GOLD: Improving Out-of-Scope Detection in Dialogues using Data Augmentation

Derek Chen
ASAPP, New York, NY 10007
dchen@asapp.com

Zhou Yu
Columbia University, NY
zy2461@columbia.edu

Abstract

Practical dialogue systems require robust methods of detecting out-of-scope (OOS) utterances to avoid conversational breakdowns and related failure modes. Directly training a model with labeled OOS examples yields reasonable performance, but obtaining such data is a resource-intensive process. To tackle this limited-data problem, previous methods focus on better modeling the distribution of in-scope (INS) examples.

We introduce GOLD as an orthogonal technique that augments existing data to train better OOS detectors operating in low-data regimes. GOLD generates pseudo-labeled candidates using samples from an auxiliary dataset and keeps only the most beneficial candidates for training through a novel filtering mechanism. In experiments across three target benchmarks, the top GOLD model outperforms all existing methods on all key metrics, achieving relative gains of 52.4%, 48.9% and 50.3% against median baseline performance. We also analyze the unique properties of OOS data to identify key factors for optimally applying our proposed method.[1]

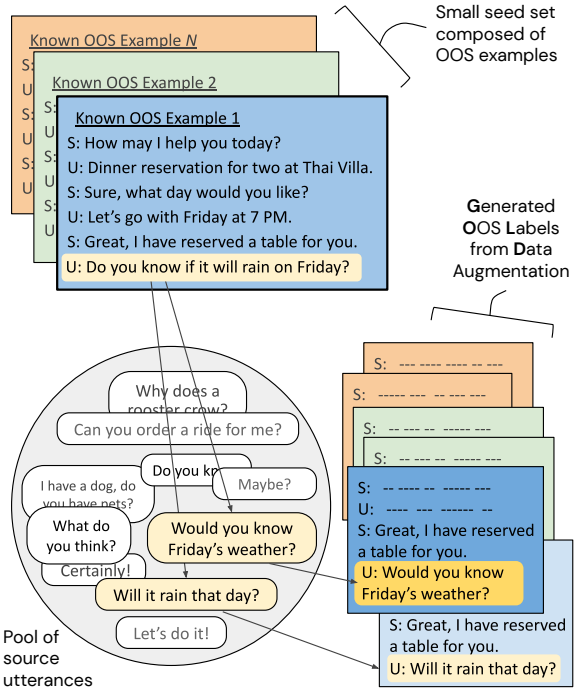

Figure 1: GOLD performs data augmentation by extracting utterances from a source dataset and merging those sentences with known OOS samples from the target dataset to generate pseudo-labeled OOS examples.

1 Introduction

Detecting out-of-scope scenarios is an essential skill of dialogue systems deployed into the real world. While an ideal system would behave appropriately in all conversational settings, such perfection is not possible given that training data is finite, while user inputs are not (Geiger et al., 2019). Out-of-distribution issues occur when the model encounters situations not covered during training, including novel user intents, domain shifts or custom entities (Kamath et al., 2020; Cavalin et al., 2020). Unique to conversations, dialogue breakdowns represent cases where the user cannot continue the interaction with the system, perhaps due to ambiguous requests or prior misunderstandings (Martinovsky and Traum, 2003; Higashinaka et al., 2016). Such breakdowns might fall within the distribution of plausible utterances, yet still fail to make sense due to the given context. OOS detection aims to recognize both out-of-distribution problems and dialogue breakdowns.

Prior methods tackling OOS detection in text have shown great promise, but typically assume access to a sufficient amount of labeled OOS data during training (Larson et al., 2019), which is unrealistic in open-world settings (Fei and Liu, 2016). Alternative methods have also been explored which train a supporting model using in-scope data rather than directly training a core model to detect OOS instances (Gangal et al., 2020). As a result, they

[1] All code and data for major experiments are available at https://github.com/asappresearch/gold

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 429–442
November 7–11, 2021.

suffer from a mismatch where the objective during training does not line up with the eventual inference task, likely leading to suboptimal performance.

More recently, data augmentation techniques have been applied to in-scope (INS) data to improve out-of-domain robustness (Ng et al., 2020a; Zheng et al., 2020). However, we hypothesize that since INS data comes from a different distribution as OOS data, augmentation on the former will not perform as well as augmentation on the latter.

In this paper, we propose a method of **G**enerating **O**ut-of-scope **L**abels with **D**ata augmentation (GOLD) to improve OOS detection in dialogue. To create new pseudo-labeled examples, we start with a small seed set of known OOS examples. Next, we find utterances that are similar to the known OOS examples within an auxiliary dataset. We then generate candidate labels by replacing text from the known OOS examples with the similar utterances uncovered in the previous step. Lastly, we run an election to filter down the candidates to only those which are most likely to be out-of-scope. Our method is complementary to other indirect prediction techniques and in fact takes advantage of progress by other methods.

We demonstrate the effectiveness of GOLD across three task-oriented dialogue datasets, where our method achieves state-of-the-art performance across all key metrics. We conduct extensive ablations and additional experiments to probe the robustness of our best performing model. Finally, we provide analysis and insights on augmenting OOS data for other dialogue systems.

2 Related Work

2.1 Direct Prediction

A straightforward method of detecting out-of-scope scenarios is to train directly on OOS examples (Fumera et al., 2003). These situations are encountered more broadly by the insertion of any out-of-distribution response or more specifically when a particular utterance does not make sense in the current context.

