 2019. Learning with noisy labels for sentence-level sentiment classification. In *Proceedings of the 2019 Conference on Empirical Methods* *in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6285–6291. Association for Computational Linguistics.

Peter Welinder, Steve Branson, Serge J. Belongie, and Pietro Perona. 2010. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 2424–2432. Curran Associates, Inc.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3731–3741. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *CoRR*, abs/2104.00773.

Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain humanmachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1103–1120. Association for Computational Linguistics.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7620–7632. Association for Computational Linguistics.

A Dataset Descriptions

In no particular order, the datasets we study are:

1. DailyDialog (DD) - a collection of conversations from the web about everyday events, curated for teaching English. (Li et al., 2017)

2. Wizard of Wikipedia (WoW) - a wizard reads an article on Wikipedia and then talks about it with their partner (Dinan et al., 2019b)

3. Google Simulated Dialogue (GSIM) - a large scale Machine-to-Machine (M2M) dataset build through paraphrasing, covers movie and restaurant domains. (Shah et al., 2018)

4. Action Based Conversations Dataset (ABCD) - customer service conversations that mimic agents in real-life where actions are taken to resolve customer issues based on instructions in agent guidelines (Chen et al., 2021)

5. MultiWoz 2.0 (MWoz) - a multi-domain dialogue dataset. Note that we use the original version for initial analysis because it contains true noise, before any additional cleaning. (Budzianowski et al., 2018)

6. TicketTalk (TT) - As part of the third installment of TaskMaster, this dataset also uses the M2M style, but focuses on the single vertical of movie ticket booking. (Byrne et al., 2019)

7. Empathetic Dialogues (ED) - a set of dialogues that aim to teach models to be empathetic by being more attuned to what a user is feeling. (Rashkin et al., 2019)

8. Machine Interaction Dialog Act Schema (MIDAS) - created for the Amazon Alexa challenge with Gunrock. Transcribed conversations are with actual Alexa users, and not crowdworkers. (Yu and Yu, 2021)

9. Schema Guided Dialogue (SGD) - the most comprehensive DST dataset to date, with a heavy focus on slot-filling for API calls. Contains natural OOD splits. (Rastogi et al., 2020)

10. Key-Value Retrieval for In-Car Assistant (KVRET) - Task oriented dataset with a knowledge base for querying items. Covers navigation, weather and scheduling domains. (Eric et al., 2017)

B Label Error Details

Class Level Examples are labeled incorrectly due to confusion with another class.

- *Uniform Label Swapping*: symmetric noise where all classes have equal likelihood to be confused with any other class. The assumption is that noise is injected through a randomly initialized noise transition matrix.
- *Structured Label Swapping*: asymmetric noise where certain classes are more likely to confused with other related classes. For example, a cheetah is more likely to be confused with leopard than a refrigerator when performing image recognition. Alternatively, dogs and wolves are likely to be confused for each other much more often than with horses since those animals are similar to each other.

Instance Level Noise comes from the example itself due to the complexity of interpreting natural language. This is the realization that even when annotators act in good-faith, mistakes are still made since the instances themselves are difficult to label. Errors must be determined on a case-by-case basis.

- *Over Labeling*: annotator added a label, but should be removed since it is unnecessary. Example: carrying over a slot-value from the previous turn to the current dialogue state when it is not warranted.
- *Under Labeling*: annotator missed the label, when most people would include it. Example: failing to notice a newly mentioned criteria in the dialogue state. This also includes cases where a better label could have been used, but the option is missing from the ontology and consequently prevents the example from being properly labeled.
- *Partial Labeling*: part of the label is correct, but other parts are not. For multi-intent utterances, the annotator may have captured one intent, but not the other. For slot-filling tasks, the annotator may have selected the appropriate value, but assigned it to the wrong slot.

Annotation Level Noise arises due to the labeler or data collection process. (Snow et al., 2004)

- *Distant Supervision*: the noise results from the fact that the label is not from a human, but rather weakly labeled from distant supervision (Sun et al., 2017). For example, using a gazetteer for labeling named entities in NER. As another example, you use the SQL results to train a semantic parser, rather than an annotated SQL query.

