New Intent Discovery with Attracting and Dispersing Prototype

Shun Zhang[1,2]**, Jian Yang**[1*]**, Jiaqi Bai**[1,2]**, Chaoran Yan**[1]**,**
Tongliang Li[3]**, Zhao Yan**[4]**, Zhoujun Li**[1,2]

[1] State Key Lab of Software Development Environment, Beihang University, Beijing, China
[2] School of Cyber Science and Technology, Beihang University, Beijing, China
[3] Beijing Information Science and Technology University, [4] Tencent Youtu Lab
{shunzhang, jiaya,bjq,ycr2345,lizj}@buaa.edu.cn
{tonyliangli}@bistu.edu.cn;{zhaoyan}@tencent.com

Abstract

New Intent Discovery (NID) aims to recognize known and infer new intent categories with the help of limited labeled and large-scale unlabeled data. The task is addressed as a feature-clustering problem and recent studies augment instance representation. However, existing methods fail to capture cluster-friendly representations, since they show less capability to effectively control and coordinate within-cluster and between-cluster distances. Tailored to the NID problem, we propose a **R**obust and **A**daptive **P**rototypical learning (RAP) framework for globally distinct decision boundaries for both known and new intent categories. Specifically, a robust prototypical attracting learning (RPAL) method is designed to compel instances to gravitate toward their corresponding prototype, achieving greater within-cluster compactness. To attain larger between-cluster separation, another adaptive prototypical dispersing learning (APDL) method is devised to maximize the between-cluster distance from the prototype-to-prototype perspective. Experimental results evaluated on three challenging benchmarks (CLINC, BANKING, and StackOverflow) of our method with better cluster-friendly representation demonstrate that RAP brings in substantial improvements over the current state-of-the-art methods (even large language model) by a large margin (average $+5.5\%$ improvement).

Keywords: New Intent Discovery, Robust Prototypical Attracting, Adaptive Prototypical Dispersing

1. Introduction

Due to the success of conventional intent detection in dialogue systems (Wu et al., 2019; Liu and Mazumder, 2021; Zhang et al., 2023a, 2022), the vast majority of learning algorithms under the closed-world scenario with static data distribution only consider pre-defined intents. To handle the new intents outside the existing known intent categories, it is necessary to equip dialogue systems with new intent discovery (NID) abilities (Raedt et al., 2023; Mou et al., 2022; Siddique et al., 2021; Fini et al., 2021; Chrabrowa et al., 2023).

Early works (Hakkani-Tür et al., 2013, 2015; Shi et al., 2018; Padmasundari and Bangalore, 2018) mainly adopt unsupervised clustering with unlabeled data. They always ignore prior knowledge of the available labeled data and fail to generate highly accurate and granular intent groups, leading to inapplicability in the open-world scenario. Recent studies are adept to semi-supervised settings to efficiently utilize the limited labeled data, such as pairwise similarities (Lin et al., 2020), iterative pseudo-labeling (Zhang et al., 2021a), probabilistic architecture (Zhou et al., 2023b) and prototypical network (An et al., 2023). Due to the transferability of injecting structural knowledge from known categories into the intent representation, the semi-supervised methods can be extended to real-world scenarios with better NID performance.

(a) Lacking Within-cluster Compactness   **(b) Lacking Between-cluster Separation**
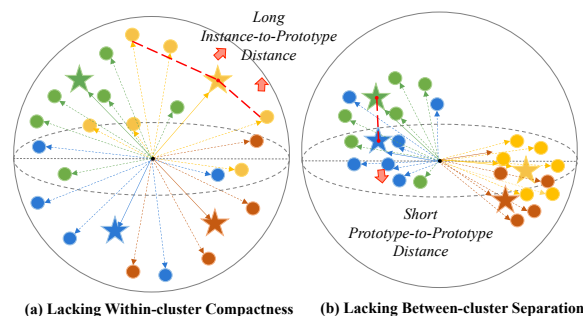
Figure 1: Embedding distribution of intent instances and the prototype of each class in a shared sphere semantic space. The circle and star shape denote the instance and the prototype, respectively. The discriminative representations fail to be extracted due to insufficient (a) within-cluster compactness and (b) between-cluster separation.

However, existing semi-supervised methods still face two challenges: **(C1)**: Lacking sufficient within-cluster compactness within the learned intent representations. **(C2)**: Requiring the explicit modeling of between-cluster dispersion in the representation space. In Figure 1(a), different intent categories overlap with each other with varying distributions, where the large class-specific variances result in distant instance embeddings from the cluster centers. In Figure 1(b), embeddings in different categories tend to gather in the same region, due to the narrow distance among cluster

---

*Corresponding author.

centers. Previous baselines encounter difficulty in capturing clear and accurate cluster boundaries for known and novel categories. Therefore, *Jointly constraining both within- and between-cluster distance to yield cluster-friendly discriminative representations still requires further exploration.*

In this work, we propose a **R**obust and **A**daptive **P**rototypical learning (RAP) framework for the joint identification of known intents and discovery of the novel intents. To mitigate the issue of insufficient within-cluster compactness, we introduce a robust prototypical attracting learning (**RPAL**) method to reduce the overlarge variances from an instance-to-prototype perspective. Specifically, we first compute class prototypes as normalized mean embeddings and enforce each instance embedding to be closer to its corresponding class prototypes. To avoid the negative impact of pseudo-label noise, a novel interpolation training strategy is used to construct virtual training samples for the maintenance of the same linear relationship with its prototypes. To address the concern of insufficient between-cluster separation, we propose an adaptive prototypical dispersing learning (**APDL**) to explicitly enlarge the between-cluster distance from a prototype-to-prototype perspective. By minimizing the total prototype-to-prototype similarities, APDL adaptively maximizes the distance between prototypes to form well-separated clusters. A weighted training objective is used to adaptively impose large penalties on nearer prototypes to push them further apart, helping achieve better dispersion among prototypes. Finally, RPAL and APDL are jointly optimized with multitask learning to guide the model to learn cluster-friendly intent representations for both known and novel intents.