Out-of-Distribution Recognition An utterance may be out-of-scope because it was not included in the distribution the dialogue model was trained on. Distribution shifts may occur due to unknown user intents, different domains or incoherent speech. We differ from such methods since they either operate on images (Kim and Kim, 2018; Hendrycks et al., 2019; Mohseni et al., 2020) or assume access to an impractically large number of OOS examples in relation to INS examples (Tan et al., 2019; Kamath et al., 2020; Larson et al., 2019)

Dialogue Breakdown In comparison to out-of-distribution cases, dialogue breakdowns are unique to conversations because they depend on context (Higashinaka et al., 2016). In other words, the utterances fall within the distribution of reasonable responses but are out-of-scope due to the state of the particular dialogue. Such breakdowns occur when the conversation can no longer proceed smoothly due to an ambiguous statement from the user or some misunderstanding made by the agent (Ng et al., 2020b). GOLD also focuses on dialogue, but additionally operates under the setting of limited access to OOS data during training (Hendriksen et al., 2019).

2.2 Indirect Prediction

An alternative set of methods for OOS detection assume access to a supporting model trained solely on in-scope data. There are roughly three ways in which a core detector model can take advantage of the pre-trained supporting model.

Probability Threshold The first class of methods utilize the output probability of the supporting model to determine whether an input is out-of-scope. More specifically, if the supporting model's maximum output probability falls below some threshold τ, then it is deemed uncertain and the core detector model labels the input as OOS (Hendrycks and Gimpel, 2017). The confidence score of the supporting model can also be manipulated in a number of ways to help further separate the INS and OOS examples (Liang et al., 2018; Lee et al., 2018). Other variations include setting thresholds on reconstruction loss (Ryu et al., 2017) or on likelihood ratios (Ren et al., 2019).

Outlier Distance Another class of methods define out-of-scope examples as outliers whose distance is far away from known in-scope examples (Gu et al., 2019; Mandelbaum and Weinshall, 2017). Variants can tweak the embedding function or distance function used for determining the degree of separation. (Cavalin et al., 2020; Oh et al., 2018; Yilmaz and Toraman, 2020). For example, Local Outlier Factor (LOF) defines an outlier as a point whose density is lower than that of its nearest neighbors (Breunig et al., 2000; Lin and Xu, 2019).

Bayesian Ensembles The final class of methods utilize the variance of supporting models to make decisions. When the variance of the predictions is high, then the input is supposedly difficult to recognize and thus out-of-distribution. Such ensembles can be formed explicitly through a collection of models (Vyas et al., 2018; Shu et al., 2017; Lakshminarayanan et al., 2017) or implicitly through multiple applications of dropout (Gal and Ghahramani, 2016).

2.3 Data Augmentation

Our method also pertains to the use of data augmentation to improve model performance under low resource settings.

Augmentation in NLP Data augmentation for NLP has been studied extensively in the past (Jia and Liang, 2016; Silfverberg et al., 2017; Fürstenau and Lapata, 2009). Common methods include those that alter the surface form text (Wei and Zou, 2019) or perturb a latent embedding space (Wang and Yang, 2015; Fadaee et al., 2017; Liu et al., 2020), as well as those that perform paraphrasing (Zhang et al., 2019). Alternatively, masked language models generate new examples by proposing context-aware replacements for the masked token (Kobayashi, 2018; Wu et al., 2019).

Data Augmentation for Dialogue Methods for augmenting data to train dialogue systems are most closely related to our work. Previous research has used data augmentation to improve natural language understanding (NLU) and intent detection in dialogue (Niu and Bansal, 2019; Hou et al., 2018). Other methods augment the in-scope sample representations to support out-of-scope robustness (Ryu et al., 2018; Ng et al., 2020a; Lee and Shalyminov, 2019). Recently, generative adversarial networks (GANs) have been used to create out-of-domain examples that mimic known in-scope examples (Zheng et al., 2020; Marek et al., 2021). In contrast, we operate directly on OOS samples and consciously generate data far away from anything seen during pre-training, a decision which our later analysis reveals to be quite important.

3 Background and Baselines

In this section we formally describe the task of out-of-scope detection and the different approaches to handling this issue.

3.1 Problem Formulation

Let $\mathcal{D}_{direct} = \{(x_1, y_1), ..., (x_n, y_n)\}$ be a target dataset containing a mixture of in-scope and out-of-scope dialogues. The input context $x_i = \{(S_1, U_1), ..., (S_t, U_t)\}$ is a series of system and user utterances within t turns of a conversation. The desired output $y_i \in [0, 1]$ is a binary label representing whether that context is out-of-scope. We define OOS to encompass both out-of-distribution utterances, such as out-of-domain intents or gibberish speech, as well as in-distribution utterances spoken in an ambiguous manner. A model given access to such a dataset is an OOS detector $P_\theta(y_i|x_i)$ performing *direct* prediction.

In contrast, the problem we tackle in this paper is *indirect* prediction, where only a limited or nonexistent number of OOS examples are available during training. Instead, the training data is sampled from in-scope dialogues $\mathcal{D}_{indirect} \sim \mathcal{P}_{INS}$, and the labels $y_j \in \mathcal{Y}$ represent a set of known user intents. This data may be used to train an intent classifier which then acts as a supporting model to the core OOS detector during inference. Critically, the supporting model $P_\psi(y_j|x_i)$ has never encountered out-of-scope utterances during training.

3.2 Baselines

Prior methods for approaching indirect prediction generally fall into three categories: probability threshold, outlier distance and Bayesian ensemble. In all cases, the supporting model trained on the intent classification task uses a pretrained BERT model as its base (Devlin et al., 2019).

