OUT-OF-DOMAIN INTENT DETECTION CONSIDERING MULTI-TURN DIALOGUE CONTEXTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-Domain (OOD) intent detection is vital for practical dialogue systems, and it usually requires considering long dialogue histories. However, previous OOD intent detection approaches are limited to single-turn contexts since it is non-trivial to gather or synthesize high-quality OOD samples in multi-turn settings, and the long distance obstacle exhibited in multi-turn contexts hinders us from obtaining robust features for intent detection. In this paper, we introduce a context-aware OOD intent detection (Caro) framework that aims to consider multi-turn contexts in OOD intent detection tasks. Specifically, we follow the information bottleneck principle to extract robust representations from multi-turn dialogue contexts by eliminating superfluous information that is not related to intent detection tasks. We also propose to synthesize pseudo OOD samples with the help of unlabeled data under the constraint of dialogue contexts, i.e., candidate OOD samples are retrieved from unlabeled data based on their context similarities and representations of these candidates are mixed-up to produce pseudo OOD samples. A three stage training process is introduced in Caro to combine above approaches. Empirical results validate the superiority of our method on benchmark datasets.

1 INTRODUCTION

Intent detection is vital for dialogue systems as it controls the pipelines of the entire system (Chen et al., 2017). It is important to explicitly model multi-turn dialogue contexts in the intent detection process since conversations usually last several turns to complete (Qin et al., 2021). Recently, promising results are reported for intent detection under the *closed-world assumption* (Shu et al., 2017), i.e., the training and testing distributions are assumed to be identical and all intents involved in testing are seen in the training process. However, this assumption may not be valid in practice (Dietterich, 2017), where a deployed system usually confronts an *open-world* (Fei & Liu, 2016; Scheirer et al., 2012), i.e., the testing distribution is subject to change and Out-of-Domain (OOD) intents that are not seen in the training process may emerge in testing. It is important to equip intent detection modules with OOD detection abilities so that they can accurately classify seen In-Domain (IND) intents while rejecting unseen OOD intents (Yan et al., 2020a; Shen et al., 2021).

Numerous methods have been proposed to develop OOD intent detection models (Yang et al., 2021), and among which the most straightforward and effective one is to build a $(k+1)$-way classifier (k is the number of IND intents) (Larson et al., 2019). The $(k+1)$-th intent is regarded as a special OOD intent (Fei & Liu, 2016). Various studies are carried out to improve this approach and state-of-the-art results are reported on a wide range of benchmarks (Shu et al., 2021; Zhan et al., 2021). Generally, the success of these attempts attributes to two key ingredients: 1. extracting robust features for intent detection (Vaze et al., 2021); and 2. gathering representative OOD samples in training (Zhan et al., 2021).

Most previous OOD intent detection approaches are limited to single-turn inputs (Yan et al., 2020a) without considering multi-turn dialogue contexts. This downgrades the OOD intent detection performance and prevents us from applying these approaches in real applications since multi-turn contexts are generally critical for practical intent detection tasks (Weld et al., 2021). However, multi-turn dialogues with OOD intent annotations are generally expensive to obtain, and it is non-trivial to synthesize high-quality OOD samples in multi-turn settings (Lee & Shalyminov, 2019). Further, the *long distance obstacle* (Qin et al., 2021) exhibited in multi-turn contexts also hinders us from

directly migrating previous OOD intent detection approaches to multi-turn settings because long dialogue histories may carry noises that are irrelevant for intent detection (Ohsugi et al., 2019).

To tackle above problems, we propose a novel context-aware OOD intent detection framework **Caro** to explore OOD intent detection in multi-turn settings. Specifically, Caro introduces two approaches to improve multi-turn OOD intent detection performance: **1.** We follow the information bottleneck principle (Tishby et al., 2000) to extract robust representations from multi-turn contexts by discarding superfluous information that is not related to intent detection, i.e., different views of inputs are generated and only information shared by these views are retained; **2.** We synthesize pseudo OOD samples that are coherent to the given contexts for the $(k+1)$-th intent with the help of unlabeled data, which can be collected almost *for free* from a deployed system in the open world (Katz-Samuels et al., 2022). Specifically, a pool of candidate samples are first gathered based on their context similarities and then a quality inspection scheme is implemented to filter out IND samples with the help of an existing OOD detector. High-quality pseudo OOD samples are synthesized using a mix-up scheme on these filtered candidates. Extensive empirical results validate the superiority of Caro in building OOD intent detectors considering multi-turn dialogue contexts.