- *Adversarial Actors*: meant to mimic spammers, this is characterized by repeating patterns or irrational behavior. For example, the annotator selects "greeting" dialogue act as the label for every single utterance regardless of the underlying text. (Raykar et al., 2009; Hovy et al., 2013; Khetan et al., 2018) Other examples include bad actors in social media who provoke chatbots into producing unsafe content or labelers who mark every review as possessing positive sentiment without actually reading the passage.
- *Formatting Mistakes*: Caused by non-experts making human mistakes, which are independent of the dialogue context. For example, typos or off-by-one errors, such as when the labeler failed to highlight the entire phrase during span selection. (See Table 2)

C Ontology Inconsistency Details

Another source of noise comes from inconsistent formatting when constructing the ontology. More specifically, the creators of the dataset did not set a canonical format for each type of slot being tracked. While we can imagine many other slot-types causing issues, the types of errors which actually occurred in practice include:

- **Dates**: tomorrow, Jan 3rd, 1/3/2022, Monday, January 3, mon
- **Times**: 14:15, 2:15 PM, quarter past 2, 215pm
- **Locations**: NYC, New York, ny, the big apple
- **Numbers**: three, 'wife daughter & I', 3, 'Me and my two buddies'.

Other ontology issues which we thought might occur more often, turn out to happen very rarely. For example, naming inconsistency such as [Fred Miyato, Mr. Miyato, fred miyato, my father] did not really occur. Titles of people or places [Macdonalds, MickeyD's, McDonald's, mcdonalds] also were not present. To minimize the amount of noise from ontology inconsistency, a recommendation is to declare the allowable slot-values upfront before data collection begins.

D Paraphrasing Examples

Paraphrasing can take on three general forms:

1. **Simplification** – the request may be simplified so much that it becomes unclear what the user wants. For a restaurant scenario:

   `Agent: What part of town would you like to eat?`
   `User: W`
   (as a shorthand for West side)

2. **Non Sequitur** – response is plausibly in-distribution, but does not reasonably answer the question.

   `Agent: What part of town would you like to eat?`
   `User: I would like Italian food.`

   Note that the user's response is still in distribution since it could have been a reasonable answer to "What cuisine do you prefer?". However, in this instance, this type of response is very noisy because it fails to answer the agent's question.

3. **Verbosity** – the request contains extra words or entities, which makes it confusing as to exactly what the answer may be.

   `Agent: What part of town would you like to eat?`
   `User: I prefer food in the East, but I live in the South right now.`

   In this case, the user's response is not necessarily long, but it is verbose enough to make it unclear whether the user wants food in the east side of town or the south side of town.

True paraphrasing noise should alter the text without altering the user's underlying intent. If the text has changed so much that the user's intent has also shifted, then it should be considered adversarial behavior beyond the scope of typical dialogue noise.

`Agent: What part of town would you like to eat?`
`User: The Northern Lights are beautiful this time of year.`

The example above displays positive sentiment, but the user has completely ignored the agent's request. This case borders on being incoherent and fails to move the dialogue forward.

E Results Breakdown

Aggregated amounts of noise by each sub-category:

	Average	Median	Std. Dev.
Class-level	4.9%	3.8%	0.7%
Instance-level	9.7%	6.9%	5.4%
Annotator-level	1.8%	0.7%	2.1%
Dates	3.6%	0.5%	6.3%
Times	1.1%	<0.1%	2.0%
Locations	1.3%	0.3%	2.1%
Numbers	2.3%	0.2%	4.6%
Incoherent	3.4%	3.8%	1.9%
Disfluent	2.6%	2.4%	2.0%
Inconsistent	1.7%	1.3%	1.5%
Nonsensical	2.0%	2.6%	1.1%
Offensive	0.2%	<0.1%	0.9%
Unnatural	4.8%	5.8%	1.6%
Overall	11.2%	10.6%	3.7%

Table 5: Breakdown across noise sub-categories

F Noise Injection Methods

Class-level Label Errors We create a noise transition matrix to mimic structured confusion. Specifically, given a certain class label, we want to determine what is likely to be confused with it so we can substitute the current label for that other class. To fill the noise transition matrix, we embed all class labels into bag-of-word GloVe embeddings and measure their similarity to other classes by cosine distance. Then, for 10% of examples, we sample an incorrect label given the original class according to the likelihood in the transition matrix.