Starting with *Probability Threshold* baselines, (1) **MaxProb** declares an example as OOS if the maximum value of the supporting model's output probability distribution falls below some threshold τ (Hendrycks and Gimpel, 2017). (2) **ODIN** enhances this by adding temperature scaling and small perturbations to the input which help to increase the gap between INS and OOS instances (Liang et al., 2018). (3) **Entropy** considers an example to be OOS if the supporting model is uncertain, as determined by the entropy level rising above a threshold τ (Lewis and Gale, 1994).

Outlier Distance baselines find OOS examples by casting the problem as detecting outliers. Inputs are considered outliers when their embeddings are too far away from clusters of INS embeddings as measured by some threshold τ. The (4) **BERT** baseline embeds utterances uses the supporting

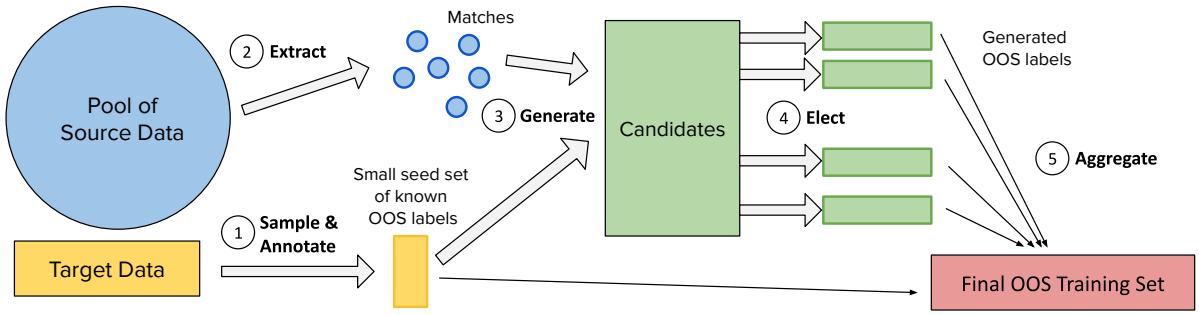

Figure 2: Full GOLD pipeline: (1) Sample and annotate a small seed set from unlabeled target data. (2) Extract similar matches from the source dataset. (3) Generate candidates by swapping utterances in the seed data with match sentences. (4) Elect the top candidates to become pseudo-labeled OOS examples. (5) Aggregate all elected labels to form the final OOS training set.

model pre-trained on intent classification and measures separation by Euclidean distance. Based on the success in (Podolskiy et al., 2021), the (5) **Mahalanobis** method embeds examples with a vanilla RoBERTa model and uses the Mahalanobis distance (Liu et al., 2019). Finally, inspired by BADGE for active learning (Ash et al., 2020), the (6) **Gradient** method sets the embedding of each example as the gradient vector of the input tokens as computed by back-propagation.

Bayesian Ensembles predict labels by the amount of variation formed by the estimates of an ensemble. More specifically, (7) **Dropout** implicitly creates a new model whenever it randomly drops a percentage of its nodes (Gal and Ghahramani, 2016). During inference, each input is passed through the supporting model k times to estimate the user intent. If the ensemble fails to reach a majority vote on the intent classification task, then the example is assigned as out-of-scope.

4 GOLD

To avoid a mismatch between training and inference, we are motivated to explore the direct prediction paradigm in a way the does not violate the OOS data restriction inherent to indirect prediction methods. Concretely, GOLD performs data augmentation on a small sample of labeled OOS examples to generate pseudo-OOS data. This weakly-labeled data is then combined with INS data for training a core OOS detector. We limit the number of OOS samples to be only 1% of the size of in-scope training examples. Note that indirect methods also typically have access to a modest number of OOS samples for tuning hyper-parameters, such as thresholds, so this adjustment is not an exclusive advantage of our method.

In addition to a small seed set of OOS examples, we assume access to an external pool of utterances, which serve as the source of data augmentations, similar to Hendrycks et al. (2019). We refer to this auxiliary data as the *source* dataset $\mathcal{S}$, as opposed to the *target* dataset $\mathcal{T}$ used for evaluating our method. GOLD now proceeds in three basic steps. (See Algorithm 1 for full details.)

4.1 Match Extraction

Our first step is to find utterances in the source data that closely match the examples in the OOS seed data. We encode all source and seed data into a shared embedding space to allow for comparison. When the seed example is a multi-turn dialogue, we embed only the final user utterance. Then for each seed utterance, we extract d similar utterances from source $\mathcal{S}$ as measured by cosine distance,[2] where d is the desired number of matches. For example, as seen in Figure 1, the seed text "*Do you know if it will rain on Friday?*" extracts "*Will it rain that day?*" as a match. We discuss different types of embedding mechanisms in section 5.3.

4.2 Candidate Generation

Since dialogue contexts often contain multiple utterances, we want our augmented examples to also span multiple turns. Accordingly, our next step involves generating candidates by carefully crafting new conversations using the existing dialogue contexts in the seed data. Each new candidate is formed by swapping a random user utterance in the seed data with a match utterance from the source data. Notably, agent utterances in the seed data are left untouched during this process.

[2]We also considered Euclidean distance and found that to yield negligible difference in preliminary testing.

4.3 Target Election

Candidates are merely pseudo-labeled as OOS, so relying on such data as a training signal might be quite noisy. Accordingly, we apply a filtering mechanism to ensure that only the candidates most likely to be out-of-scope are "elected" to become target OOS data. Elections are held by running all the candidates through an ensemble of baseline detectors. Specifically, we choose the top detectors from each of the major indirect prediction categories which results in three voters. If the majority of voters agree that an example is out-of-scope, then we include that candidate in our target pool.

