Learn to Discover Dialog Intents via Self-supervised Context Pretraining

Anonymous NAACL submission

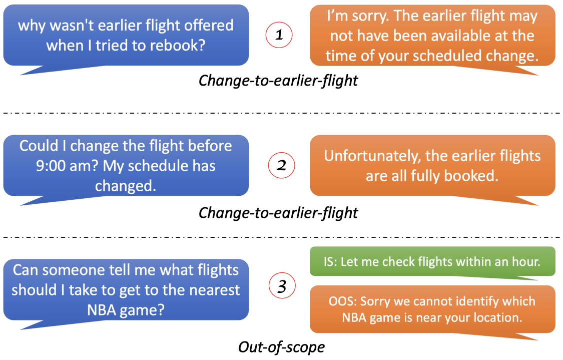
Abstract

Intent detection is one of most critical tasks in prevalent task-oriented dialog systems. However, most systems could only identify a fixed set of intents, without covering a ubiquitous space of real-world semantics. Inducing new dialog intents or excluding out-of-scope (OOS) queries are crucial particularly in complex domains like customer support. We present a simple yet effective intent induction schema via pre-training and contrastive learning. In particular, we first transform pretrained LMs into conversational encoders with in-domain dialogs. Then we conduct context-aware contrastive learning to reveal latent intent semantics via coherence from dialog contexts. By composing a fine-grained intent subspace from in-scope domain data, we demonstrate the effectiveness of our approach to induce intents with simple clustering algorithms and detect outliers with probabilistic linear discriminant analysis (pLDA). The experimental results validate the robustness and versatility of our framework, which also achieves superior performances over competitive baselines without label supervision.

1 Introduction

Recent advances of task-oriented agents such as Apple Siri, Amazon Alexa or Google DialogFlow have prompted the success of improving more natural customer service automation (Zhang et al., 2020b). They have been widely adapted in several domains like booking flights, restaurants, and customer support. Usually, the very first step of establishing such systems is to determine appropriate ontologies that constrain the dialog intent and state space for a specific task (Weld et al., 2021).

However, most works with such assumptions usually categorize utterances into few simple intents; yet exclude further pragmatic intent discovery in the real world (Hendrycks and Gimpel, 2018). Dialogs with different levels of stylized interactions may experience disparate variations of intent fluidity, where the scope of intents varies. In complex domains like customer support and healthcare, users may not be fully cognizant of the system capabilities, where 'out-of-scope' (OOS) queries would reasonably emerge during dialogs that fall out of bounds of system-supported intent scopes (Larson et al., 2019). Systems which can barely enumerate a complete view of user intents will easily fail to recognize an emerging intent during shifts between domains or chit-chats, as shown in Figure 1.


why wasn't earlier flight offered when I tried to rebook?

I'm sorry. The earlier flight may not have been available at the time of your scheduled change.

Could I change the flight before 9:00 am? My schedule has changed.

Unfortunately, the earlier flights are all fully booked.

Can someone tell me what flights should I take to get to the nearest NBA game?

IS: Let me check flights within an hour.

OOS: Sorry we cannot identify which NBA game is near your location.

Figure 1: Example of dialogs in TwACS dataset. First two examples have different narration but expressed the same intents. The last example has two different responses after identifying the query as in-scope (IS) or out-of-scope (OOS).

Second, unlike (written) texts, spoken utterances do not come with clear sentential segmentation by topographic means (Jonathan Ginzburg, 2010). This poses a more challenging understanding task for recent systems. Perkins and Yang (2020) introduced the dialog induction task to cluster user query utterances. It aims to learn a good discriminative classifier that predicts the same intents for utterances and their corresponding contexts. However, the performance may be degraded if representations of encoders trained from misleading objectives are hard to separate with accurate decision boundaries (Vulić et al., 2021). We argue that discovering fine-grained representations may be a more critical aspect of retrieving their latent intents that could be easily differentiated with simple clustering. In addition, OOS queries may still be allocated in one of intent clusters in their method.

Recently, contrastive learning based on pretrained Language Models (LM) such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), offers striking state-of-the-art (SOTA) performances across multiple natural language processing tasks (NLP) (Casanueva et al., 2020). They aim to enhance discrimination abilities of models by relying on semantic similarity between utterances, which exactly aligns with the clustering objectives based on the distance metrics. Hence, instead of emphasizing on algorithms that seek decision boundaries between data, is it possible to directly learn a good intent space where projected samples are distributed densely in clusters and easily segregated? Second, how could we quickly adapt pretrained LMs into a task-specific operator that generate such representations and discriminate OOS samples?

To solve the above two concerns and mitigate the data scarcity issue, we propose an entirely unsupervised approach without labeled instances to induce new dialog intents and exclude OOS queries from a large non-annotated corpus. Intuitively, we first deploy two-phase pre-training on LMs to learn a good representation that both inherits the coherence characteristics within dialogs and is easily discriminated semantically. Then we introduce a cluster operator to assign each query representation with their intents and exclude OOS ones with Probabilistic Linear Discriminant Analysis (pLDA) (Ioffe, 2006). We show that our pipeline is highly versatile to incorporate different LMs, pretraining strategies and cluster algorithms to tailor in each domain. The threshold in pLDA could be flexible chosen to trade off between accuracies of detecting IS/OOS queries.