Instance-level Label Errors To match the behavior of over-labeling, we keep a running tally of recent labels and occasionally insert an extra one from this pool into the example. Partial-labeling is achieved by replacing a label from the recent pool, and under-labeling is achieved by simply dropping a random label from the example.

Annotator-level Label Errors We mimic spammers who apply preset answers to every occasion without considering the actual dialogue. For the classification tasks, we assume a spammer randomly picks from one of the three most common labels for that task as the noisy target label. For response generation tasks, we assume a spammer randomly responds with one of three generic phrases.

Undesirable Discourse Attributes We replace a subset of the utterances with noisy versions 10% of the time. Incoherent utterances are randomly selected sentences from other dialogues within the dataset. Disfluent utterances are generated by shuffling the tokens within the current utterance. Unnatural utterances are generated by selecting from a list of awkward sentences referencing the task.

Ontology Inconsistency To clean the data, we manually remove entries that do not comply to the proper format. We also merge similar categories to create more compact ontologies. Training examples that are covered by the remaining entries are considered the clean version, while the full, original dataset is considered the noisy version.

Out-of-Distribution Multi-domain data is divided such that training data contains a subset of domains while the test set includes examples from all domains. Choosing the domains to exclude was straightforward for ABCD and SGD since they are given by the task design. Rather than choosing an arbitrary domain to leave out for MWOZ, we instead run the experiment once for each domain, and report the average of the five results.

Dialogue Breakdown We reproduce this behavior by pre-training a paraphrase model and applying it to perturb 10% of utterances. Paraphrase model is trained on QQP, MRPC and PAWS corpora.

G Denoising Procedure for MultiWOZ

We identify the highest likelihood sources of noise for any given dataset and dealing with each one accordingly. MultiWOZ in particular has (1) ontology issues, (2) instance level label errors and (c) out-of-distribution examples caused by low coverage in the training set. In turn, we proceed to deal with each of these issues as follows:

(1) To clean up the *ontology*, we drop values that do not conform to the correct format for their given slots, and remove the associated examples from training. For example, if the slot is a time of day expecting the HH:MM format, then we remove all values referencing ‘Friday’ or ‘afternoon’ which are incorrectly formatted.

(2) To deal with possible label errors, we filter out individual *instances* where the predicted label from a pre-trained GPT2-medium model disagrees with the annotator label (Cuendet et al., 2007; Jiang et al., 2018; Chen et al., 2019).

(3) Lastly, we augment our training data to counteract issues caused by *OOD* cases. In order to augment, we pseudo-labeling the datapoints that have been stripped out in the first two steps. However,

Noise Source	**MultiWoz 2.3**	**Dataset 2**	**Dataset 3**	**Dataset 4**
Label Noise by Class	MWOZ (TLC on intents)	DD (CLC on topics)	ED (CLC on emotions)	MIDAS (TLC on dialog acts)
Label by Instance	MWOZ (DST on slot-values)	SGD (DST w/ slot-values)	TT (DST w/ slot-values)	GSIM (TLC on user acts)
Label by Annotator	MWOZ (RG of agent utt)	DD (CLC on topics)	TT (TLC on APIs)	WOW (RG on wizard utt)
Discourse Attributes	MWOZ (RG of agent utt)	WOW (IR on wizard utt)	ABCD (IR on agent utt)	KVRET (IR on KB entries)
Ontology Inconsistency	MWOZ (DST on slot-values)	GSIM (TLC on user acts)	ED (CLC on emotions)	SGD (DST on slot-values)
Out-of-Distribution	MWOZ (DST on slot-values)	SGD (DST on slot-values)	ABCD (CLC on subflows)	SGD (TLC on intents)
Dialogue Breakdown	MWOZ (RG of agent utt)	WOW (RG on wizard utt)	ED (RG on agent utt)	TT (DST on slot-values)

Figure 5: Mapping of model performance to datasets and dialogue tasks. Parentheses also includes the target of the task. For example, 'CLC on topics' means that the task is to classify the associated topic label at a conversation level, while 'TLC on intents' means the task is to classify the intent of each user turn.