As a last step, we aggregate the pseudo-labeled OOS examples, the small seed set of known OOS examples and the original INS examples to form the final training set for our model. We train a classifier with this data to directly predict out-of-scope instances.

Algorithm 1 GOLD

Require: ensemble of baseline detectors e
external source dataset $\mathcal{S}$

Input: Labeled, in-scope data from target data $\mathcal{T} = \{(x_1, y_1) \dots (x_n, y_n)\}$
Unlabeled data from target distribution $\mathcal{T}' = \{(x_1) \dots (x_m)\}$
Desired number of matches d

function SWAPAUGMENT($\mathcal{T}, \mathcal{T}', d$)
  seed set $\mathcal{A} \leftarrow$ sample and annotate $\mathcal{T}'$
  $\mathcal{S}' \leftarrow$ embed all items in $\mathcal{S}$
  for instance $i \in \mathcal{A}$ **do**:
    initialize $A_i = \{\}$
    while size(A_i) $< d$: **do**
      $i' \leftarrow$ embed instance i
      extract m nearest neighbors of i' from $\mathcal{S}'$ by cosine distance
      for $j \in m$ matches **do**:
        candidate $c \leftarrow$ generate(j, i)
        votes $\leftarrow$ ensemble e holds an election on candidate c
        if majority(votes) **then**:
          $A_i \leftarrow A_i \cup c$
        end if
      end for
    end while
  end for
  A' = aggregate(A_i)
  augmented dataset $\mathcal{D} \leftarrow \mathcal{T} \cup \mathcal{A} \cup \mathcal{A}'$
  return $\mathcal{D}$
end function

Split	*STAR*	*FLOW*	*ROSTD*
Train	22,051/1,248	60,119/4,499	30,521/3,200
Dev	2,751/178	3,239/228	4,181/453
Test	2,708/168	3,227/239	8,621/937

Table 1: Data count for each target dataset, broken down by number of in-scope/out-of-scope examples.

5 Experimental Setup

5.1 Target Datasets

We test our detection method on three dialogue datasets. Example counts shown in Table 1.

***S*chema-guided Dialog Dataset for *T*ransfer Le*ar*ning** STAR is a task-oriented dataset containing 6,651 multi-domain dialogues with turn-level intents (Mosig et al., 2020). Following the suggestion in Section 6.3 of their paper, we adapt the data for out-of-domain detection by selecting responses labeled as "ambiguous" or "out-of-scope" to serve as OOS examples. After filtering out generic utterances (such as greetings), we are left with 29,104 examples consisting of 152 user intents. Since the corpus does not strictly define a train and test set, we perform a random 80/10/10 split of the dialogues and other minor pre-processing to prepare the data for training.

SM Calendar *Flow* FLOW is also a task-oriented dataset with turn-level annotations (Andreas et al., 2020). Originally built for semantic parsing, FLOW is structured as a novel dataflow object that takes the form of a computational graph. For our purposes, we take advantage of the 'Fence' related labels found in the dataset, which represent situations where a user is straying too far away from discussions within the scope of the system, and thus need to be "fenced-in". We focus on utterances associated with a clear intent, once again dropping turns representing greetings and other pleasantries, which results in 71,551 examples spanning 44 total intents. The test set is hidden behind a leaderboard, so we divide the development set in half, resulting in an approximate 90/5/5 split for train, dev and test, respectively.

***R*eal *O*ut-of-Domain *S*entences From *T*ask-oriented *D*ialog** ROSTD is a dataset explicitly designed for out-of-distribution recognition (Gangal et al., 2020). The authors constructed sentences to be OOS examples with respect to a separate dataset collected by Schuster et al. (2019). The dialogues found in the original dataset then represent

the INS examples. ROSTD contains 47,913 total utterances spanning 13 intent classes and comes with a pre-defined 70/10/20 split which we leave unaltered. The dataset is less conversational since each example consists of a single turn command, while its labels are higher precision since each OOS instance is human-curated.

5.2 Evaluation Metrics

Following prior work on out-of-distribution detection (Hendrycks and Gimpel, 2017; Ren et al., 2019), we evaluate our method on three primary metrics. (1) Area under the receiver operating characteristic curve (AUROC) measures the probability that a random OOS example will have a higher probability of being out-of-scope than a randomly selected INS example (Davis and Goadrich, 2006). This metric averages across all thresholds and is therefore threshold independent. (2) The area under the precision-recall curve (AUPR) is another holistic metric which summarizes performance across multiple thresholds. The AUPR is most useful in scenarios containing class imbalance (Manning and Schütze, 2001), which is precisely our case since INS examples greatly outnumber OOS examples. (3) The false positive rate at recall of N (FPR@N) is the probability that an INS example raises a false alarm when $N\%$ of OOS examples are detected (Hendrycks et al., 2019). Thus, unlike the first two metrics, a lower FPR@N is better. We report FPR at values of N=$\{0.90, 0.95\}$.

5.3 Experiments on Model Variants

In addition to testing against baseline methods, we also run experiments to study the impact of varying the auxiliary dataset and the extraction options.

5.3.1 Source Datasets

We consider a range of datasets as sources of augmentation, starting with known out-of-scope queries (OSQ) from the Clinc150 dataset (Larson et al., 2019). Because our work falls under the dialogue setting, we also consider Taskmaster-2 (TM) as a source of task-oriented utterances (Byrne et al., 2019) and PersonaChat (PC) for examples of informal chit-chat (Zhang et al., 2018). Upon examining the validation data, we note that many examples of OOS are driven by users attempting to ask questions that the agent is not able to handle. Thus, we also include a dataset composed of questions extracted from Quora (QQP) (Iyer et al., 2017). Finally, we consider mixing all four datasets together into a single collection (MIX).