Caro provides several advantages compared to existing methods: **1.** Caro is an effective learning framework for OOD intent detection that considers multi-turn dialogue contexts, whereas previous methods primarily focus on single-turn inputs. We bridge a critical research gap since leveraging multi-turn contexts for OOD intent detection is yet under explored; **2.** Caro captures robust representations for intent detection from multi-turn contexts, and we are the first to employ the information bottleneck principle in multi-turn OOD intent detection tasks to migrate the long distance obstacle; **3.** Caro synthesizes high-quality OOD samples under the constraint of given dialogue contexts with the help of "cheap" unlabeled data. This approach relieves the burden of collecting expensive annotations for OOD samples.

We summarize our main contributions as follows.

1. We propose a novel framework Caro to build OOD intent detection models considering multi-turn dialogue contexts. This setting is more practical in real applications and our framework can classify IND intents and detect OOD intents simultaneously.
2. We follow the information bottleneck principle to learn robust representation for intent detection and synthesize pseudo OOD samples under the constraint of multi-turn dialogue contexts.
3. We conduct extensive experiments on two benchmark datasets to empirically demonstrate the effectiveness of our proposed framework.

2 Problem Setup

The OOD intent detection task investigated in our study aims to reject OOD inputs while being able to detect intents of IND inputs. Concretely, given k IND intent classes $\mathcal{I} = \{I_i\}_{i=1}^{k}$, we denote all samples that do not belong to these k classes as the $(k+1)$-th intent I_{k+1}. Our training data contain a set of labeled IND samples $\mathcal{D}_I = \{(x_i, y_i)\}$ and a set of unlabeled samples $\mathcal{D}_U = \{\widetilde{x_i}\}$, where $y_i \in \mathcal{I}$ is the label of input sample x_i and the label of $\widetilde{x_i}$ belongs to $\mathcal{I} \cup I_{k+1}$. Our testing data contain a mixture of IND and OOD samples $\mathcal{D}_T = \{(\widetilde{x_i}, \widetilde{y_i})\}$, where $\widetilde{y_i} \in \mathcal{I} \cup \{I_{k+1}\}$. Moreover, each input sample x from $\mathcal{D}_I$, $\mathcal{D}_U$ and $\mathcal{D}_T$ consists of an utterance u and a dialogue history $h = u_1, u_2, \ldots, u_t$, $(t \geq 0)$: $x = \langle h, u \rangle$. In this study, we follow the most widely applied OOD intent detection approaches to build a $(k+1)$-way classifier as our OOD intent detector.

3 Method

Our framework Caro attempts to extend previous two key ingredients for building OOD detectors in multi-turn settings. Specifically, Caro mainly addresses the following two issues: (1) how to alleviate the long distance obstacle and learn robust representations from multi-turn dialogue histories; (2) how to synthesize high-quality pseudo OOD samples that fit the constraints of given dialogue contexts.

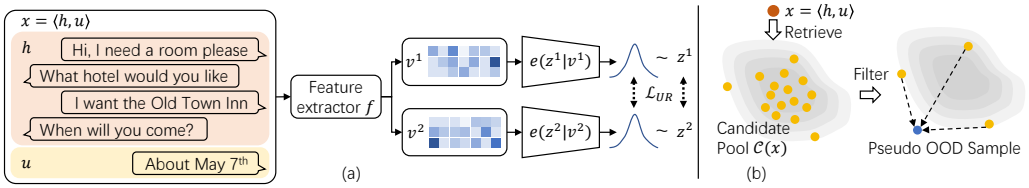


Figure 1: Two key ingredients introduced in Caro: (a) Learn robust representations from multi-turn dialogues based on two views of an input x using the information bottleneck principle; (b) Synthesize pseudo OOD samples using the feature mixup approach under the constrain of a given context.

3.1 Robust Representation Learning

The major challenge for extracting robust representations from multi-turn dialogue histories is the long distance obstacle, i.e., the dialogue history is too long so that it may yield superficial representations that are irrelevant to intent detection tasks, and OOD intent detectors build upon these representations are liable to break under distribution shift (Wang et al., 2018).

Our framework Caro tackles above issues with the help of the multi-view information bottleneck principle (Tishby et al., 2000), which is built based on the basic assumption of the multi-view literature: each view provides the same task-relevant information (Zhao et al., 2017). Robust representations are obtained by only capturing information shared between these views so that more modeling capacity is allocated to label-related information.