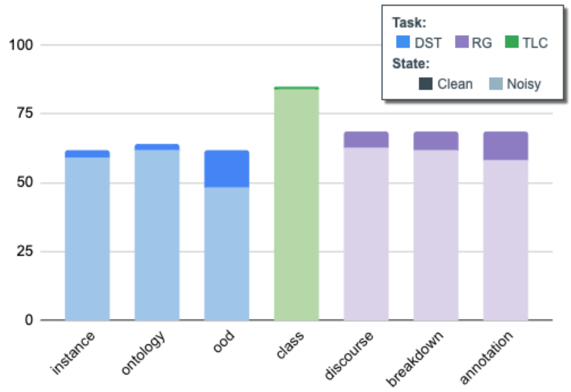

Figure 6: Impact of the different noise types on the MultiWoz2.3 dataset. DST is dialogue state tracking, RG is response generation and TLC is turn level classification.

the pretrained model's predictions are unlikely to be all correct, so rather than keep all the new labels, we only keep the examples where the probability of the max value crosses the 0.5 threshold. Then, since neural networks are often over-confident, we perform calibration with temperature scaling using a λ parameter (Guo et al., 2017). However, pseudo-labeling with the same model that is used to perform filtering causes errors to propagate which hinders performance gains. As a result, inspired by co-teaching (Han et al., 2018), we instead use a different model to force divergence of model parameters and avoid the existing biases. In more detail, we rely on a BART-base model rather than the original GPT-2 medium, which works even though BART-base has much fewer parameters.

H Noise as Uncertainty

An interesting way to view the impact of noise is through the lens of Bayesian uncertainty. In particular, aleatoric and epistemic uncertainty can be seen caused by different types of noise. Kendall and Gal (2017) describe aleatoric uncertainty as uncertainty which "captures noise inherent in the observations." In contrast "epistemic uncertainty accounts for uncertainty in the model parameters which can be explained away given enough data."

Roughly speaking, labeling errors cause epistemic uncertainty since these errors produce uncertainty in the model parameters. If given enough clean data to train a model, the issues caused by the noisy labels should largely be erased. In other words, epistemic uncertainty describes what the model does not know because training data was not appropriate, so by resolving the labeling errors, the training data is now appropriate and the dialogue system can be trained successfully.

On the other hand, ontology inconsistencies cause aleatoric uncertainty since they can lead to situations where it is impossible to fix the problem by altering the training data alone. Suppose we want the dialogue model to predict the desired time for a restaurant reservation (such as 11 AM, 6PM or 8PM), but options such as 'Sunday' or 'afternoon' keep appearing, which are never correct. This would make it harder for a classifier to choose the correct time. In the degenerate case, suppose the ontology only consisted of days of the week such as 'Monday', 'Wednesday' or 'Friday', such that the classifier would only have the ability to choose from seven incorrect options. In this case, adding any amount of extra data (even those labeled in the correct format) would do nothing to resolve the issue since the problem itself has been modeled incorrectly.

Accordingly, a model developer should focus on eliminating certain types of noise based on the type of uncertainty they are seeing in their dialogue system. If the model is consistently making a handful of random mistakes, then relabeling some data or collecting new data may resolve the issue. Alternatively, if the model is a making systematic errors then looking into the ontology or data collection procedure might be a better route.

I Additional Noise Examples

Examples were carefully selected to give good coverage of the different types of noise that occurred frequently within the data.

Dataset	Noise Type	Dialogue	Comments
WOW	Labeling Error ↪ Instance level ↪ Under	*Apprentice* I have visited the United States. To New York City, Los Angeles, and Seattle for work and vacation. Every city was unique with its own culture and loved every one of them. *Wizard* I haven't been to the East coast yet, but I have been to Los Angeles, which is Spanish for "The Angels" *Apprentice* Oh I never knew. The East coast always felt busier, the West coast felt more relaxed. *Wizard* Agreed! I grew up in Hawaii, where the life expectancy is amongst the highest in the nation. Do you like large cities or smaller towns?	*Correct label*: {topic: 'Los Angeles'} *Possible:* *missing labels:* {topic:'Hawaii' 'longevity'}
DD	Labeling Error ↪ Class level ↪ Uniform	*A* Hello Mike ! Would you like a drink ? *B* No, thank you. I had too much to drink yesterday evening. I had a bad hangover this morning. My head felt terrible. (happiness) *A* Were you celebrating something ? *B* Yes. It was a friend's birthday party. We drank all kinds of things - beer, wine and spirits. After midnight, we even drank cocktails! *A* It's a bad idea to drink a combination of alcoholic drinks. You should stick with one for the whole evening.	*Revise label*: happiness → disgust
ED	Discourse Attribute ↪ Disfluent	*A* I've got popcorn kernels to last me through retirement. I wonder how long they keep for. *B* That is nice. *A* Yea, it is. Do you like popcorn? *B* Yes. Why did you bought that many popcorn kernels?	*grammar mistake*