Experimental results on multiple NID benchmarks demonstrate our method brings substantial improvements over previous state-of-the-art methods by a large margin of $+5.5\%$ points. Our key contributions are summarized as follows:

- A new prototype-guided learning framework is designed to learn cluster-friendly discriminative representations with stronger within-cluster compactness and larger between-cluster separation.

- We propose a robust prototypical attracting learning method and an adaptive prototypical dispersion learning method, which solve the problems of insufficient within-cluster compactness and between-cluster separation.

- Extensive experiments on three benchmark datasets show that our model establishes state-of-the-art performance on the semi-supervised NID task (average $+5.5\%$ improvement), which demonstrates competitive NID performance.

2. Related Work

2.1. New Intent Discovery

Existing NID methods can be divided into two categories: unsupervised and semi-supervised. For the former, pioneering works (Hakkani-Tür et al., 2013, 2015) primarily rely on statistical features of the unlabeled data to cluster similar queries for discovering new user intents. Subsequently, some studies (Xie et al., 2016; Yang et al., 2017; Shi et al., 2018) endeavor to leverage deep neural networks to learn robust representations conducive to new intent clustering. However, these methods lack the capacity to leverage prior knowledge for clustering guidance. Addressing this limitation, some studies (Basu et al., 2004; Hsu et al., 2018, 2019; Han et al., 2019) begin to explore the use of semi-supervised clustering methods to better leverage prior knowledge. For example, Lin et al. (2020) combines the pairwise constraints and target distribution to discover new intents while Zhang et al. (2021a) introduces an alignment strategy to improve the clustering consistency. Further, Shen et al. (2021); Kumar et al. (2022); Vaze et al. (2022); Zhang et al. (2022) designs contrastive learning strategies in both the pre-training phase and the clustering stage to learn discriminative representations of intents. Recently, Zhou et al. (2023b) introduces a principled probabilistic framework and An et al. (2023) proposed a decoupled prototypical network to enhance the performance of the NID. However, these methods fail to effectively capture discriminative representations with strong within-cluster compactness and large between-cluster separation. This difficulty makes it challenging to differentiate between the characteristics of known and novel intents.

2.2. Prototypical Learning

Prototypical learning (PL) methods (Snell et al., 2017) have become promising approaches due to their simplicity and effectiveness and they have been widely applied in various scenarios, such as unsupervised domain adaptation (Yue et al., 2021), out-of-domain detection (Zhang et al., 2023b,c), machine translation (Yang et al., 2021b, 2019, 2021c, 2020; Chai et al., 2024), and named entity recognition (Zhou et al., 2023a; Yang et al., 2022; Mo et al., 2023). Among them, prototypical contrastive learning (Li et al., 2021) is proposed to generate compact clusters. It employs cluster centroids as prototypes and trains the network by drawing instance embeddings closer to its assigned prototypes. Here, we explore the utilization of the interpolation training strategy to enhance cluster compactness while mitigating the effects of label noise, thus rendering it more robust.

3. Approach

In this section, we describe the proposed RAP framework for new intent discovery in detail. As shown in Figure 2, the architecture of the RAP framework contains four main components: (1) *Intent representation learning*, which pre-trains a feature extractor E_θ on both labeled and unlabeled intent data to optimize better representation learning (Sec. 3.2); (2) *Categorical prototypes generation*, which derives the prototypes of the training data by a clustering method (Sec. 3.3); (3) *Robust prototypical attracting*, which mitigates the effects of noisy pseudo-labels while minimizes the instance-to-prototype distance for stronger within-cluster compactness (Sec. 3.4); (4) *Adaptive prototypical dispersing*, which maximizes the prototype-to-prototype distance for larger between-cluster dispersion (Sec. 3.5).

3.1. Problem Definition

New Intent Discovery follows an open-world setting, which aims to recognize all intents with the aid of limited labeled known intent data and unlabeled data containing all classes. Let $\mathcal{I}_k$ and $\mathcal{I}_n$ represent the sets of known and novel intents respectively, where $\{\mathcal{I}_k \cap \mathcal{I}_n\} = \varnothing$ and $|\mathcal{I}_k| + |\mathcal{I}_n| = C$, where C represents the total number of intent categories. A typical NID task comprises a set of labeled training set $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^N$, wherein each intent $y_i \in \mathcal{I}_k$, and a set of unlabeled intent utterances $\mathcal{D}_u = \{(x_i)\}_{i=1}^M$, where the intent of each utterance x_i belongs to $\{\mathcal{I}_k \cup \mathcal{I}_n\}$. The goal of semi-supervised NID is to use $\mathcal{D}_s$ as prior knowledge to help learn clustering-friendly representations to recognize known and discover new intent groups. After training, the model performance will be evaluated on the testing set $\mathcal{D}_t = \{x_i | y_i \in \mathcal{I}_k \cup \mathcal{I}_n\}$.

3.2. Intent Representation Learning

Considering the excellent generalization capability of the pre-trained model, we use the pretrained language model BERT (Devlin et al., 2019) as our feature extractor ($E_\theta : \mathcal{X} \to \mathbb{R}^H$). Firstly, we feed the ith input sentence x_i to BERT, and take all token embeddings $[T_0, \ldots, T_L] \in \mathbb{R}^{(L+1) \times H}$ from the last hidden layer (T_0 is the embedding of the [CLS] token). The sentence representation $s_i \in \mathbb{R}^H$ is first obtained by applying mean-pooling operation on the hidden vectors of these tokens:

$$s_i = \text{mean-pooling}([[\text{CLS}], T_1, ..., T_L]) \qquad (1)$$

where [CLS] is the vector for text classification, L is the sequence length, and H is the hidden size. Motivated by (Zhang et al., 2021a), we aim to effectively generalize prior knowledge through pre-training to unlabeled data, we fine-tuned BERT on labeled data ($\mathcal{D}_s$) using the cross-entropy (CE) loss. Furthermore, we follow (Zhang et al., 2022) to use the masked language modeling (MLM) loss on all available data ($\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_u$) to learn domain-specific semantics. We concurrently pre-train the model with the aforementioned two types of loss:

$$\mathcal{L}_{pre} = \mathcal{L}_{ce}(\mathcal{D}_s) + \mathcal{L}_{mlm}(\mathcal{D}_s \cup \mathcal{D}_u) \qquad (2)$$

where $\mathcal{D}_s$ and $\mathcal{D}_u$ are labeled and unlabeled intent corpus, respectively. The masked language model is trained on the whole corpus $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_u$. After pretraining, models can acquire diverse general knowledge for both known and novel intents, enabling them to learn meaningful semantic representations for subsequent tasks.