Our contributions are summarized as follows:
1) We make the first attempt of introducing contrastive learning in the dialog induction task without any label supervision.
2) We introduce a flexible algorithm with a tunable threshold to detect OOS queries by forming clusters with only in-domain queries.
3) Experimental results verify the SOTA performance of our approach on two intent induction datasets and demonstrate the effectiveness of detecting OOS queries with self-supervised learning.

2 Problem Formulation

In this section, we first follow Perkins and Yang (2020) to formulate the dialog induction task as the clustering task. Suppose we have a training corpus of unannotated dialogs $X \in \mathcal{X}$ where each sample has the following format: $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{c}_i)$, $\mathbf{u}_i$ is the first query utterance and $\mathbf{c}_i$ with $N-1$ length is the following contexts. We hope to find an optimal operator $f : \mathcal{X} \rightarrow \mathcal{Y}$, $Y \subseteq \mathbb{R}^K$ to assign a new query $\mathbf{u}_{test}$ an one-hot vector y representing its cluster id k. We introduce a function $\phi_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ that maps an input into a hidden space $\mathcal{Z} \in \mathbb{R}^H$, where we restrict ϕ_θ in a language model function class Φ. We will finally try to find a good set of parameters θ and best f s.t. $y^*_{test} = \arg\min_{\phi\in\Phi,f,y} J(\phi_\theta(\mathbf{u}_{test}), f)$ over some clustering objective function J. We are also interested in detecting an OOS signal $y_{oos} \in \mathbb{R}^2$ for $\mathbf{u}_{test}$ whether a query is in domain, i.e. $y_{test} \in \mathcal{Y}$.

3 Methodology

3.1 P1: Self-supervised Pre-training (SSP)

We first deploy three language learning strategies to adapt pretrained LMs into conversational tasks: Mask Language Modeling (**MLM**), Mask Context Modeling (**MCM**) and Next Context Prediction (**NCP**). We first concatenate each utterance $\mathbf{u}_i$ and their contexts $\mathbf{c}_i$ with '[SEP]' as $\mathbf{x}_i$. Then, we will retrieve the feature representation $\mathbf{h}_i$ for the entire dialog $\mathbf{x}_i$ with each token as $\mathbf{x}_i^t$ with a pretrained LM, $\phi_\theta(\cdot)$ that implicitly learns sentential and conversational structures. We use two different projection matrices $\mathbf{W}_v \in \mathbb{R}^{H\times|V|}$ and $\mathbf{W}_c \in \mathbb{R}^{H\times 2}$ to transform $\mathbf{h}_i$ into one-hot vectors $\hat{\mathbf{y}}_i$, where H is BERT hidden size and $|V|$ is vocabulary size. We then update the parameters θ and $\mathbf{W}_v$ or $\mathbf{W}_c$ to minimize the cross entropy losses as follows:

$$\mathcal{L}_v(\theta, \mathbf{W}_v) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{y}_i log\langle\phi_\theta(\mathbf{x}_i^t), \mathbf{W}_v\rangle \tag{1}$$

$$\mathcal{L}_c(\theta, \mathbf{W}_c) = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i log\langle\phi_\theta(\mathbf{x}_i), \mathbf{W}_c\rangle \tag{2}$$

1) MLM: Similarly in BERT, we perform masking on 15% of input tokens $\mathbf{x}_i$ with 80% of the time a special token '[MASK]', 10 % of the time random word tokens from vocabulary and others remaining the same; and update LM with loss in Eq 1.
2) MCM: **MLM** does not take any account of sequential relations between utterances and their contexts. Instead, we only mask tokens in the contexts $\mathbf{c}_i$ to explicitly condition on utterances for token prediction and update LM with loss in Eq 1.
3) NCP: We directly predict whether the entire dialog is reasonable based on $\mathbf{u}_i$ and the given $\mathbf{c}_i$. And update LM with the loss in Eq 2. For each $\mathbf{u}_i$,

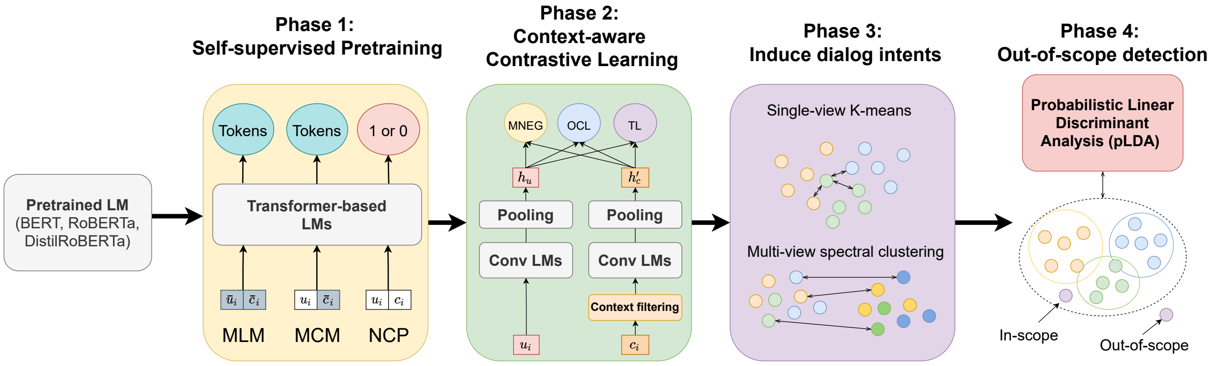


Figure 2: Illustration of the proposed framework for dialog intent induction and out-of-scope detection. There are four phases including pretraining and detection. $\bar{u}_i$ means it is randomly masked. Conv is short for Conversational.

we also randomly sample a negative context $\mathbf{c}_{j\setminus i}$ to construct a negative pair $(\mathbf{x}_i, \mathbf{c}_{j\setminus i})$.