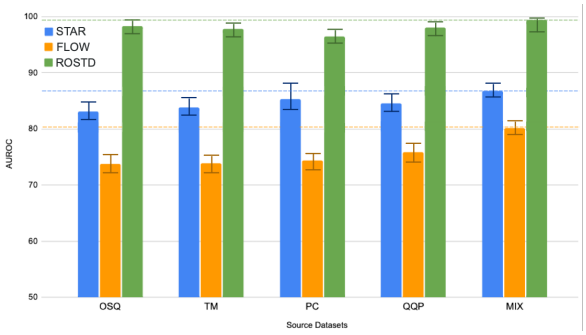

Figure 3: AUROC performance across source datasets

5.3.2 Extraction Techniques

To optimize the procedure of extracting matches from the source data, we try four different mechanisms for embedding utterances. (1) We feed each OOS instance into a SentenceRoBERTa model pretrained for **paraphrase** retrieval to find similar utterances within the source data (Reimers and Gurevych, 2019). (2) As a second option, we encode source data using a static BERT **Transformer** model (Devlin et al., 2019). Then for each OOS example encoded in the same manner, we extract the nearest source utterances. (3) We embed OOS and source data as a bag-of-words where each token is a 300-dim **GloVe** embedding (Pennington et al., 2014). (4) As a final variation, we embed all utterances with **TF-IDF** embeddings of 7000 dimensions. The spectrum of extraction techniques aim to progress from methods that capture strong semantic connections to the OOS seed data towards options with weaker relation to original seed data.

6 Key Results

We now present the results of our main experiments. As evidenced by Figure 3, MIX performed as the best data source across all datasets, so we use it to report our main metrics within Table 2. Also, given the strong performance of GloVe extraction technique across all datasets, we select this version for comparison purposes in the following analyses.

6.1 STAR Results

Left columns of Table 2 present STAR results. Models trained with augmented data from GOLD consistently outperform all other baselines across all metrics. The top model exhibits gains of 8.5% in AUROC and 40.0% in AUPR over the

nearest baseline. Performance is even more impressive in lowering the false positive rate with improvements of 24.2% and 29.8% at recalls of 0.95 and 0.90, respectively. Among the different baselines, we observe the Outlier Distance methods generally outperforming the others, with the Mahalanobis method doing the best. Among GOLD variations, there are mixed results as GloVe and TF-IDF both produce high overall accuracy. Notably, the Paraphrase method meant to extract matches most similar to the seed data performed the worst.

6.2 FLOW Results

Central columns of Table 2 present results on FLOW data. Once again, GOLD models outperform all baselines across all metrics. This time around, there is not an obvious winner among baselines. On the other hand, GloVe stands out as the clear overall top performer, with Transformer following closely behind. Models trained on data augmented by GloVe show improvements of 11.1% in AUROC, 71.9% in AUPR and 19.5% for FPR@0.95 over the nearest baseline. We again notice that the Paraphrase variation does not perform quite as well among GOLD methods.

6.3 ROSTD Results

As seen in Tables 2 and 3, GOLD outperforms not only all baselines, but also prior work on ROSTD across all metrics. The GloVe method cements its standing at the top with gains of 1.7% in AUROC, 13.8% in AUPR and 97.9% in FPR@0.95 against the top baselines. Given the consistently poor performance of Paraphrase yet again, we conclude that unlike traditional INS data augmentation, augmenting OOS data should *not* aim to find the most similar examples to seed data. We hypothesize that producing pseudo-labeled OOS data that are too similar to given known-OOS data causes the model to overfit since it is simply optimizing towards the same examples over and over again.

7 Discussion and Analysis

In this section, we conduct follow-up experiments to analyze the impact of our method's components and identify best practices when applying data augmentation for OOS detection.

7.1 Ablations

How much does augmentation help? Given the extra labels from the seed set, it is natural to ask whether the augmented data add any value. Furthermore, if the augmented data are useful, then we might want to know what an ideal number of additional datapoints would be. Figure 4 displays the AUROC of a model trained on varying the number of augmented datapoints, where "0" represents including only known OOS examples. We see a trend that accuracy improves for all target datasets as we add more pseudo-labeled examples, showing that augmentation helps. Improvement reaches a max around 24 matches per seed example, which suggests that the benefit of adding more datapoints has a limit. Accordingly, we use 24 matches for all results listed in Table 2.

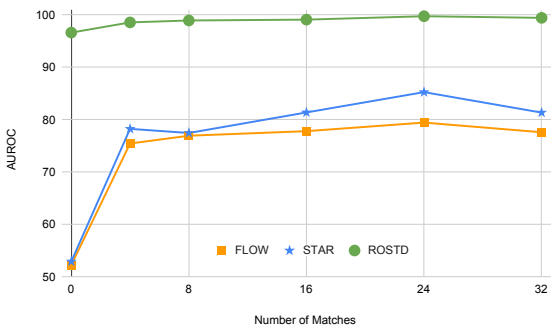

Figure 4: AUROC across datasets as the number of matches increases. A setting of $d = 8$ means for each seed example, 8 augmented examples are generated.