MWOZ	Ontology Inconsistency
	↪ Location
	
	Labeling Error
	↪ Annotator
	↪ Formatting
		User I'm looking for a special place, can you help?
		`attraction(type=Special)`
		System I need just a little more information to help. I think all places in Cambridge are special ...
		User I am looking specifically for Saint John's College. `attraction(type=special, name=Saint John's College.)`
		System sorry i dont have that in our list. is there something else i can do for you?
		User Okay, well I also need a train departing for Cambridge on Wednesday.
		`train(dest=cambridge,day=wednesday)`
		System I have several trains headed to Cambridge on Wednesday. Where will you be departing from? <*truncated*>
			ontology:
			uppercase
			
			ontology:
			lowercase
			formatting:
			added an extra period

ABCD	Labeling Error ↪ Class level ↪ Structured	*Agent* Thank you for contacting acmebrands. how can I help you? *Customer* I need to check the status of an order *Agent* Okay. Could I have your name or Account ID please? *Customer* alessandro phoenix *Action* Account has been pulled up for alessandro phoenix. `[Pull-up Account]` *Agent* Thank you. Do you just need to know the order's shipping status? *Customer* The email that I received showed it is being shipped to the wrong address. is it being shipped to <street_address>, jacksonville, mi	*Revision:* The agent followed the 'manage shipping' flow when instead they should have followed the 'shipping status' flow
ED	Labeling Error ↪ Instance level ↪ Partial	*A* I liked the way Lebron was snatched up by the Lakers *B* Will be a game changer. *A* I hope so. I hope they can crush the warriors *B* Did the Warriors get a replacement?	*Emotion label*: impressed *Missing labels*: hopeful, anticipating
SGD	Discourse Attribute ↪Inconsistent, Nonsensical	*User* I want to check in on March 2nd, and stay in 1 room until the 12th. *System* Okay, please confirm: You want to stay in New York at 1 hotel central park, checking in tomorrow, and checking out on March 12th. you also want 1 room. *User* Actually 3 rooms please. *System* Okay, confirm: You want 3 rooms and you check in tomorrow? *User* No. I want to check out on the 13th. *System* Please confirm: You want to check out of your hotel in New York on March 13th. *User* Yes, works for me. *System* THe reservation is made. *User* Does it allow pets? How much does it cost? *System* It is pet friendly, and costs $4,488. *User* Thank you, i'm all set.	*Inconsistent* since the user wants 1 room than arbitrarily changes to 3. *Nonsensical* since you would want to know the price of a hotel room before booking, not afterwards.
MIDAS	Labeling Error ↪ Class level ↪ Structured	*User* actually *Agent* what was that experience like *User* i have not	*Revise label*: statement → neg answer

GSIM	Ontology Inconsistency
	↪ General
		System please confirm the date is next monday at 8 pm for ephesus restaurant for 3 people
		User yes i confirm the date is next monday at 8pm for ephesus restaurant for 3 people
		(multi-part label not needed)
			Original label:
			`AFFIRM_AFFIRM_`
			`AFFIRM_AFFIRM`
			Revised label:
			`AFFIRM`

KVRET	Ontology Inconsistency ↪ Date	*User* what is the weather like in the city *Assistant* Which city do you want the weather for? *User* Will it hail in Alameda today? *Assistant* Today in Alameda is gonna be drizzle, with low temperature of 70F, and high of 90F *User* Thank you! *Assistant* You are welcome.	`request - {date:today, loc:Alameda, weather:hail}` Date is annotated as "today", but should be a day of the week such as Monday, Tuesday