3.2 P2: Context-aware contrastive Learning (CACL)

To better align with the downstream metric-based clustering, inspired by the recent success of contrastive learning in few-shot setting (Vulić et al., 2021; Zhang et al., 2021b), we found it is rather appropriate to introduce similarity-based pretraining for retrieving distant-sensitive[1] representations for clustering, instead of pseudo labeling samples directly. As shown in Lee et al. (2021) of contrastive learning setting, if we have the input variable X_1 and its target random variable X_2 for the pretext task, given the knowledge of their latent label Y, one can possibly predict X_2 without knowing much about X_1. At this case, X_2 is approximately independent of X_1 conditional on the label Y ($X_1 \perp X_2|Y$). In other words, if we try to learn a good representation $\phi(\mathbf{x}_1)$ close to ϕ^* in Eq 3 by semantically minimizing distances between X_1 and X_2, we implicitly force LMs to first predict Y; and predict X_2 from Y. And we could possibly linearly separate such learned representations ϕ^* with small representation errors.

$$\phi^* = \arg\min_{\phi \in \Phi} \mathbb{E}||\phi(X_1) - \phi(X_2)||^2 \qquad (3)$$

In our case, X_1 and X_2 are the given utterance queries and their corresponding contexts. Intuitively, successful dialogs should preserve coherence where X_1 and X_2 should be driven by the same underlying intent Y, regardless of being causes or effects of the intent. It elucidates that with a good representation ϕ^*, we shall easily cluster X_1 based on similarity by inducing their underlying intents Y. Therefore, we leverage the dual-encoder architectures to model the relevance between an utterance and their context pair $(\mathbf{u}, \mathbf{c})$. We will use the pretrained encoder $\phi(\cdot)$ from section 3.1 to obtain $(\mathbf{h}_u, \mathbf{h}_c)$. And we take the final encoding from the '[CLS]' token via a *pooling* operation.

Context Filtering Before sending data pairs into the encoder, some redundant sentences $\mathbf{c}_i$ in the dialog contexts $\mathbf{c} = \{\mathbf{c}_i\}$ may cause frictions to obfuscate user intents. Therefore, instead of encoding all dialog contexts, we first obtain the hidden representations for each sentence in a dialog $(\mathbf{h}_u, \mathbf{h}_c^1, ..., \mathbf{h}_c^{L-1})$ where L is the dialog length. Then we calculate the distance (cosine similarity) between each dialog context representation and the utterance. Eventually, we take the closest context as the final sentence pair $(\mathbf{h}_u, \mathbf{h}_c^{i'})$, where $i' = \arg\min_{i \in [1, L-1]} cos(\mathbf{h}_u, \mathbf{h}_c^i)$.

For a given pair $(\mathbf{h}_u, \mathbf{h}_c^{i'})$, we then introduce three strategies to finetune our encoder ϕ.

1) Context Ranking: We leverage the standard multiple negatives ranking loss (MNEG) (Henderson et al., 2017) to rank the correct context $\mathbf{c}_i'$ for the utterance u_i over other negative contexts $\mathbf{c'}_{j\setminus i'} \notin \{\mathbf{c}_i\}$ in a single batch. MNEG loss for a single batch $(u_1, c_1), ..., (u_{N_B}, c_{N_B})$ is calculated as the following:

$$\mathcal{L}_{MNEG} = -\sum_{i=1}^{N_B} S \cdot cos(u_i, c_i) + \sum_{i=1}^{N_B} log \sum_{j=1, j \neq i}^{N_B} e^{S \cdot cos(u_j, c_j)} \qquad (4)$$

where S is a scaling factor. It will try to maximize correct match score and abate negative scores.

2) Online Contrastive Learning (OCL): We select only hard positive and negative pairs to update the contrastive loss (Robinson et al., 2021), where

[1] Here we mean the mapped representations preserve the *topological information* within a latent intent space.

$d(u_i, c_j) = 1 - cos(u_i, c_j)$ and γ is the margin:

$$\mathcal{L}_{OCL} = \begin{cases} d(u_i, c_j)^2 & \text{if } j = i' \\ (ReLU(\gamma - d(u_i, c_j)))^2 & \text{if } j \neq i' \end{cases} \tag{5}$$

3) Triplet Loss (TL): Finally we also try to compare positive and negative pairs at once by minimizing the following triplet loss:

$$\mathcal{L}_{TL} = \max(\gamma + cos(u_i, c'_i) - cos(u_i, c_{j \neq i'}), 0) \tag{6}$$

It will prompt the minimization of the distance from the anchor example and approximate that of the negative example to the margin γ.

3.3 Clustering

After two-phase context learning, our main goal is to induce the intents from a collection of new user queries. We follow Perkins and Yang (2020) to formulate it as an unsupervised clustering task that involves two popular clustering algorithms *k-means* and *spectral clustering* as an operator $f : x \rightarrow k(x) = m^{k(x)}$, where $k(x)$ is the cluster id associated with its mean $m^{k(x)}$. *k-means* only requires the query-view vectors from our pretrained encoder ϕ while *multi-view spectral clustering (MVSC)* will solicit both query-view and content-view representations for matching.