3.3. Categorical Prototypes Generation

The prototype of an intent class computed as the text average embeddings within the class is defined as a representative embedding for a group of semantically similar instances. We first obtain the intent embedding $E_\theta(x_i)$ for each $x_i \in \mathcal{D}_a$ and then perform k-Means clustering on the training instances $\mathcal{D}_a$ to generate C clusters $Q = \{Q_c\}_{c=1}^C$, where C represents the total number of intent categories. The prototype matrix $P_r = \{\mu_c\}_{c=1}^C$ are generated based on their clusters. Concretely, for each cluster Q_c, the prototype μ_c is computed by:

$$\mu_c = \frac{1}{|Q_c|} \sum_{x_i \in Q_c} E_\theta(x_i) \qquad (3)$$

where we presume prior knowledge of C following previous works (Zhang et al., 2021a) to make a fair comparison and we tackle the problem of estimating this parameter in the experiment (refer to Sec. 5.4 for a detailed discussion on more accurately estimating C within RAP).

3.4. Robust Prototypical Attracting

The key idea of the robust prototypical attracting learning (RPAL) method aims to acquire cluster-friendly discriminative representations with strong within-cluster compactness. To minimize the distance between instances and their corresponding prototypes, we employ prototypical contrastive learning (PCL). PCL brings instance representations closer to their matched prototypes while pushing them away from other prototypes as:

$$\mathcal{L}_p = -\mathbb{E}_{i \leq N_b} \log \frac{\exp(z_i \cdot \mu^{y_i}/\tau)}{\sum_{k=1}^C \exp(z_i \cdot \mu^k/\tau)} \qquad (4)$$

where the normalized sentence embeddings z_i matches the prototype μ^{y_i} of its ground truth label y_i, N_b is the size of the training set and τ is a scalar temperature.
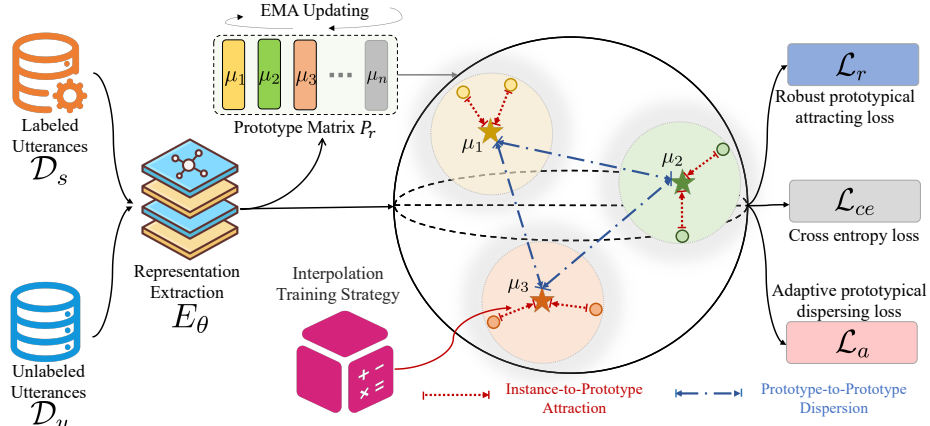
Figure 2: Overview of RAP. Our method is jointly optimized by L_r, L_a, and L_{ce}. L_r mitigates the effects of noisy pseudo-labels while minimizes the instance-to-prototype distance, while L_a maximizes the prototype-to-prototype distance. L_{ce} is a cross-entropy loss to prevent knowledge forgetting.

Interpolation Training Strategy Since the mapping between instances and prototypes derives from pseudo-labels generated by k-means clustering, the standard PCL objective is susceptible to noise influence, resulting in suboptimal performance (Jiang et al., 2020). The encoder necessitates a regularization technique to avoid overfitting from forcefully memorizing hard training labels. To further address this issue, we extend the PCL objective by introducing the interpolation training strategy (ITS). By constructing virtual training samples that are linear interpolations of two random samples, the model is forced to predict less confidently on interpolations and produce smoother decision boundaries. Specifically, we first perform convex combinations of instance pairs as:

$$\boldsymbol{x}^{mix} = \eta \boldsymbol{x}_a + (1 - \eta)\boldsymbol{x}_b \qquad (5)$$

where $\eta \in [0,1] \sim \text{Beta}(\alpha, \alpha)$ and $\boldsymbol{x}^{mix}$ denotes the training sample that combines two samples $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, which are randomly chosen from the same minibatch. We then impose a linear relation in the contrastive loss, which is defined as a weighted combination of the two $\mathcal{L}_p$ with respect to class y_a and y_b. It enforces the embedding for the interpolated input to have the same linear relationship with its corresponding prototypes:

$$\mathcal{L}_r = \eta \mathcal{L}_p\left(\boldsymbol{z}^{mix}, \boldsymbol{y}_a\right) + (1 - \eta)\mathcal{L}_p\left(\boldsymbol{z}^{mix}, \boldsymbol{y}_b\right) \qquad (6)$$

where $\boldsymbol{z}^{mix}$ be the normalized embedding for $\boldsymbol{x}^{mix}$, $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$ are the classes of $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$.

3.5. Adaptive Prototypical Dispersing

To ensure that the generated intent representations with adequate between-cluster separation to establish distinct cluster boundaries, we draw in-

spiration from instance-wise contrastive learning:

$$\mathcal{L}_c = \underbrace{-\frac{1}{\tau}\mathbb{E}_{i,j}s\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)}_{\text{Instance Alignment}} + \underbrace{\mathbb{E}_i \log \sum_{k=1}^{2N_b} \mathbb{1}_{[k \neq i]} e^{s(\boldsymbol{z}_i, \boldsymbol{z}_k)}}_{\text{Instance Uniformity}} \qquad (7)$$

where the second term is referred to `uniformity` since it encourages instance representation to be uniformly distributed in the hypersphere. But instance-wise constraints may inevitably lead to the class collision issue (Saunshi et al., 2019).