Specifically, for each input x, we produce two different views of x based on a label preserving data augmentation scheme and use a feature extractor f to extract two feature vectors v^1 and v^2 from these two views, respectively. To learn compact representation without superfluous information, we assume there is a latent variable z^j for the representation of each view $v^j (j = 1, 2)$. Formally, we assume an encoder $e(z^j|v^j)$ can predict the distribution of latent variable z^j, i.e., $z^j \sim e(z^j|v^j)$, $(j = 1, 2)$. To remove superfluous information from representations yielded by the encoder, we follow the information bottleneck principle (Tishby et al., 2000) by optimizing the following unsupervised representation learning loss (Federici et al., 2019) on unlabeled data $\mathcal{D}_U$:

$$\mathcal{L}_{UR} = \sum_{x \in \mathcal{D}_U} -I(z^1; z^2) + (D_{KL}[e(z^1|v^1)||e(z^2|v^2)] + D_{KL}[e(z^2|v^2)||e(z^1|v^1)])/2, \quad (1)$$

where z^1 and z^2 are random variables for these two views' representations, I calculates mutual information (MI) of two random variables, and D_{KL} calculates the KL divergence between two distributions. Eq. 1 ensures that the representation z^1 for v^1 is sufficient for v^2 and it also helps to increase the robustness of the representation by discarding irrelevant information (Federici et al., 2019).

To facilitate the computation of Eq. 1, we model the distributions of z^1 and z^2 as factorized Gaussian distributions $z^j \sim \mathcal{N}(\mu(v^j), \Sigma(v^j))$, $(j = 1, 2)$, in which $\mu(v^j)$ and $\Sigma(v^j)$ are two neural networks that produces the mean and deviation, respectively. These two views v^1 and v^2 of x are obtained through dropout operations on the feature extractor f. In this study, we implement f using the BERT model (Devlin et al., 2018). f takes the concatenation of all utterances in x as the input.

3.2 Pseudo OOD Synthesis

When implementing OOD intent detectors with $(k + 1)$-way classifiers, we need annotated OOD samples for the $(k+1)$-th intent. However, it is usually expensive to manually collect these annotated data. Previous studies propose to tackle this issue by synthesizing pseudo OOD samples, for which the most effective approach is the feature Mixup (Zhan et al., 2021), i.e., convex combinations of IND features are used as pseudo OOD samples.

However, despite the reported feasibility, previous pseudo OOD sample synthesizing approaches only model single-turn inputs. That means these synthesized pseudo OOD samples are unconstrained. In multi-turn settings, an ideal pseudo OOD sample x' should be *constrained* by a dialogue

history h, i.e., x' should be coherent to h while carrying OOD intents. Utterances that are not coherent to the current context h are highly unlikely to be seen in practice and can be easily determined. In fact, we can regard OOD samples that are constrained by h as "hard" OOD samples, which are reported to be more effective at improving the performance of OOD detectors (Zhan et al., 2021).

To tackle above issues, we propose to synthesize pseudo OOD samples by only mixing-up features under the constraint of similar contexts. Specifically, for each sample x in $\mathcal{D}_I$, we retrieve M samples from the unlabelled dataset $\mathcal{D}_U$ that share similar contexts with x, and construct a candidate pool $\mathcal{C}(x)$ that contains a mixture of IND and OOD samples. The feature v of the constructed pseudo OOD sample is obtained by mixing-up features of candidates in $\mathcal{C}(x)$ by their distance to x:

$$v = \sum_{x_i \in \mathcal{C}(x)} w_i f(x_i), \quad w_i = \frac{\exp ||f(x), f(x_i)||}{\sum\limits_{x_j \in \mathcal{C}(x)} \exp ||f(x), f(x_j)||}, \tag{2}$$

where $||\cdot, \cdot||$ calculates the L2 distance between two features. With these pseudo OOD samples, we can train a $(k+1)$-way classifier $p(y|f(x))$ as our OOD detector.

3.3 Multi-stage Training

For a synthesized pseudo OOD sample x', we usually expect x' does not carry any IND intents so that it can be more effectively applied to optimize the $(k+1)$-th intent. However, we observe that pseudo OOD samples obtained through Eq. 2 may violate this expectation because samples from the candidate pool $\mathcal{C}(x)$ may carry IND intents. Therefore it is a desiderata to filter these IND samples from $\mathcal{C}(x)$ when building high-performance OOD detectors. In this study, we introduce a training process that involves three stages to tackle this issue.

In the **first stage**, we aim to equip our detector with the basis OOD intent detection ability. Specifically, a set of pseudo OOD samples are first constructed following the approach proposed by Zhan et al. (2021) and a $(k+1)$-way classifier p is trained with labeled IND samples from $\mathcal{D}_I$ and these pseudo OOD samples. This classifier assigns probability masses for the OOD intent on testing samples.

In the **second stage**, we aim to enhance the robustness of extracted representations with the help of unlabeled data in $\mathcal{D}_U$. Specifically, we infer pseudo-labels for samples in D_U and collect samples that are labeled with intent I_{k+1} as a set of pseudo OOD samples $\mathcal{D}_{PL}$. A self-training scheme is used in this stage to further optimize the classifier p using the cross-entropy loss on $\mathcal{D}_I \cup \mathcal{D}_{PL}$ and the unsupervised representation learning loss (Eq. 1) on $\mathcal{D}_U$.