3.4 Out-of-scope (OOS) Detection

After forming the query clusters, we are further interested whether such in-domain gaussian-like distributions could help us exclude the future out-of-scope queries. Here we propose to use the idea of **Probabilistic Linear Discriminant Analysis (pLDA)** (Ioffe, 2006) with predicted labels generated from our *kmeans* operator $f(\cdot)$. The goal of pLDA is to further project query samples onto a latent space such that samples from the same class remain in the same distribution. It can generate class center using continuous non-linear functions even from single example of unseen class. By first obtaining the predicted labels y_i from clustering, we could model both the distribution of utterances $x \sim N(x|y, \phi_w)$ and class labels $y \sim N(y|m, \Phi_b)$, where m is the parameter, Φ_w and Φ_b are the within/between-class covariance matrices. Latent variables u, v will represent the example of class and class variable in the projected space. For each class y_k, we compute the mean latent representations $\bar{u}_k$ from utterances belonging to class y_k. Finally, for a new test query x_{test}, we could calculate the likelihood ratio R_k with each class's $\bar{u}_k$:

$$R(u_{test}, \bar{u}_k) = \frac{P(u_{test}, \bar{u}_k)}{P(u_{test})P(\bar{u}_k)} \tag{7}$$

If all of $R(u_{test}, \bar{u}_k)$ are below a threshold T, we could conclude the query is out-of-scope that does not belong to any of the in-domain clusters.

Dataset	# Samples	# Intent	# Domain
TwACS (Perkins and Yang, 2020)	28,857	14	1 (Airline)
AskUbuntu (Perkins and Yang, 2020)	69,927	20	1 (Ubuntu)
CLINC (Larson et al., 2019)	23,700	150	10
BANKING (Casanueva et al., 2020)	13,083	77	1 (Banking)

Table 1: Statistics for Intent Datasets.

Supervised Fine-tuning To further validate the pLDA-based OOS detection performance from trained representations, we also perform similarity-based contrastive learning on datasets with K labels available. We generate a sets of n positive pairs $(\mathbf{x}_i, \mathbf{x}_j)$ for each intent y_k and obtain totally $K \times n$ samples. Then, we use LM encoder $\phi(\cdot)$ to generate representations $\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)$. Then we adopt MNEG loss mentioned above to conduct the ranking task. Finally, we use the pairs $(\phi(\mathbf{x}_i), y_i)$ to train pLDA model and perform OOS detection.

4 Experimental Setup

4.1 Datasets

We collect four datasets with two of them (TwACS, AskUbuntu) having smaller amount of annotated user intents (Perkins and Yang, 2020) and two of them (CLINC, BANKING) (Larson et al., 2019) with much larger intent annotated corpus. Table 1 shows the data statistics. For pretraining and contrastive learning purposes, we only use unlabeled corpus for training. For supervised contrastive learning, we use labeled training corpus from CLINC and BANKING. For evaluation datasets, there are three experimental settings:

1) Clustering: we use all annotated dialogs in TwACS and AskUbuntu to verify our model's performance on the dialog intent induction task.

2) Unsupervised OOS: we first generate pseudo training set by first sampling $n = 6000$ dialogs $\mathbf{x}_i \in \mathcal{X}$ as $\mathcal{X}_s$; then obtain their representations $\phi(\mathbf{x}_i)$ and their assigned clusters $\hat{y}_i = f(\phi(\mathbf{x}_i))$ via *k-means* as data pairs $(\phi(\mathbf{x}_i), \hat{y}_i)$. We use these training datasets to assess necessary parameters from pLDA. For evaluation, we randomly select 1000 samples from OOS test sets from CLINC and BANKING as OOS samples and select 1000 samples from IS training sets $\mathcal{X}_t \in \mathcal{X}$ $(\mathcal{X}_t \cap \mathcal{X}_s = \emptyset)$.

3) Supervised OOS: we directly use the training data $(\phi(\mathbf{x}_i), y_i)$ where y_i is the true label for pLDA

Dataset		TwACS				AskUbuntu			
Pretraining Method	**Clustering**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**
Autoencoder (Perkins and Yang, 2020)	k-means	27.1	18.7	22.2	15.9	27.9	19.3	22.8	14.4
Autoencoder (Perkins and Yang, 2020)	MVSC	29.3	21.2	24.6	18.9	28.6	13.0	17.9	11.2
Autoencoder (Perkins and Yang, 2020)	Avkmeans	46.0	33.6	38.9	31.6	50.2	43.8	46.8	36.2
QT (Logeswaran and Lee, 2018)	Avkmeans	51.1	38.1	43.7	35.9	56.6	55.1	55.8	43.2
DialoGPT (Zhang et al., 2020a)	k-means	28.2	16.2	20.6	15.1	20.1	15.4	17.5	13.3
BERT (Devlin et al., 2019)	k-means	26.3	19.5	22.4	15.8	22.7	15.6	18.5	13.8
TOD-BERT (Wu et al., 2020)	k-means	28.7	22.8	25.4	16.4	36.8	30.5	33.3	28.6
SBERT (Reimers and Gurevych, 2019)	k-means	37.6	28.0	32.1	26.7	53.3	44.8	48.7	41.5
BERT* (Devlin et al., 2019)	k-means	35.4	23.8	28.5	22.8	58.1	45.6	51.1	45.1
TOD-BERT* (Wu et al., 2020)	k-means	33.7	21.4	26.2	20.6	55.0	47.7	51.1	45.6
BERT (NCP + CACL) (ours)	k-means	**55.1**	**46.8**	**50.6**	42.4	68.3	52.6	59.5	47.8
BERT (MLM + CACL) (ours)	k-means	53.8	45.5	49.3	**44.0**	72.2	63.5	67.6	62.2
BERT (MCM + CACL) (ours)	k-means	47.5	37.0	41.6	34.6	**76.8**	63.0	69.2	61.2
BERT (NCP + CACL) (ours)	MVSC	53.0	43.5	47.8	42.0	74.2	61.2	67.1	58.0
BERT (MLM + CACL) (ours)	MVSC	53.6	40.3	46.0	39.0	75.8	63.6	69.2	61.9
BERT (MCM + CACL) (ours)	MVSC	51.6	40.7	45.5	38.1	76.1	**64.3**	**69.7**	**63.2**