Does the extraction technique matter? Previous sections have established that extracting matches based on maximizing similarity to known OOS examples might not be ideal. We now ask what would happen if we went to the extreme by extracting matches that have no discernible relation to known OOS examples. The final row of Table 2 reveals the result of using random selection as an extraction technique. While Random is not always the worst, its poor performance across all metrics strongly suggests that augmented data should have at least some connection to the original seed set.

Is filtering even necessary? Since the source data distribution is obviously distinct from the target data distribution, perhaps it is possible to bypass elections and simply accept all candidates as OOS, similar to Outlier Exposure from Hendrycks et al. (2019). As seen in row 4 of Table 4, we observe that skipping elections leads to a drop in the AUROC of all the models on all datasets. The effect is most pronounced for STAR, where some of the QQP dialogues overlap with in-scope STAR domains.

Methods	STAR Data				FLOW Data				ROSTD Data			
	AUROC	*AUPR*	*FPR@.95*	*FPR@.90*	*AUROC*	*AUPR*	*FPR@.95*	*FPR@.90*	*AUROC*	*AUPR*	*FPR@.95*	*FPR@.90*
Oracle	0.9869	0.8837	11.2%	2.6%	0.8931	0.6471	62.5%	39.8%	0.9999	0.9992	0.03%	0.02%
MaxProb	0.6891	0.1824	85.8%	72.9%	0.6881	0.1581	75.4%	67.8%	0.7969	0.4554	54.3%	54.2%
ODIN	0.7012	0.1860	87.3%	75.3%	0.6893	0.1714	76.9%	69.5%	0.8087	0.4938	54.3%	54.2%
Entropy	0.7206	0.1915	87.0%	75.7%	0.6875	0.1887	77.4%	70.2%	0.8125	0.5250	54.3%	54.2%
BERT	0.7170	0.1958	85.3%	74.1%	0.5570	0.1685	98.6%	96.0%	0.9754	0.8126	8.80%	4.45%
Mahalanobis	0.8002	0.3179	73.9%	58.2%	0.7004	0.1757	82.7%	72.1%	0.9583	0.7338	19.1%	11.7%
Gradient	0.7255	0.1402	72.0%	62.7%	0.7223	0.1850	81.4%	70.0%	0.9821	0.8729	7.97%	3.87%
Dropout	0.5332	0.0631	99.9%	99.8%	0.5091	0.0707	99.9%	99.8%	0.5036	0.1991	99.9%	99.8%
Paraphrase	0.8537	0.4133	62.6%	47.5%	0.7767	0.2743	72.8%	60.4%	0.9967	0.9897	0.26%	0.16%
Transformer	0.8542	0.4251	65.9%	45.7%	**0.8059**	0.3228	62.7%	51.3%	0.9981	0.9904	0.21%	**0.09%**
GloVe	**0.8683**	0.4450	**56.0%**	40.9%	0.8022	**0.3243**	**60.6%**	**49.5%**	**0.9990**	**0.9933**	**0.17%**	**0.09%**
TF-IDF	0.8614	**0.4539**	68.1%	**40.3%**	0.7790	0.2758	74.2%	58.0%	0.9987	0.9905	0.56%	0.19%
Random	0.8531	0.4378	68.8%	45.4%	0.7692	0.2889	73.1%	62.0%	0.9984	0.9893	0.40%	0.19%

Table 2: Experimental results across all target datasets where bold items indicate best results, and underlined items indicate the runner-up. First seven rows are baselines, while the bottom five rows are models trained with GOLD.

Methods	AUROC↑	AUPR↑	FPR@.95↓
Likelihood (Gangal et al., 2020)	0.9822	0.9647	7.41%
OodGAN (Marek et al., 2021)	0.9899	0.9626	2.59%
GOLD w/ GloVe extraction	**0.9990**	**0.9933**	**0.17%**

Table 3: ROSTD results against previous works

Methods	STAR	FLOW	ROSTD
GloVe w/ QQP	0.3885	0.3173	0.9884
w/ US	0.2186	0.1514	0.9633
w/ MQA	0.2722	0.2283	0.9873
(4) No Election	0.2487	0.2548	0.9312
(5) Tiny Seed Set	0.1983	0.2111	0.8856
(6) Swap Last	0.3678	0.2980	0.9656

Table 4: Additional AUROC results with data augmented from QQP source and GloVe extraction

7.2 Applicability

How well would a direct classifier perform? Indirect prediction is often necessary in real-life because while in-scope data may be trivial to obtain, out-of-scope data is typically lacking. Accordingly, we artificially limited the amount of data available to mimic this setting. If such a limitation were to be lifted such that a sufficient amount of known OOS data were available, we could train a model to directly classify such examples. The first row in Table 2 shows the results of using all the available OOS data to perform direct prediction and represents an upper-bound on accuracy. This also shows there is still substantial room for improvement.

When does GOLD help the most? GOLD depends on a small seed set to perform data augmentation, so if this data is unavailable or extremely sparse, then the method will likely suffer. To test this limit, we train a model with half the size of the seed data and double the number of matches ($d = 24 \rightarrow 48$) to counterbalance the effect. Despite having an equal amount of pseudo-labeled OOS examples, the model with a tiny seed set (row 5 in Table 4) severely underperforms the original model (row 1). Separately, we note that dialogue breakdowns are more likely in conversations that contain multiple turns of context, like in STAR, as opposed dialogues consisting of single lines, as in ROSTD. Given the more prominent gains by our method in STAR, we conclude that GOLD achieves its gains partially from being able to recognize dialogue breakdowns.