Derived from Eq. 7, we devise a novel adaptive prototypical dispersing learning (APDL) method to maximize the prototype-to-prototype distance and improve distribution uniformity by extending the instance-wise contrastive loss. To facilitate large angular distances among different class prototypes, APDL utilizes the distances between prototypes as adaptive weights. This imposes stronger penalties on close prototypes and produces well-separated clusters. The APDL loss is given by:

$$\mathcal{L}_a = \underbrace{\mathbb{E}_{i \leq C} \log \frac{\sum_{j=1}^{C} \mathbb{1}_{[j \neq i]} D(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) e^{s(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)}}{C - 1}}_{\text{Prototypical Uniformity}} \qquad (8)$$

where $s(\cdot, \cdot)$ is the cosine similarity to evaluate semantic similarities among prototypes ($s(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \cos(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)/\tau$). It is worth noting that $D(\cdot, \cdot)$ is an adaptive constraint term (ACT) that adaptively maximizes the distance between nearer prototypes by taking the reciprocal of their distances:

$$D(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2} \qquad (9)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ represent prototypes of any two intent classes in the intent representation space.

3.6. Dynamic Prototypes Update

It is crucial to continuously update the class prototypes over the course of training. Although $\boldsymbol{\mu}_k$ can

be calculated by averaging the representations of class k across the entire dataset at the end of an epoch, it means the prototypes will remain static during the next full epoch. This is not ideal, as distributions of intent representations and clusters are rapidly changing, especially in the earlier epochs. So we use the exponential moving average (EMA) algorithm to learn more robust prototypes:

$$\boldsymbol{\mu}_k = \sigma(\lambda\boldsymbol{\mu}_k + (1-\lambda)\boldsymbol{z}_i) \tag{10}$$

where the σ is the layer normalization, λ is a momentum factor and $\boldsymbol{z}_i$ is normalized embeddings.

3.7. Multitask Learning

Our approach learns cluster-friendly intent representations with stronger within-cluster compactness and larger between-cluster separation by jointly performing RPAL and APDL. To mitigate the risk of catastrophic forgetting of knowledge gained from labeled data, we integrate cross-entropy loss into the training process. The multitask learning objective for NID can be denoted as:

$$\mathcal{L}_{all} = \omega\mathcal{L}_r + \mathcal{L}_a + \mathcal{L}_{ce} \tag{11}$$

where ω is a hyperparameter that controls the weight of loss. For inference, we perform the non-parametric clustering method k-means to obtain cluster assignments for testing data.

4. Experiments

4.1. Datasets

To validate the effectiveness and generality of our method, we conducted experiments on three diverse and challenging real-world datasets. Detailed statistics are described in Table 1.

- **CLINC** (Larson et al., 2019) is a dataset specially designed for OOD detection and intent discovery, which contains 22.5K samples of user queries in total and 150 unique labeled intents from 10 domains.

- **BANKING** (Casanueva et al., 2020) is a dataset about banking. The dataset provides user queries and labeled intents from the banking domain, with a total of 13K samples and 77 types of intents.

- **StackOverflow** (Xu et al., 2015) is a dataset published in Kaggle.com. It contains 20K samples across 20 classes of technical questions collected from the Kaggle website.

Dataset	$|\mathcal{I}_k|$	$|\mathcal{I}_n|$	$|\mathcal{D}_s|$	$|\mathcal{D}_u|$	$|\mathcal{D}_t|$
CLINC	113	37	1344	16656	2250
BANKING	58	19	673	8330	3080
StackOverflow	15	5	1350	16650	1000

Table 1: Statistics of datasets. We set the known class ratio $|\mathcal{I}_k|/|\mathcal{I}_k \cap \mathcal{I}_n|$ to 75%. The columns represent the number of known categories, novel categories, labeled data, unlabeled data, and testing data, respectively.

4.2. Baselines

We compare our approach with unsupervised and semi-supervised models:

- **Unsupervised:** k-means (MacQueen et al., 1967), Agglomerative Clustering (AC) (Gowda and Krishna, 1978), SAE-KM and Deep Embedded Cluster (DEC) (Xie et al., 2016), Deep Clustering Network (DCN) (Yang et al., 2017), DAC (Chang et al., 2017), DeepCluster (Caron et al., 2018).

- **Semi-supervised:** PCK-means (Basu et al., 2004), KCL (Hsu et al., 2018), MCL (Hsu et al., 2019), DTC (Han et al., 2019), CDAC+ (Lin et al., 2020), GCD (Vaze et al., 2022), DeepAligned (Zhang et al., 2021a), MTP-CLNN (Zhang et al., 2022), ProbNID (Zhou et al., 2023b), DPN (An et al., 2023).

4.3. Evaluation Metrics

To evaluate the quality of the discovered intent clusters, we use three broadly used evaluation metrics (Zhang et al., 2021a, 2022; Zhou et al., 2023b): (1) Normalized Mutual Information **(NMI)** measures the normalized mutual dependence between the predicted labels and the ground-truth labels. (2) Adjusted Rand Index **(ARI)** measures how many samples are assigned properly to different clusters. (3) Accuracy **(ACC)** is measured by assigning dominant class labels to each cluster and taking the average precision. Higher values of these metrics indicate better performance. Specifically, NMI is defined as:

$$\text{NMI}(\mathbf{y}^{gt}, \mathbf{y}^p) = \frac{MI(\mathbf{y}^{gt}, \mathbf{y}^p)}{\frac{1}{2}(H(\mathbf{y}^{gt}) + H(\mathbf{y}^p))}, \tag{12}$$

where $\mathbf{y}^{gt}$ and $\mathbf{y}^p$ are the ground-truth and predicted labels, respectively. $MI(\mathbf{y}^{gt}, \mathbf{y}^p)$ represents the mutual information between $\mathbf{y}^{gt}$ and $\mathbf{y}^p$, and $H(\cdot)$ is the entropy. $MI(\mathbf{y}^{gt}, \mathbf{y}^p)$ is normalized by the arithmetic mean of $H(\mathbf{y}^{gt})$ and $H(\mathbf{y}^p)$, and the values of NMI are in the range of [0, 1]. ARI is