Table 2: Evaluation results (%) on TwACS and AskUbuntu datasets for different pretrain and cluster strategies. QT is short for Quick Thoughts. $*$ indicates the model has been finetuned with in-domain corpus. NCP, MLM, MCM denote the different pretraining methods, followed by contrastive learning (CACL). The best results are in **bold**.

training. For evaluation, we combine the in-domain test set and OOS test set as the final testing set.

4.2 Model Variants and Baselines

We experiment with competitive baselines from some popular transformer-based LMs and clustering approaches. The pretraining methods include:
1) Autoencoders (Perkins and Yang, 2020) which are finetuned on TwACS and AskUbuntu datasets.
2) Quick thoughts (QT) (Logeswaran and Lee, 2018) which is a strong representation learning baseline that is adopted in BERT.
3) DialoGPT (Zhang et al., 2020a) which is generative model trained with dialog corpus.
4) BERT (Devlin et al., 2019) which is a contextualized representation model based on transformers.
5) TOD-BERT (Wu et al., 2020) which is finetuned on conversational data based on BERT.
6) Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) which deploys dual-encoder structure for similarity training.

We also experiment with a range of model variants enabled by our framework, and compare their performances against an array of cutting-edge conversational sentence encoders. The entire pipeline is specified as LM+P1+P2-LOSS, where we adopt a LM structure and pretrain it with Phase-1 (P1) pretraining and Phase-2 (P2) learning with specific loss (LOSS). For LMs, we use 1) BERT, 2) RoBERTa (ROB) as an improved variant of BERT, LM-pretrained with more data and 3) DistilRoBERTa (DROB), a distilled version of RoBERTa, trained with around 4 times fewer data than the teacher RoBERTa model. All of models contain 768-dimensional transformer layers with 12 (BERT, ROB) or 6 (DROB) attention layers. For P1, we have pretraining methods: MLM, MCM, NCP. And we use MNEG, TL, OCL for P2-LOSS.

For OOS detection, we use binary classification with pretrained LMs as our baselines. BERT-all indicates that we train all in-scope data and 250 out-of-scope queries. To avert class imbalancing, we further subsample only 1000 in-scope data for BERT, ROB and DROB model. We also pretrained BERT with in-scope data beforehand as BERT*. Eventually we train different LMs as specified in § 3.4 and detect OOS queries with pLDA.

4.3 Experimental Setting

We use Sentence-Transformers package (Reimers and Gurevych, 2019) to implement our framework. We first pretrain the LMs with P1-SSP. We train for 4 epochs in batches of 16. We also set the max length as 200 and mask probability of 0.15. For P2-CACL, we train each model with 2 epochs in batches of 8. And for supervised contrastive learning, we randomly sample $n = 500$ positive query pairs for each class and train models with 10 epochs in batches of 10. Margin γ is set to be 5. We set the threshold T in pLDA as -1.7 for CLINC and 0 for BANKING empirically. We use the AdamW optimizer and $2e-5$ learning rate and weight decay rate is 0.01. We run each experiment 5 times and report the average. We follow the metrics used in Perkins and Yang (2020) for clustering precision/recall, f1 score and accuracy.

5 Results and Discussion

5.1 Main Clustering Results

The main results are summarized in Table 2. First we observe that by directly deploying general

Dataset		TwACS				AskUbuntu			
Pretraining LM	Loss	P	R	F1	ACC	P	R	F1	ACC
DROB (Sanh et al., 2020)	MNEG	53.8	42.2	47.3	39.4	73.8	61.6	67.2	57.9
	TL	51.4	42.5	46.5	40.0	71.5	59.7	65.1	56.8
	OCL	47.9	34.6	40.2	32.8	66.3	54.8	60.0	52.4
ROB (Liu et al., 2019)	MNEG	51.4	40.7	45.4	38.7	73.9	62.4	67.7	59.2
	TL	**55.1**	42.9	48.3	40.9	72.9	62.8	67.5	60.1
	OCL	43.5	38.3	40.8	35.9	69.5	56.8	62.5	53.4
BERT (Devlin et al., 2019)	MNEG	53.8	**45.5**	**49.3**	**44.0**	72.2	**63.5**	67.6	**62.2**
	TL	54.9	41.6	47.3	39.0	**74.4**	63.4	**68.4**	62.0
	OCL	48.8	35.9	41.4	33.0	72.0	59.6	65.2	57.5

Table 3: Clustering Performance (%) of our proposed approach with different pretraining LMs and losses. The best results are in **bold**. We use MLM pretraining for all BERT-based structures.