In the **third stage**, we use Eq. 2 to construct another set of pseudo OOD samples $\mathcal{D}_{DC}$ and add these samples in the optimization process of the cross-entropy loss on the basis of the second stage's training target. Specifically, before applying Eq. 2, the obtained classifier p is used to filter out candidates in $\mathcal{C}(x)$ that lie close to IND intents to ensure the quality of synthesized pseudo OOD samples, i.e., samples with high classification confidence scores are filtered out. The following loss is optimized in this stage:

$$\mathcal{L} = \sum_{(x,y) \in \mathcal{D}_I \cup \mathcal{D}_{PL}} CE[y, p(f(x))] + \sum_{v \in \mathcal{D}_{DC}} CE[I_{k+1}, p(v)] + \lambda \mathcal{L}_{UR}, \tag{3}$$

where λ is a scalar hyper-parameter to control the weight of the representation learning loss.

The whole training process of Caro is summarized in Algorithm 1. In the inference phase, we use the prediction of the classifier p to implement the OOD detector, i.e., $p(y|f(x))$.

Discussions. The two key components of Caro, i.e., robust representation learning (Section 3.1) and pseudo OOD synthesis (Section 3.2) work collaboratively in the training process. First, learning robust representation makes IND and OOD samples more separable and helps synthesize high-quality representations for pseudo OOD samples. Second, synthesized pseudo OOD samples boost the performance of the $(k+1)$-way classifier. Third, an effective and evolving classifier helps to more effectively filter out IND samples from the candidate pool.

Algorithm 1: Caro: OOD Detection Considering Multi-turn Contexts

Input: IND data $\mathcal{D}_I = \{(x_i, y_i)\}$, unlabelled data $\mathcal{D}_U = \{\widetilde{x_i}\}$, classifier p, randomly initialized model with parameter θ, weight for representation learning λ.
Output: model parameter $\theta^\star$.
while *train_stage1* **do**
  Construct a set of pseudo OOD samples $\mathcal{D}_{Mix}$ (Zhan et al., 2021).
  Optimize the classifier p using the cross-entropy loss on $\mathcal{D}_I \cup \mathcal{D}_{Mix}$.
end
while *train_stage2* **do**
  Calculate the unsupervised representation learning loss using Equation 1.
  Construct a set of OOD samples $\mathcal{D}_{PL}$ from $\mathcal{D}_U$ by pseudo-labels.
  Calculate the cross-entropy loss on $\mathcal{D}_I \cup \mathcal{D}_{PL}$.
  Update the parameters θ.
end
while *train_stage3* **do**
  Calculate the unsupervised representation learning loss using Equation 1.
  Construct a set of OOD samples $\mathcal{D}_{PL}$ from $\mathcal{D}_U$ by pseudo-labels.
  Synthesize OOD samples $\mathcal{D}_{DC}$ using Equation 2.
  Calculate the cross-entropy loss on $\mathcal{D}_I \cup \mathcal{D}_{PL} \cup \mathcal{D}_{DC}$.
  Update the parameters θ based on Equation 3.
end
while *eval* **do**
  Predict the intent class using the classifier p.
end

STAR	Intent	IND train	OOD train	IND valid	OOD valid	IND test	OOD test
Full	150	22,051	1,248	2,751	0	2,708	168
Small	150	11,000	621	2,751	0	2,708	168

Table 1: Dataset statistics.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets We perform experiments on two variants of a benchmark dataset STAR (Mosig et al., 2020), i.e., STAR-Full and STAR-Small. STAR is a task-oriented dialog dataset, consisting of 5,820 dialogues in 13 domains with turn-level intents. The dataset is designed to be strong history dependent and each dialogue contains 21.71 turns on average. We adopt the data as OOD samples by selecting turns labeled as "out_of_scope", "custom", or "ambiguous", suggested by the authors. Following Chen & Yu (2021), we construct STAR-Full by filtering out generic utterances (e.g., greetings). We also make STAR-Small by down-sampling (50%) the training set of STAR-Full to evaluate the performance in the low-resource scenario. Standard splits of above datasets are followed (see Table 1).

To simulate the unlabelled data, we mix 30% of IND data and all of the OOD data in the training set. The number of IND/OOD samples of unlabelled data is 6614/1248 in STAR-Full and 3302/621 in STAR-Small, respectively. Note that in the training and validation process, we do not use the label information of the unlabelled data and the labeled data only contains samples from the IND intents.