	Methods	CLINC			BANKING			StackOverflow		
		NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
Unsupervised	K-means	70.89	26.86	45.06	54.57	12.18	29.55	8.24	1.46	13.55
	AC	73.07	27.70	44.03	57.07	13.31	31.58	10.62	2.12	14.66
	SAE-KM	73.13	29.95	46.75	63.79	22.85	38.92	32.62	17.07	34.44
	DEC	74.83	27.46	46.89	67.78	27.21	41.29	10.88	3.76	13.09
	DCN	75.66	31.15	49.29	67.54	26.81	41.99	31.09	15.45	34.56
	DAC	78.40	40.49	55.94	47.35	14.24	27.41	14.71	2.76	16.30
	DeepCluster	65.58	19.11	35.70	41.77	8.95	20.69	-	-	-
Semi-supervised	PCK-Means	68.70	35.40	54.61	48.22	16.24	32.66	17.26	5.35	24.16
	KCL (BERT)	86.82	58.79	68.86	75.21	46.72	60.15	8.84	7.81	13.94
	MCL (BERT)	87.72	59.92	69.66	75.68	47.43	61.14	66.81	57.43	72.07
	CDAC+	86.65	54.33	69.89	72.25	40.97	53.83	69.84	52.59	73.48
	DTC (BERT)	90.54	65.02	74.15	76.55	44.70	56.51	63.17	53.66	71.47
	GCD$^\heartsuit$	91.13	67.44	77.50	77.86	46.87	58.95	64.74	47.70	67.71
	DeepAligned	93.95	80.33	87.29	79.91	54.34	66.59	76.47	62.52	80.26
	MTP-CLNN$^\heartsuit$	94.88	84.77	88.25	84.22	63.10	73.98	77.03	69.50	83.18
	ProbNID	95.01	83.00	88.99	84.02	62.92	74.03	77.32	65.70	80.50
	DPN$^\heartsuit$	95.14	84.30	89.22	84.31	63.26	74.45	79.89	70.27	84.59
	RAP(Ours)	**95.93**	**86.28**	**91.24**	**85.16**	**65.79**	**76.27**	**82.36**	**71.73**	**86.60**

Table 2: Comparison against the unsupervised and semi-supervised baselines on three benchmarks. $\heartsuit$ denotes results obtained from running the provided code and other results are retrieved from Zhou et al. (2023b). Results are averaged over different random seeds and the bold fonts denote the best scores.

defined as:

$$\text{ARI} = \frac{\sum_{i,j} \binom{n_{i,j}}{2} - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{u_i}{2} + \sum_j \binom{v_j}{2}] - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{n}{2}}$$
(13)

where $u_i = \sum_j n_{i,j}$, and $v_j = \sum_i n_{i,j}$. n is the number of samples, and $n_{i,j}$ is the number of the samples that have both the ith predicted label and the jth ground-truth label. The values of ARI are in the range of [-1, 1]. ACC is defined as:

$$\text{ACC}(\mathbf{y}^{gt}, \mathbf{y}^p) = \max_m \frac{\sum_{i=1}^n \mathbb{I}\left\{y_i^{gt} = m\left(y_i^p\right)\right\}}{n}$$
(14)

where m is a one-to-one mapping between the ground-truth label $\mathbf{y}^{gt}$ and predicted label $\mathbf{y}^p$ of the ith sample. The Hungarian algorithm is used to obtain the best mapping m efficiently. The values of ACC are in the range of [0, 1].

4.4. Experimental Settings

In experiments, we utilize the pre-trained 12-layer bert-uncased BERT model[1] (Devlin et al., 2019) as the backbone encoder and only fine-tune the last transformer layer parameters to expedite the training process. For model optimization, we adopt the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning of 1e-5. The wait patience for early stopping is set to 20. For masked language modeling, the mask probability is set to 0.15 following previous work. For MTP-CLNN, the external

---

dataset is not used as in other baselines, the parameter of top-k nearest neighbors is set to $\{100, 50, 500\}$ for CLINC, BANKING, and StackOverflow, respectively, as utilized in Zhang et al. (2022). For all experiments, we split the datasets into train, valid, and test sets, and randomly select 25% of categories as unknown and only 10% of training data as labeled (Zhang et al., 2021a). The number of intent categories is set as ground truth. we set the temperature scale as τ = 0.1 in Eq. (4) and Eq. (8), the parameter in beta distribution α = 1 in Eq. (5) (i.e. η is sampled from a uniform distribution), the momentum factor λ = 0.9 in Eq. (10). All the experiments are conducted on 4 Tesla V100 GPUs and averaged over 10 different seeds.

4.5. Main Results

Table 2 shows the main results on three datasets. It is observed that RAP achieves the overall best performances compared to other baselines across all datasets. The in-depth observations can be derived from the results: (1) Compared with unsupervised methods (Unsupervised), semi-supervised methods (Semi-supervised) achieve much better results, which demonstrates the advantage of prior knowledge transfer for subsequent tasks. (2) Under the semi-supervised setting (Semi-supervised), our method achieves new state-of-the-art results across all datasets and metrics. The effectiveness of our method can be attributed to its ability to efficiently control and coordinate both within-cluster and between-cluster

Methods	CLINC			BANKING			StackOverflow		
	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
RAP	**95.93**	**86.28**	**91.24**	**85.16**	**65.79**	**76.27**	**82.36**	**71.73**	**86.60**
① w/o $\mathcal{L}_{rpal}$	93.77	81.42	87.05	80.82	59.15	71.33	76.06	66.55	80.72
② w/o $\mathcal{L}_{apdl}$	93.13	82.58	88.39	82.37	62.03	72.15	79.15	69.07	83.67
③ w/o $\mathcal{L}_{ce}$	95.02	85.27	90.17	84.49	65.27	75.00	81.18	71.06	85.44
④ w/o ITS	94.93	85.09	90.90	83.72	63.18	74.96	81.77	70.50	85.21
⑤ w/o ACT	95.49	85.70	89.84	84.58	63.91	75.77	82.19	70.44	86.34

Table 3: Experimental results of ablation study on the CLINC, BANKING, and StackOverflow datasets.

distances, even with limited labeled data as prior knowledge. It is helpful to establish distinct global decision boundaries between known and novel intent categories, thereby excelling in NID. We also find our method yields competitive results on imbalanced datasets like BANKING, underscoring the robust generalization capabilities of our model. This indicates that our method is better tailored for real-world NID tasks.