BERT or DialoGPT, universal representations do not reveal clear boundaries for intent separation, causing relatively low performances. Training with conversational data like TOD-BERT offers better disentanglement of intent features for each sample. We then observe even competitive performance when we apply SBERT dual-encoder structure. The results suggest that the pretraining objectives and corpus may substantially influence the downstream clustering task. This prompts us to finetune pretrained LMs like BERT and TOD-BERT with P1-SSP. The results unanimously suggest the improved performance even over the baselines by using previous SOTA Avkmeans algorithm in AskUbuntu.

The power of contrastive learning By introducing the P2-CACL, the proposed approaches achieve the best performance across two datasets, particularly at NCP+CACL+k-means in TwACS and MCM+CACL+MVSC in AskUbuntu. They successfully establish an intent subspace where utterances are close to their contexts, which congruently approximates the process of disclosing their underlying latent intents. Such cross-instance similarities are especially useful when we compare these representations based on distance metrics, rather than simply learning a MLP-based mapping from each sentence to its class. In addition, we also see that NCP by differentiating correct utterance-context pair has slight improvement in TwACS over MLM, while we see MCM surpasses MLM more in AskUbuntu setting with MVSC approach. The context-aware pretraining (MCM, NCP) seem to be more beneficial in MVSC where contexts are considered, than single-view *k-means* approach.

Impacts of input LMs and losses Table 3 displays the results by adopting different pretrained LMs and losses from § 3.1 during P2-CACL. It verifies the versatility and wide applicability of our proposed method. But the model variants may still naturally impact the absolute clustering performance. Surprisingly, we see a better performance of BERT-based structure over RoBERTa, which deduces the pretraining objectives may play crucial roles of dominating how clusters are formed. The comparison between DROB and BERT reveals that loss regime influences more than the parameter capacity in TwACS; reversely in AskUbuntu. Overall, both MNEG and TL yield similar strong results, while OCL is poor by using only hard examples.

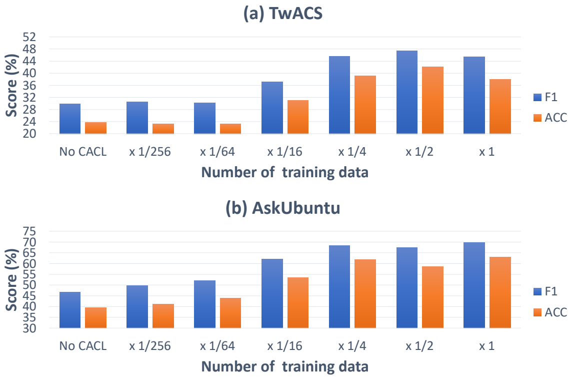


Figure 3: Varying the amount of the training data for Phase-2 CACL; ×1 refers to the training size used in all experiments, while other data sizes are relative to this corpus size (e.g. $\times 1/64$ denotes we only use $1/64$ of all data for training.)

Amount of training examples We also analyze what amount of in-domain corpus is adequate to fairly discriminate utterances' intents, i.e. to be decently separable upon projection onto the latent intent space. We reduce the amount of data through subsampling The scores are provided in Figure 3. As expected, more training data yield better performance on average while the absolute scores reach the bottleneck after increasing up to 1/4 of total training data. It implies that even few samples may fairly discriminate the utterance and surpass the models skipping Phase-2.

5.2 Ablation Study

A more careful ablation comparison is conducted in Table 4 to investigate the effects of different stages of pretraining and filtering method. Experimental

Dataset				TwACS				AskUbuntu			
SSP	**CACL**	**Filtered**	**Clustering**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**
MLM	✓	✓	k-means	53.8	**45.5**	49.3	**44.0**	72.2	63.5	67.6	**62.2**
MLM	✓	✓	MVSC	-0.2	-5.2	-3.3	-5	**+3.6**	**+0.1**	**+1.6**	-0.3
MLM	✓		k-means	+1.1	-1.3	-0.3	-3.3	-0.8	-3.4	-2.3	-3.8
	✓	✓	k-means	-0.2	-2.4	-1.5	-4.2	-3.1	-3.0	-3.0	-5.1
	✓		k-means	**+3.3**	+0.0	**+1.4**	-2.9	-3.7	-5.8	-5.0	-6.6
MLM			k-means	-18.4	-21.7	-20.8	-21.2	-14.1	-17.9	-16.5	-17.1
NCP			k-means	-16.2	-17.5	-17.2	-18.8	-22.8	-20.1	-21.4	-22.6
MCM			k-means	-17.9	-19.7	-19.3	-20.2	-19.8	-21.1	-20.7	-22.5
			k-means	-27.5	-26.0	-26.9	-28.2	-49.5	-47.9	-49.1	-48.4

Table 4: Ablation study (%) of our proposed approach with different component combination. **SSP** is the Phase-1 pretraining. **CACL** is the Phase-2 contrastive learning. **Filtered** is whether to use whole context or only filtered one. **Clustering** denotes the cluster method. Data points are specified as difference compared with the first row comprehensive model.

results indicate that all of components are necessary to achieve the peak performance in both datasets. Around 3-4 % performance drop is observed when we do not perform context filtering where redundant noises may disturb. Forfeiting P1-SSP will result in further drops where we believe it will benefit the model discrimination ability of semantically similar utterances. In addition, we found large performance drops by only adopting P1-SSP in all three pretraining methods, which indicates the necessity of contrastive learning in differentiating samples with aid of their contexts. We also found that NCP is more beneficial in TwACS and MLM in AskUbuntu than P2-only models.