Metrics Following Zhang et al. (2021b); Zhan et al. (2021); Shu et al. (2021), we use macro F1-score (**F1-All**) calculated over all intents (IND and OOD intents) to evaluate the OOD detection performance. We also calculate macro F1-scores over IND intents (**F1-IND**) and OOD intent (**F1-OOD**) to evaluate fine-grained performances. We do not use the metric of accuracy, because the test set is unbalanced.

	Model	STAR-Full			STAR-Small		
		F1-All	F1-OOD	F1-IND	F1-All	F1-OOD	F1-IND
Oracle	K+1	50.1	64.46	50	46.54	58.23	46.46
$\mathcal{D}_I$	MSP w/o h	17.29	14.12	17.31	17.12	13.49	17.14
	MSP	40.83	19.74	40.97	37.17	18.1	37.31
	SEG w/o h	0.06	2.77	0.04	0.05	2.27	0.04
	SEG	17.45	6.85	17.53	11.66	7.39	11.69
	DOC w/o h	11.31	14.16	11.29	0.08	11.04	0
	DOC	26.53	16.80	26.60	3.47	11.78	3.41
	ADB w/o h	23.27	17.63	23.30	20.08	21.27	20.07
	ADB	44.64	20.56	44.80	41.36	18.23	41.51
	DA-ADB w/o h	17.87	15.15	17.88	16.34	17.03	16.33
	DA-ADB	37.27	22.87	37.37	34.81	20.43	34.91
	Outlier w/o h	23.35	16.75	23.39	19.56	15.42	19.59
	Outlier	43.84	19.53	44.01	39.51	19.92	39.64
	CDA	43.76	5.26	44.03	40.02	10.48	40.22
$\mathcal{D}_I$+$\mathcal{D}_U$	Pseudo-label	45.34	50.22	45.31	42.59	41.33	42.6
	Caro w/o $\mathcal{D}_{DC}$	46.14	52.58	46.10	43.24	44.66	43.23
	Caro (ours)	**48.18**± 0.9	**54.26**± 3.1	**48.13**± 0.9	**44.09**± 1.9	**47.51**± 2.7	**44.06**± 1.9

Table 2: **Main results.** Comparison between our method and baselines. All values are percentages. **Bold** numbers are superior results. We report standard deviations estimated across 3 runs.

Training Details Our feature extractor f is implemented using BERT Devlin et al. (2018) with a mean-pooling layer. The classification head in p is implemented as two-layer MLPs with the LeakyReLU activation Xu et al. (2020), while the projection heads in $\mu(v^j)$ and $\Sigma(v^j)$ are implemented as three-layer MLPs. The optimizer AdamW and Adam Kingma & Ba (2014) is used to finetune BERT and all the heads with a learning rate of 1e-5 and 1e-4, respectively. Jensen-Shannon mutual information estimator (Hjelm et al., 2018) is used to maximize the MI between two latent variables. We use $\lambda = 2$ in all experiments. All results reported in our paper are averages of 3 runs with different random seeds. Hyper-parameters are searched based on IND intent classification performances on validation sets. See Appendix A for more implementation details.

4.2 Main Results

We compare Caro with two types of OOD detection methods. The first type uses only labeled IND data in training, while the second type trains models on both IND and unlabelled data. We compare with competitive baselines of the first type: **MSP:** (Hendrycks & Gimpel, 2017) utilizes the maximum Softmax probability of a k-way classifier to detect OOD inputs; **SEG:** (Yan et al., 2020b) proposes a semantic-enhanced Gaussian mixture model; **DOC:** (Shu et al., 2017) employs k 1-vs-rest Sigmoid classifiers and use the maximum predictions to detect OOD intents; **ADB:** (Zhang et al., 2021b) learns an adaptive decision boundaries for OOD detection; **DA-ADB:** (Zhang et al.) learns distance-aware intent representations and adaptive decision boundaries for open intent detection; **Outlier:** (Zhan et al., 2021) mixes convex interpolated outliers and open-domain outliers to train a $(k+1)$-way classifier; **CDA:** (Lee & Shalyminov, 2019) performs detection by using counterfeit OOD turns. We also implement their variants which ignore dialogue contexts (**w/o** h). Note that CDA is designed for multi-turn contexts, hence we do not implement its single-turn variant. For fair comparisons, all baselines are implemented with codes released by their authors, and use BERT as the backbone. See Appendix B for more details about baselines.