4.6. Ablation Study

To investigate the contribution of each component within RAP, we conduct an ablation study across all datasets in Table 3 and consider five submodules as variants. Removing any component will lead to performance degradation, emphasizing the essence of each independent component. Specifically, (1) RAP (w/o $\mathcal{L}_{rpal}$) refers to removing the robust prototypical attracting learning objective. Experiment ① indicates that our method is adept at minimizing the instance-to-prototype distances, thereby enhancing within-cluster compactness. (2) RAP (w/o $\mathcal{L}_{apdl}$) involves removing adaptive prototypical dispersing learning objective. Experiment ② suggests that augmenting between-cluster dispersion is pivotal for optimizing NID performance. Without explicitly constraining prototype-to-prototype distances, prior methods hinder the model from acquiring cluster-friendly representations. (3) RAP (w/o $\mathcal{L}_{ce}$) indicates the exclusion of the cross-entropy loss term during the joint optimization process for our method. Experiment ③ highlights the importance of $\mathcal{L}_{ce}$ in mitigating catastrophic forgetting of knowledge learned from labeled data. (4) RAP (w/o ITS) refers to removing the interpolation training strategy in Eq. (6), which implies substituting Eq.(4) in place of Eq. (6). Experiment ④ connotes the efficacy of the ITS in mitigating pseudo-label noise. (5) RAP (w/o ACT) refers to removing the adaptive constraint term $\text{dist}(\cdot, \cdot)$ in Eq. (8). Experiment ⑤ verifies the importance of imposing stricter penalties on nearer prototypes to optimize the between-cluster distances.

5. Discussion

5.1. Compactness and Separability

RAP aims to generate cluster-friendly representations with two vital characteristics: Compactness and Separability. We investigate these two properties from the following perspectives.

Stronger within-cluster compactness. To show the capability of our method in promoting tight within-cluster representations, we measure the within-cluster distance by average cosine similarity calculation between each intent embedding and its corresponding prototype in Table 4. We can see that our method achieves a significantly lower within-cluster distance compared to previous leading methods. This phenomenon may be attributed to the pivotal role of RPAL in enhancing within-cluster compactness.

Larger between-cluster dispersion. To further analyze whether our method truly enlarges the distances among prototypes, we compute the mean cosine similarity for all pairs of class prototypes. In Table 4, the proposed RAP consistently obtains larger between-cluster distances compared to previous competitive methods. We speculate that the primary reason for this finding is that APDL plays a crucial role in enlarging the between-cluster dispersion. This also fully conforms to our expectations that the APDL effectively improves the uniformity of the intent representation space.

5.2. Representation Visualization

To further validate the effectiveness of our method in learning discriminative intent representations, we adopt the t-SNE to visualize projected representation on the StackOverflow dataset. Comparing Figure 3(a) and Figure 3(d), we can clearly see that the clusters obtained by our method are generally more compact and well-separated than those obtained by the strong baseline model. This indicates that our model learns cluster-friendly features for NID. Looking at Figure 3(b) and Figure 3(d), it evidently shows that the RPAL effectively pulls instances closer to their corresponding
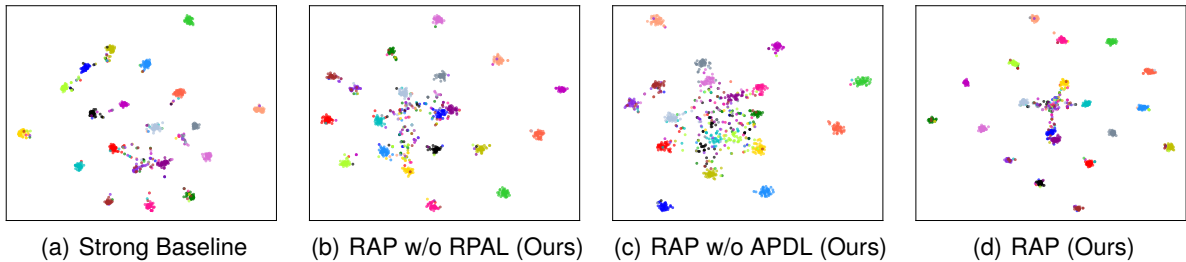
(a) Strong Baseline   (b) RAP w/o RPAL (Ours)   (c) RAP w/o APDL (Ours)   (d) RAP (Ours)

Figure 3: t-SNE visualization of learned representation.

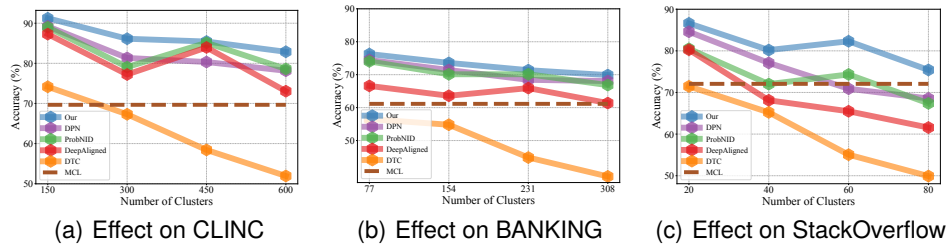(a) Effect on CLINC   (b) Effect on BANKING   (c) Effect on StackOverflow

Figure 4: Sensitivity of the models to the number of initial clusters on three datasets.