5.3 Out-of-scope Detection

Unsupervised detection Figure 4 shows the ROC curves of our unsupervised outlier detectors in both datasets. We could observe with varying threshold T, the specificity (True negative rate) will be inversely proportional to recall where our classifier predicts more outliers correctly and incorrectly. But surprisingly, without any label supervision, pLDA could still achieve around average 0.7 AUC score. We observe models pretrained with MLM perform better in TwACS and MCM in AskUbuntu. It alludes that models rely more on contexts in AskUbuntu with more turns than TwACS. In addition, we care more about the recall than the precision since precision errors will prompt fallback responses to make users reiterate. Hence, we could sacrifice little on in-scope accuracy and achieve higher recall by tuning T.

Supervised detection Table 5 indicates the results of supervised OOS detection. We could observe that binary classification with LMs may give a high in-domain accuracy and low recall since the model tends to overfit, significantly depending on the amount of OOS data during training (extreme low recall by leveraging all in-scope data). Instead, even not relying on OOS train data, our method could still achieve competitive true negative rates while maintaining superior recall especially in BANKING. In this case, Phase-1 pretraining seems to be less conducive to assist in out-of-scope detection. Models with RoBERTa backbone perform much better than the other model variants. Figure 4 also shows the PR curve for OOS detection. By specifying the same range of available thresholds, we observe that models with RoBERTa backbone have both higher precision and recall. Even models with DistilRoBERTa backbone may have competitive results with BERT-based model.

Dataset		CLINC		BANKING	
Classifier	**Detector**	**TN**	**R**	**TN**	**R**
BERT-all	binary	99.5	16.4	98.8	3.70
BERT	binary	94.4	46.5	93.3	32.6
BERT*	binary	98.9	43.9	95.0	18.9
ROB	binary	97.3	52.5	94.6	36.8
DROB	binary	97.5	51.8	93.5	43.6
S-BERT	pLDA	94.3	52.2	92.5	51.6
S-ROB	pLDA	**96.4**	**55.6**	93.5	**67.0**
S-DROB	pLDA	94.4	53.2	91.8	62.0
S-BERT*	pLDA	93.4	55.5	**93.7**	55.9

Table 5: Experimental results (%) of OOS detection using different supervised pretraining strategies. $*$ indicates that the model is finetuned with in-domain corpus. **TN** refers to the true negative rate and **R** refers to Recall.

5.4 t-SNE visualization

To understand how well the latent space is formulated where encoded utterances are mapped onto, we perform t-SNE visualization in Figure 5 on three types of representations generated with disparate pretrained encoders. First, we could see without any finetuning, utterances with distinct intents are entangled into a single mixed cluster. After Phase-1 of transforming LMs into conversational encoders, we start to see some clusters with intents apart. Eventually, Phase-2 further specializes the sentence encoder to learn meaningful task-related semantic clusters even without any label supervision.

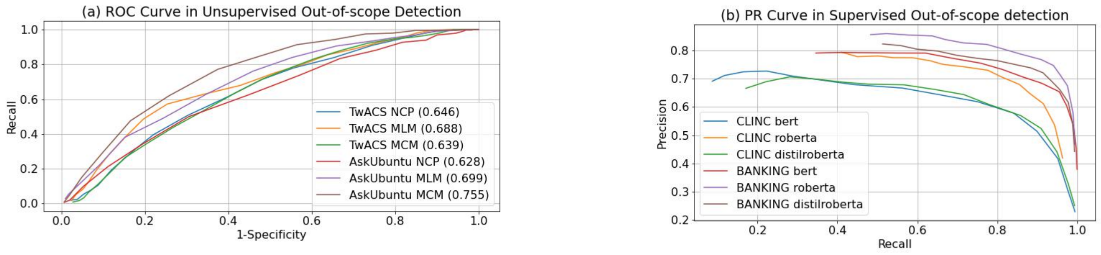
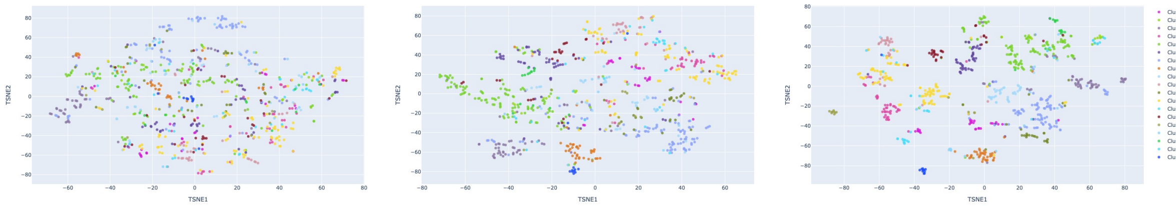


Figure 4: (a) ROC curves for proposed unsupervised OOS approach. The score behind the model name in each legend indicates the Area-Under-Curve (AUC) score. (b) PR curves for proposed supervised OOS approach.