We compare with variants of Caro, which belong to the second type of method: **Pseudo-label:** (Lee et al., 2013) learns a $(k+1)$-way classifier which constructs a set of OOD samples $\mathcal{D}_{PL}$ from $\mathcal{D}_U$ by pseudo-labels and optimizes the cross-entropy loss on $\mathcal{D}_I \cup \mathcal{D}_{PL}$; **Caro w/o** $\mathcal{D}_{DC}$**:** removes the synthesized OOD samples under the constraint of dialogue contexts, i.e., the model is only trained by two stages; We further report the performance of a $(k+1)$-way classifier (**K+1**) trained on $\mathcal{D}_I$ and $\mathcal{D}_U$, in which the label information of $\mathcal{D}_U$ is leveraged. Note that the performance of K+1 is generally regarded as the upper bound.

Model		STAR-Full			STAR-Small		
		F1-All	F1-OOD	F1-IND	F1-All	F1-OOD	F1-IND
IND+Unlabelled	InfoMax	44.71	50.5	44.67	42.13	41.07	42.14
	MV-InfoMax	44.77	51.43	44.73	42.95	43.20	42.95
	Contrastive Learning	44.68	51.94	44.63	42.72	42.78	42.72
	Caro w/o $\mathcal{D}_{DC}$	**46.14**	**52.58**	**46.10**	**43.24**	**44.66**	**43.23**

Table 3: Ablation study on the representation learning loss.

Model		STAR-Full			STAR-Small		
		F1-All	F1-OOD	F1-IND	F1-All	F1-OOD	F1-IND
IND+Unlabelled	FM	41.20	22.67	41.33	33.72	19.60	33.81
	LG	46.67	50.57	46.65	43.03	44.76	43.01
	OS	46.4	49.02	46.38	43.19	44.90	43.17
	Caro w/o $\mathcal{Q}$	42.39	27.17	42.50	34.38	19.17	34.48
	Caro (ours)	**48.18**	**54.26**	**48.13**	**44.09**	**47.51**	**44.06**

Table 4: Ablation study on OOD synthesis approaches.

Table 2 shows the OOD detection performance of all baselines and our method. We can observe that: **1.** Methods (e.g., MSP, SEG, DOC, ADB, DA-ADB, and Outlier) using dialogue contexts, in general, show strong OOD detection performance over the counterparts (i.e., w/o h). **2.** Our method Caro significantly outperforms all baselines in terms of IND and OOD effectiveness on all the two datasets, by making better use of the dialogue contexts on IND and unlabelled data; **3.** Our method outperforms Pseudo-label and Caro w/o $\mathcal{D}_{DC}$. This proves the importance of robust representation learning for dialogue contexts and shows the effectiveness of pseudo OOD samples synthesized by Caro.

4.3 Ablation Studies

This section provides comprehensive ablation studies to understand the effectiveness of Caro.

Ablation on the representation learning loss. We perform ablation on three alternatives for $\mathcal{L}_{UR}$: 1. **InfoMax** (Poole et al., 2019) maximizes mutual information between input and its latent variable $I(v; z)$; 2. **MV-InfoMax** (Bachman et al., 2019) maximizes mutual information between latent variables of an input's two views $I(z_1; z_2)$; 3. **Contrastive Learning** (Caron et al., 2020) formulates the contrastive loss for two views of an input. Here, we focus on studying the efficacy of representation learning, and do not use the synthesized OOD samples $\mathcal{D}_{DC}$ produced by Caro. Table 3 indicates that our method outperforms all ablation models. We can observe that: **1.** representation learned by other losses degenerates the model performance by a large margin. This shows the effectiveness of the robust representation learned by Caro. **2.** InfoMax achieves the lowest performance compared to other models. This further proves the importance of eliminating superfluous information.

Ablation on OOD synthesis approaches. We compare Caro with four pseudo OOD synthesis approaches: 1. *Feature Mixup* (**FM**): follows Zhan et al. (2021) to produce OOD features using convex combinations of IND features; 2. *Latent Generation* (**LG**): follows Zheng et al. (2020) to decode pseudo OOD samples from a latent space; 3. *Open-domain Sampling* (**OS**): follows Zhan et al. (2021) to use sentences from other corpora as OOD samples; 4. **Caro w/o** $\mathcal{Q}$: a variant of Caro which synthesize OOD samples without a quality inspection scheme to filter out IND samples. Table 4 shows that Caro outperforms all alternative approaches. We can further observe that: **1.** synthesized OOD samples without a quality inspection scheme would significantly hurt the performance, because the candidate samples in the pool might contain IND samples which would make the produced OOD samples overlap with IND samples in the feature space. **2.** FM achieves the lowest performance compared to other models. The produced OOD samples by mixing IND features might also overlap with IND samples.