Methods	Within ($\downarrow$)	Between ($\uparrow$)
DTC (BERT)	22.40	11.82
DeepAligned	11.75	13.91
DPN	9.03	17.38
Ours (CLINC)	**8.68**	**23.36**
DTC (BERT)	20.31	22.49
DeepAligned	10.28	13.14
DPN	7.96	11.08
Ours (BANKING)	**7.22**	**17.50**
DTC (BERT)	17.11	7.98
DeepAligned	9.73	11.53
DPN	7.30	14.66
Ours (StackOverflow)	**4.07**	**20.81**

Table 4: Statistics of within-cluster and between-cluster distances (Islam et al., 2021).

Methods	CLINC		BANKING		StackOverflow	
	C	E	C	E	C	E
DTC (BERT)	112	25.33	58	24.68	26	30.00
DeepAligned	129	14.00	67	12.99	17	15.00
ProbNID	130	13.30	73	5.48	-	-
DPN	137	8.67	71	7.80	18	10.00
Ours	**141**	**6.00**	**75**	**2.60**	**22**	**10.00**

Table 5: Estimation of the number of clusters C, where E represents the error rate, which is obtained by calculating the estimated C and the ground truth number.

prototypes, achieving strong within-cluster compactness. Moreover, the difference between Figure 3(c) and Figure 3(d) shows that the APDL significantly pushes prototypes away from each other and builds distinct cluster boundaries.

5.3. Effect of the Number of Clusters

To explore the sensitivity of the models to the initial number of clusters C, we adjust C from its ground-truth value up to four times that amount. In Figure 4, we observe that most methods show a performance drop with an increasing initial value of C. This is because the unreasonably assigned C leads to the generation of noisy pseudo-labels, which substantially impacts the clustering results. Notably, our method still achieves the best performance on three datasets, validating the capacity of

our model to mitigate the impact of noisy pseudo-labels and augment robustness. This also shows the superiority of our method in the same semi-supervised NID setting.

5.4. Estimate Number of Clusters

The above experiments assume the number of clusters C to be the ground truth. But this is unrealistic in practice. Therefore, in order to further validate the effectiveness of our method in practical scenarios, we conduct experiments to estimate the number of clusters. We use the same settings as Zhang et al. (2021a) and firstly assign the number of intents as two times the ground truth number to investigate the ability to estimate C. In Table 5, we notice that our method can predict the number of intents more accurately and achieve better results at the same time. The results indicate that our method more easily learns cluster-friendly discriminative representations that assist in accurately estimating the number of clusters.

12200
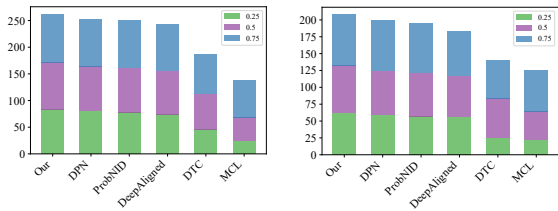
(a) Impact on CLINC (b) Impact on BANKING

Figure 5: Impact of varying the known class ratio on two datasets. The x-axis represents different models and the y-axis denotes their corresponding accuracy values.
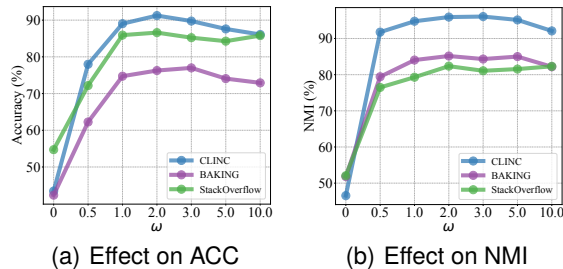
(a) Effect on ACC (b) Effect on NMI

Figure 6: Weight of the multitask learning ω.

5.5. Effect of Known Class Ratio

To investigate the influence of the number of known intents, we vary the known class ratio in the range of 25%, 50% and 75% during training. From Figure 5, it is evident that the performance of all strong methods gradually decreases as the known intent rate decreases. As the known intent rate decreases, there is less labeled data available to guide model training, which complicates the transfer of prior knowledge for discovering new intents. However, with the decrease in the known intent rate, our proposed RAP demonstrates more significant improvements. We surmise that as the number of intent categories increases, the pivotal factor for enhancing performance is the learning of cluster-friendly representations, which establish distinct boundaries for both known and novel categories. This highlights the proficiency of our model in optimizing both within-cluster and between-cluster distances, resulting in well-defined cluster boundaries.

5.6. Weight of Multitask Learning

The weight of the multitask learning ω in Eq. 11 adjusts the contribution of two objectives RPAL and APDL. To pursue the optimal performance, we conduct experiments varying ω across $\{0.0, 0.5, 1.0, 2.0, 3.0, 5.0, 10.0\}$. In Figure 6, the model performance continues improving with the value of ω increasing to 1.0. The model keeps a relatively stable performance after reaching 3.0 but encounters performance degradation with a large value of $\omega = 10.0$. The ability of the RPAL

Methods	KACC	NACC	NMI	ARI
ChatGPT3.5	**84.14**	25.45	73.58	37.88
Ours (CLINC)	68.56	**82.00**	**88.53**	**47.06**
ChatGPT3.5	**80.41**	32.69	87.69	46.68
Ours (BANKING)	71.42	**85.00**	**89.60**	**58.66**
ChatGPT3.5	71.53	50.00	72.20	47.17
Ours (StackOverflow)	**82.35**	**89.41**	**81.50**	**62.13**

Table 6: Comparison between RAP and LLM.

cannot be fully exploited with a small value of ω, while the ability of the APDL is suppressed leading to a worse clustering performance with a large value of ω. Empirically, we choose $\omega = 2.0$ for all datasets.

5.7. Comparison with LLM

To conduct a comprehensive performance comparison with the large language model (LLM) on the NID task, we randomly select 1.5% of training data as labeled and choose 75% of all intents as known. For evaluation, our comprehensive assessment covers 600 instances from three datasets (200 samples randomly from each dataset). The metrics are comprised of known intents accuracy (KACC), novel detection accuracy (NACC), and clustering performance (NMI and ARI) for novel intents. In Table 6, our method consistently outperforms ChatGPT3.5[2] across all datasets and evaluation metrics with a small model size and fast inference speed, demonstrating the superior performance of our approach. Moving forward, we plan to explore the integration of RAP with LLM to boost the performance in NID.