What attributes make a source dataset useful? In studying Figure 3, we find that the most consistent single source dataset is QQP, which we use as the default for Table 4. Reading through some examples in QQP, the pattern we found was that many of the samples contained reasonable, but unanswerable questions that were beyond the skillset of the agent. One method for curating a useful source dataset then is to look for a corpus containing questions your dialogue model likely cannot answer. Furthermore, PersonaChat (PC) performed particularly well with STAR, a task-oriented dataset. We believe that since goal-oriented chatbots aim to solve specific tasks rather than engage in chit-chat, open-domain chat datasets serve as a good source of OOS examples.

The themes above suggest that good source datasets are simply those sufficiently different from the target data. We wondered if there was such as a thing as going to 'far', and conversely if there was any harm in being quite 'close'. Concretely, we expected a dataset containing medical questions would represent a substantially different dialogues compared to our target data (Ben Abacha

and Demner-Fushman, 2019). Table 4 presents results when training with source data from a medical question-answering dataset (MQA) or from unlabeled samples (US) from the same target dataset. The results show a significant drop in performance, indicating that augmentations far away from the decision boundary might not add much value. Rather, pseudo-labels near the border of INS and OOS instances are the most helpful. (Further analysis in Appendix C)

How does one create good OOS examples? As a final experiment, we replace only the last utterance with a match when generating candidates, rather than swapping any user utterance. We speculate this creates less diverse pseudo-examples, and therefore decreases the coverage of the OOS space. Indeed, row 6 in Table 4 reveals that worse candidates are generated when only the final utterance is allowed to be replaced. In conjunction with the insight from Section 6.3 that generated examples should be sufficiently different from given OOS examples, we believe that the key to producing good pseudo-OOS examples is to maximize the diversity of fake examples. OOS detection is less about finding out-of-scope cases, but rather an exercise in determining when something is not in-scope. This subtle distinction implies that the appropriate inductive biases should aim to move away from INS distribution, rather than close to OOS distribution.

8 Conclusion

This paper presents GOLD, a method for improving OOS detection when limited training examples are available by leveraging data augmentation. Rather than relying on a separate model to support the detection task, our proposed method directly trains a model to detect out-of-scope instances. Compared to other data augmentation methods, GOLD takes advantage of auxiliary data to expand the coverage of out-of-scope distribution examples rather than trying to extrapolate from in-scope examples. Moreover, our analysis reveals key techniques for further diversifying the training data to support robustness and prevent overfitting.

We demonstrate the effectiveness of our technique across three dialogue datasets, where our top models outperform all baselines by a large margin. Future work could explore detecting more granular levels of errors, as well as more sophisticated methods of filtering candidates (Welleck et al., 2020).

Acknowledgments

The authors are grateful to Tao Lei, Yi Yang, Jason Wu and Samuel R. Bowman for reviewing earlier versions of the manuscript. We would also like to thank David Gros and other members of the NLP Dialogue Group at Columbia University for their continued feedback and support.

References

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 93–104. ACM.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1:toward a realistic and diverse dialog dataset. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong. Association for Computational Linguistics.

Paulo R. Cavalin, Victor Henrique Alves Ribeiro, Ana Paula Appel, and Claudio S. Pinhanez. 2020. Improving out-of-scope detection in intent classification by using embeddings of the word graph space

of the classes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3952–3961. Association for Computational Linguistics.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and ROC curves. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 233–240. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 567–573. Association for Computational Linguistics.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 506–514. The Association for Computational Linguistics.

Giorgio Fumera, Ignazio Pillai, and Fabio Roli. 2003. Classification with reject option in text categorisation systems. In *12th International Conference on Image Analysis and Processing (ICIAP 2003), 17-19 September 2003, Mantova, Italy*, pages 582–587. IEEE Computer Society.

Hagen Fürstenau and Mirella Lapata. 2009. Semi-supervised semantic role labeling. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 220–228. The Association for Computer Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. 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 7764–7771. AAAI Press.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing fair generalization tasks for natural language inference. 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 4484–4494. Association for Computational Linguistics.

Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. 2019. Statistical analysis of nearest neighbor methods for anomaly detection. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10921–10931.

Mariya Hendriksen, Artuur Leeuwenberg, and Marie-Francine Moens. 2019. LSTM for dialogue breakdown detection: Exploration of different model types and word embeddings. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 443–453. Springer.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

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).

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of*

the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 1234–1245. Association for Computational Linguistics.

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs. *Kaggle Competition*.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5684–5696. Association for Computational Linguistics.

Joo-Kyung Kim and Young-Bum Kim. 2018. Joint learning of domain classification and out-of-domain detection with dynamic class weighting for satisficing false acceptance rates. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 556–560. ISCA.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 452–457. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. 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 6402–6413.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Sungjin Lee and Igor Shalyminov. 2019. Contextual out-of-domain utterance handling with counterfeit data augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7205–7209. IEEE.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 3–12. ACM/Springer.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. 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 5491–5496. Association for Computational Linguistics.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9031–9041. 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. *arXiv preprint arXiv:1907.11692*.

Amit Mandelbaum and Daphna Weinshall. 2017. Distance-based confidence score for neural network classifiers. *CoRR*, abs/1709.09844.

Christopher D. Manning and Hinrich Schütze. 2001. *Foundations of statistical natural language processing*. MIT Press.

P. Marek, V. Naik, Vincent Auvray, and Anuj Goyal. 2021. Oodgan: Generative adversarial network for out-of-domain data generation. *ArXiv*, abs/2104.02484.

Bilyana Martinovsky and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *International Speech Communication Association ISCA Workshop on Error Handling in Spoken Dialogue Systems - Château d'Oex, Vaud, Switzerland, 28-31 August 2003*, pages 11–16.