Ablation on the weight of representation learning loss. Tabel 5 reports the OOD detection results as we vary the weight λ for $\mathcal{L}_{UR}$. The model is evaluated on STAR-small dataset. The results

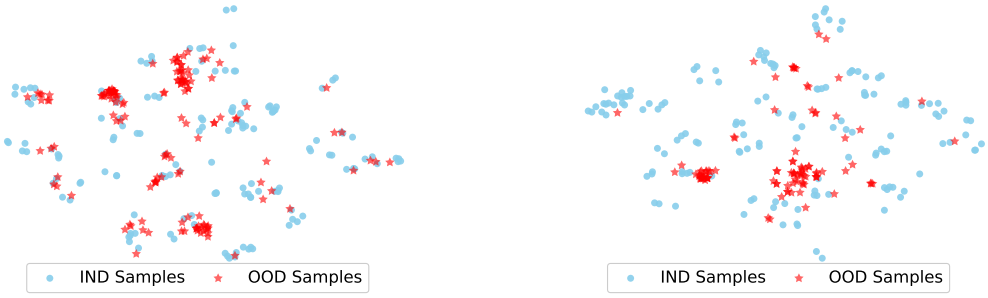
	1.8	1.9	2.0	2.1	2.2
F1-All	43.20	44.00	**44.09**	43.78	42.34
F1-OOD	49.79	**50.04**	47.51	47.25	49.98
F1-IND	43.16	43.96	**44.06**	43.76	42.29

Table 5: Ablation study on the weight λ for the representation learning loss on STAR-small dataset.

indicate that a relatively large weight is desirable. In most cases, Caro outperforms the baseline methods in Table 2.

4.4 Qualitative analysis

To further demonstrate the effectiveness of Caro, we visualized the features learned in the penultimate layer of OOD detectors that are trained by Pseudo-label and Caro, respectively. Results shown in Figure 2 demonstrate the robust representation and OOD samples produced by Caro help the OOD detector learn better representations. The learned feature space is smoother and representations for IND and OOD samples are more separable. This validates our claim that Caro helps to learn robust representation and produce high-quality OOD samples and improves the OOD detection performance.



(a) Pseudo-label (b) Caro

Figure 2: t-SNE visualization of learned features on the test set on STAR-Full.

5 Related Work

OOD Intent Detection: OOD detection problems have been widely investigated in conventional machine learning studies Geng et al. (2020). Recent neural-based methods try to improve the OOD detection performance by learning more robust representations on IND data Zhou et al. (2021; 2022); Yan et al. (2020b); Zeng et al. (2021). These representations can be used to develop density-based or distance-based OOD detectors Lee et al. (2018); Podolskiy et al. (2021); Liu et al. (2020); Tan et al. (2019). Some methods also propose to distinguish OOD inputs using thresholds based methods Gal & Ghahramani (2016); Lakshminarayanan et al. (2017); Ren et al. (2019); Gangal et al. (2020); Ryu et al. (2017), or utilizing unlabeled IND data Xu et al. (2021); Jin et al. (2022).

Multi-turn Dialogue Contexts: Modeling multi-turn contexts is the foundation for various dialogue related tasks, such as intent detection (Ghosal et al., 2021), question answering (Li et al., 2020), and dialogue summarization (Chen et al., 2021). Lee & Shalyminov (2019) propose to perform OOD Intent detection relying on the dialogue contexts. However, they do not explicitly tackle the long distance obstacle (Qin et al., 2021) exhibited in the multi-turn contexts. Chen & Yu (2021) generate pseudo OOD samples from an auxiliary dataset with a seed set of IND samples considering multi-turn contexts. However, they need to label IND samples, which are expensive to annotate in the context of multi-turns.

Pseudo OOD Sample Generation: Some works try to tackle OOD detection problems by generating pseudo OOD samples. Generally, three categories of approaches are proposed: **1.** Feature

Mixup Zhan et al. (2021): OOD features are directly produced by mixing up IND features Zhang et al. (2018); **2.** Latent Generation Marek et al. (2021): OOD samples are drawn from the low-density area of a latent space; **3.** Open-domain Sampling Hendrycks et al. (2018): data from other corpora are directly used as pseudo OOD samples. Existing approaches only attempt to generate OOD samples using single-turn inputs. Our method Caro is the first attempt to synthesize OOD samples considering multi-turn dialogue contexts.

6 CONCLUSION

In this paper, we propose Caro, a novel context-aware OOD intent detection framework to explore OOD intent detection in multi-turn settings. We follow the information bottleneck principle to learn robust representation for intent detection and synthesize pseudo OOD samples under the constraint of multi-turn dialogue contexts. A three-stage training process is introduced in Caro to construct a $(k+1)$-way classifier as the resulting OOD intent detector. We conduct extensive experiments on two benchmark datasets to empirically demonstrate the effectiveness of our proposed framework.

REFERENCES

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Derek Chen and Zhou Yu. Gold: Improving out-of-scope detection in dialogues using data augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 429–442, 2021.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5062–5074, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Thomas G Dietterich. Steps toward robust artificial intelligence. *Ai Magazine*, 38(3):3–24, 2017.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2019.

Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 506–514, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7764–7771, 2020.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1435–1449, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tür. Towards textual out-of-domain detection without in-domain labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1386–1395, 2022. doi: 10.1109/TASLP.2022.3162081.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 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)*, pp. 1311–1316, 2019.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Sungjin Lee and Igor Shalyminov. Contextual out-of-domain utterance handling with counterfeit data augmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7205–7209. IEEE, 2019.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2642–2652, 2020.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33, 2020.

Petr Marek, Vishal Ishwar Naik, Anuj Goyal, and Vincent Auvray. Oodgan: Generative adversarial network for out-of-domain data generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp. 238–245, 2021.

Johannes EM Mosig, Shikib Mehri, and Thomas Kober. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*, 2020.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. A simple but effective method to incorporate multi-turn context with bert for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pp. 11–17, 2019.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13675–13682, 2021.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. A survey on spoken language understanding: Recent advances and new frontiers. In *IJCAI*, 2021.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. Enhancing the generalization for intent classification and out-of-domain detection in slu. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2443–2453, 2021.

Lei Shu, Hu Xu, and Bing Liu. Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2911–2916, 2017.

Lei Shu, Yassine Benajiba, Saab Mansour, and Yi Zhang. Odist: Open world classification via distributionally shifted instances. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3751–3756, 2021.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 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)*, pp. 3566–3572, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1364. URL https://aclanthology.org/D19-1364.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2021.

Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2018.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys (CSUR)*, 2021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Jin Xu, Zishan Li, Bowen Du, Miaomiao Zhang, and Jing Liu. Reluplex made more practical: Leaky relu. In *2020 IEEE Symposium on Computers and communications (ISCC)*, pp. 1–7. IEEE, 2020.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. Unsupervised out-of-domain detection via pre-trained transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1052–1061, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.85. URL https://aclanthology.org/2021.acl-long.85.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam. Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1050–1060, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.99. URL https://aclanthology.org/2020.acl-main.99.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert YS Lam. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1050–1060, 2020b.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Hong Xu, and Weiran Xu. Adversarial self-supervised learning for out-of-domain detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5631–5639, 2021.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3521–3532, 2021.

Hanlei Zhang, Hua Xu, Shaojie Zhao, and Qianrui Zhou. Learning discriminative representations and decision boundaries for open intent detection.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 167–174, 2021a. doi: 10.18653/v1/2021.acl-demo.20. URL https://aclanthology.org/2021.acl-demo.20.

Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14374–14382, 2021b.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

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

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1100–1111, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.84. URL https://aclanthology.org/2021.emnlp-main.84.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5129–5141, 2022.

A Additional Experimental Details

We use the BERT model (*bert-base-uncased*) provided in the Huggingface's Transformers library (Wolf et al., 2020) to implement f. Following (Zhang et al., 2021b), we add an averaging-pooling layer on top of BERT to obtain the representation of each input utterance. The classification head in p is implemented as two-layer MLPs with the LeakyReLU activation Xu et al. (2020), while the projection heads in $\mu(v^j)$ and $\Sigma(v^j)$ as three-layer MLPs. The projection dimension is 64. Following (Federici et al., 2019), we use Jensen-Shannon mutual information estimator (Hjelm et al., 2018) to maximize the MI between two latent variables. Following (Zhan et al., 2021), We use AdamW (Kingma & Ba, 2014) to fine-tune BERT using a learning rate of 1e-5 and Adam (Wolf et al., 2019) to train the MLP heads using a learning rate of 1e-4. The batch size is 25 for IND and unlabelled data, respectively. In the training stage, 10/15 epochs of training are first conducted in stage 1 for STAR_Full and STAR_Small, respectively; then, 10 epochs of training are conducted in stage 2; finally, 10 epochs of training are conducted in stage 3 with early stopping. We tried pool size of $\{20, 50, 90\}$ for the candidate samples. All hyper-parameters are tuned according to the classification performance over the IND samples on the validation set. We find that $\lambda = 2$ works well with all datasets. Each result is an average of 3 runs with different random seeds, and each run is stopped when we reach a plateau on the validation performance. ALL experiments were conducted in the Nvidia Tesla V100-SXM2 GPU with 32G graphical memory.

B More Details about Baselines

We get the baseline results (MSP, SEG, DOC, ADB, and DA-ADB) using an OOD detection toolkit TEXTOIR (Zhang et al., 2021a). We get the baseline result of Outlier by running their released codes. We re-implement CDA by using counterfeit OOD turns. For fair comparisons, all baselines are implemented by using BERT as the backbone.