(a) BERT (no fine-tuning)

(b) BERT (Phase 1: MCM)

(c) BERT (Phase 1: MCM + Phase 2: CACL)

Figure 5: t-SNE plots of encoded utterances from AskUbuntu. The representations are created by BERT-based encoder ϕ (a) without any finetuning (b) after Phase 1 MCM pretraining and (c) after Phase 1 and phase 2 contrastive learning. The colors are true labels.

6 Related Work

User intent clustering Clustering user utterances to discover new intents is a critical task that disambiguates utterances. Early works focused on feature engineering for query traits to perform high-quality clustering (Kathuria et al., 2010; Cheung and Li, 2012; Padmanabhan, 2016). Haponchyk et al. (2018) further leveraged supervised signals to detect duplicate questions. Samir Kanaan-Izquierdo and Perera-Lluna (2018) also proposed multi-view spectral clustering to consider consistent cluster assignments across different views. Deep clustering was proposed to better represent data features (Xie et al., 2016; Zhang et al., 2021a; Chang et al., 2017; Nguyen et al., 2021). Perkins and Yang (2020) further finetuned pretrained representations for multi-view *kmeans* clustering.

Out-of-scope detection Until recently, the idea of OOS data is considered to address possibility of queries that fall out of intended output classes (Hendrycks and Gimpel, 2018). Usually the problem is formed as a binary classification task where Larson et al. (2019) provided a large-scale evaluation dataset with some benchmark classifier results. Data augmentation (Zhan et al., 2021; Chen and Yu, 2021) and outlier detection (Lin and Xu, 2019; Yilmaz and Toraman, 2020; Cavalin et al., 2020) are two main streams to tackle such problem. Here we introduce new concepts of detecting OOS queries from probabilistic formed intent clusters.

Contrastive learning Models trained with contrastive learning have rendered striking performances in many NLP tasks (Casanueva et al., 2020; Gunel et al., 2021; Zhang et al., 2021b). In particular, Gunel et al. (2021) proposed supervised contrastive learning for pretrained models on several benchmark tasks. Wu et al. (2019) trained models to learn dialog orders. Zhang et al. (2021b) and Vulić et al. (2021) proposed to transform pretrained LMs by masking tokens and response reanking; leveraging similarity-based learning for few-shot intent detection. Our work differs from them that we specifically tackle the intent clustering task with entirely unsupervised fashion.

7 Conclusion

In this work, we present a simple yet effective two-phase self-supervised pipeline to induce dialog intents with clustering and detect OOS queries accordingly. We propose to transform LMs into conversational encoders and retrain them with context-aware similarity-based learning. We demonstrate that by extracting such intent-aware representations, it is possible to separate dialogs based on their underlying intents. Without any supervision, we could flexibly obtain decent recall on precluding queries that do not belong in task-specific domains.

References

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Paulo Cavalin, Victor Henrique Alves Ribeiro, Ana Appel, and Claudio 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)*, pages 3952–3961, Online. Association for Computational Linguistics.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2017. Deep adaptive image clustering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5880–5888.

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

Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 383–392, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning.

Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti. 2018. Supervised clustering of questions into intents for dialog system applications. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2310–2321, Brussels, Belgium. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Dan Hendrycks and Kevin Gimpel. 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks.

Sergey Ioffe. 2006. Probabilistic linear discriminant analysis. In *Computer Vision – ECCV 2006*, pages 531–542, Berlin, Heidelberg. Springer Berlin Heidelberg.

Raquel Fernández Jonathan Ginzburg. 2010. *Computational Models of Dialogue*.

Ashish Kathuria, Jim Jansen, Carolyn Hafernik, and Amanda Spink. 2010. Classifying the user intent of web queries using k -means clustering. *Internet Research*, 20:563–581.

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.

Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. 2021. Predicting what you already know helps: Provable self-supervised learning.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss.

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.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations.

Xuan-Bac Nguyen, Duc Toan Bui, Chi Nhan Duong, Tien D. Bui, and Khoa Luu. 2021. Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10842–10851.

Padmanabhan. 2016. Mixkmeans : Clustering question-answer archives.

Hugh Perkins and Yi Yang. 2020. Dialog intent induction with deep multi-view clustering.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples.

Andrey Ziyatdinov Samir Kanaan-Izquierdo and Alexandre Perera-Lluna. 2018. Multiview and multi-feature spectral clustering using common eigenvectors. volume 102. Pattern Recognition Letters.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Ivan Vulić, Pei-Hao Su, Sam Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. Convfit: Conversational fine-tuning of pretrained language models.

H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han. 2021. A survey of joint intent detection and slot-filling models in natural language understanding.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Self-supervised dialogue learning.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA. PMLR.

Eyup Halit Yilmaz and Cagri Toraman. 2020. *KLOOS: KL Divergence-Based Out-of-Scope Intent Detection in Human-to-Machine Conversations*, page 2105–2108. Association for Computing Machinery, New York, NY, USA.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y. S. Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021a. Discovering new intents with deep aligned clustering.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021b. Few-shot intent detection via contrastive pre-training and fine-tuning.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020a. Dialogpt: Large-scale generative pre-training for conversational response generation.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang, and Xiaoyan Zhu. 2020b. Recent advances and challenges in task-oriented dialog system.