Sina Mohseni, Mandar Pitale, J. B. S. Yadawa, and Zhangyang Wang. 2020. Self-supervised learning for generalizable out-of-distribution detection. 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 5216–5223. AAAI Press.

Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. 2020. STAR: A Schema-Guided Dialog Dataset for Transfer Learning. *arXiv e-prints*.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020a. 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 2020, Online, November 16-20, 2020*, pages 1268–1283. Association for Computational Linguistics.

Nathan Ng, Marzyeh Ghassemi, Narendran Thangarajan, Jiacheng Pan, and Qi Guo. 2020b. Improving dialogue breakdown detection with semi-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tong Niu and Mohit Bansal. 2019. Automatically learning data augmentation policies for dialogue tasks. 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 1317–1323. Association for Computational Linguistics.

Kyo-Joong Oh, Dongkun Lee, Chan Yong Park, Young-Seob Jeong, Sawook Hong, Sungtae Kwon, and Ho-Jin Choi. 2018. Out-of-domain detection method based on sentence distance for dialogue systems. In *2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018, Shanghai, China, January 15-17, 2018*, pages 673–676. IEEE Computer Society.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. *CoRR*, abs/2101.03778.

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.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14680–14691.

Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. 2017. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognit. Lett.*, 88:26–32.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 714–718. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3795–3805. Association for Computational Linguistics.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2911–2916. Association for Computational Linguistics.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection, Vancouver, BC, Canada, August 3-4, 2017*, pages 90–99. Association for Computational Linguistics.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. 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 3564–3570. Association for Computational Linguistics.

Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke. 2018. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, volume 11212 of *Lecture Notes in Computer Science*, pages 560–574. Springer.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2557–2563. The Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. 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 6381–6387. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT contextual augmentation. In *Computational Science - ICCS 2019 - 19th International Conference, Faro, Portugal, June 12-14, 2019, Proceedings, Part IV*, volume 11539 of *Lecture Notes in Computer Science*, pages 84–95. Springer.

Eyup Halit Yilmaz and Cagri Toraman. 2020. KLOOS: KL divergence-based out-of-scope intent detection in human-to-machine conversations. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2105–2108. ACM.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1298–1308. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:1198–1209.

A Additional Results

This section shows the AUPR results corresponding to the AUROC results presented in the main paper.

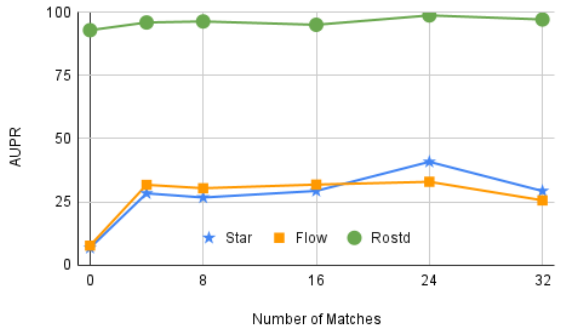

Figure 5: AUPR on with GloVe extraction method as we vary the number of matches. Compare with Figure 4 in the main paper.

We note that the trend is very similar, but just slightly harder to read since the range on the y-axis is larger. Overall, we reach the same conclusion that augmenting the examples certainly provides a benefit over simply training on the seed data alone.

B Latency Impact

Since GOLD is a data augmentation method, an OOS detector trained with this method incurs no additional cost during inference. In contrast, Probability Threshold methods will experience extra latency, albeit only minimally, from calculating whether an example falls below the threshold. Separately, the Outlier Distance methods must measure the distance to multiple clusters which takes a bit of time. Additionally, the Dropout method must pass the input through N models that form the Bayesian ensemble, leading to much slower inference.

With that said, our OOS detector only performs binary classification. So if it were to be deployed in a real-world task, such as intent classification, there would need to be an additional downstream model that separately classified the intents when the OOS detector labels a dialogue as in-scope. To mitigate this issue, a simple solution could be running the intent classifier alongside the OOS detector. Thus, rather than waiting for the result of the detector to start the prediction, the classifier would run in parallel and the classification results would be used only when the detector deemed it necessary.

Methods	STAR	FLOW	ROSTD
Random w/ MIX (default)	0.438	0.289	0.989
Random w/ MQA	0.245	0.176	0.979
Random w/ US	0.209	0.142	0.958

Table 5: AUPR results with varying source datasets and Random extraction technique

C Source Dataset vs. Technique

One might be curious to know whether choosing a source data or a technique is more important. Before answering this, we first note that source datasets (such as MIX) are not directly comparable to extraction techniques (such as GloVe) since they are different directions to improve performance. Source datasets impose the set of options to choose from, whereas extraction techniques determine how you select the options from that set. Both decisions can be combined together, and are not mutually exclusive.

With that said, there is some evidence that choosing the appropriate source dataset can make a more substantial impact. As initial evidence, notice that the Random extraction technique performs surprisingly well. This suggests that the gains come largely from using an advantageous source dataset that contains dialogue related examples near the INS and OOS border. Thus, Random extraction will naturally select some data points near the border as well, and do decently well. In contrast, Section 6.2 compares two new source datasets (MQA and US) that are not near the border, so Random selection of these points should cause the model to do poorly.

To verify this, we ran an additional experiment which extracted MQA samples using a Random approach rather than using GloVe as done originally. Table 5 reveals that indeed AUPR drops noticeably across all datasets. Similar decreases emerge when the experiment is run on the US dataset as well. Therefore, we conclude that selection of the source dataset can be fairly critical to success.