6. Conclusion

In this work, we propose a robust and adaptive prototypical learning (RAP) framework for new intent discovery, which aims to learn cluster-friendly discriminative representations. Specifically, we design the robust prototypical attracting learning (RPAL) method and the adaptive prototypical dispersion (APDL) method to control within-cluster and between-cluster distances, respectively. Experimental results on three benchmarks demonstrate that RAP significantly outperforms the previous unsupervised and semi-supervised baselines and even defeats the large language model (ChatGPT3.5). Extensive probing analysis further verifies that RPAL is helpful for realizing stronger within-cluster compactness while mitigating the effects of noisy pseudo-labels and APDL is beneficial for attaining larger between-cluster dispersion. We hope our work can provide useful insights for further research.

---

[2]We utilize OpenAI *gpt-3.5-turbo-0301*, see index.

Limitations

Despite the promising results obtained by our RAP, it is crucial to acknowledge several limitations: **(1) Improving pseudo-labels assignments.** The pseudo-labels obtained using the k-means method are not sufficiently reliable, as it is highly sensitive to noisy intent data. We plan to explore more reliable pseudo-label assignment approaches. **(2) Leveraging LLMs to facilitate interpretability.** While our clustering method can assign cluster labels to unlabeled utterances, it cannot generate meaningful and interpretable names for each identified cluster or intent. We intend to investigate the combination of our method with LLMs to assign accurate category names to newly discovered intent categories.

Acknowledgements

7. Bibliographical References

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, Qianying Wang, and Ping Chen. 2023. Generalized category discovery with decoupled prototypical network. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12527–12535. AAAI Press.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, 2004*, pages 333–344. SIAM.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 139–156. Springer.

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.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2017. Deep adaptive image clustering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5880–5888. IEEE Computer Society.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. 2019. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8928–8939.

Aleksandra Chrabrowa, Tsimur Hadeliya, Dariusz Kajtoch, Robert Mroczkowski, and Piotr Rybak. 2023. Going beyond research datasets: Novel intent discovery in the industry setting. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 895–911. Association for Computational Linguistics.

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

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. 2021. A unified objective for novel class discovery. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9264–9272. IEEE.

K. Chidananda Gowda and G. Krishna. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognit.*, 10(2):105–112.

Dilek Hakkani-Tür, Asli Celikyilmaz, Larry P. Heck, and Gökhan Tür. 2013. A weakly-supervised approach for discovering new user intents from search query logs. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 3780–3784. ISCA.

Dilek Hakkani-Tür, Yun-Cheng Ju, Geoffrey Zweig, and Gökhan Tür. 2015. Clustering novel intents in a conversational interaction system with semantic parsing. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1854–1858. ISCA.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8400–8408. IEEE.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2018. Learning to cluster in order to transfer across domains and tasks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard J. Radke, and Rogério Feris. 2021. A broad study on the transferability of visual representations with contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8825–8835. IEEE.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4804–4815. PMLR.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1836–1853. Association for Computational Linguistics.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 1311–1316. Association for Computational Linguistics.

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

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8360–8367. AAAI Press.

Bing Liu and Sahisnu Mazumder. 2021. Lifelong and continual learning dialogue systems: Learning during conversation. In *Thirty-Fifth*

AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 15058–15063. AAAI Press.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 14, pages 281–297. Oakland, CA, USA.

Ying Mo, Jian Yang, Jiahao Liu, Qifan Wang, Ruoyu Chen, Jingang Wang, and Zhoujun Li. 2023. mcl-ner: Cross-lingual named entity recognition via multi-view contrastive learning. *arXiv preprint arXiv:2308.09073*.

Yutao Mou, Keqing He, Yanan Wu, Pei Wang, Jingang Wang, Wei Wu, Yi Huang, Junlan Feng, and Weiran Xu. 2022. Generalized intent discovery: Learning from open world dialogue system. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 707–720. International Committee on Computational Linguistics.

Padmasundari and Srinivas Bangalore. 2018. Intent discovery through unsupervised semantic text clustering. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 606–610. ISCA.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. IDAS: intent discovery with abstractive summarization. *CoRR*, abs/2305.19783.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR.

Xiang Shen, Yinge Sun, Yao Zhang, and Mani Najmabadi. 2021. Semi-supervised intent discovery with contrastive learning. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 120–129.

Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng Wang, and Lintao Zhang. 2018. Auto-dialabel: Labeling dialogue data with unsupervised learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 684–689, Brussels, Belgium. Association for Computational Linguistics.

A. B. Siddique, Fuad T. Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1925–1929. ACM.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2022. Generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7482–7491. IEEE.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 808–819. Association for Computational Linguistics.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop*

and Conference Proceedings, pages 478–487. JMLR.org.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, Denver, Colorado. Association for Computational Linguistics.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3861–3870. PMLR.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. CROP: zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of EMNLP 2022*, pages 486–496.

Jian Yang, Shuming Ma, Li Dong, Shaohan Huang, Haoyang Huang, Yuwei Yin, Dongdong Zhang, Liqun Yang, Furu Wei, and Zhoujun Li. 2023. Ganlm: Encoder-decoder pre-training with an auxiliary discriminator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9394–9412, Toronto, Canada. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. 2021a. Multilingual machine translation systems from microsoft for WMT21 shared task. In *WMT 2021*, pages 446–455. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Ming Zhou. 2020. Improving neural machine translation with soft template prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5979–5989.

Jian Yang, Shuming Ma, Dongdong Zhang, Juncheng Wan, Zhoujun Li, and Ming Zhou. 2021b. Smart-start decoding for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3982–3988. Association for Computational Linguistics.

Jian Yang, Juncheng Wan, Shuming Ma, Haoyang Huang, Dongdong Zhang, Yong Yu, Zhoujun Li, and Furu Wei. 2021c. Learning to select relevant knowledge for neural machine translation. In *Natural Language Processing and Chinese Computing - 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13-17, 2021, Proceedings, Part I*, volume 13028 of *Lecture Notes in Computer Science*, pages 79–91. Springer.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021d. Multilingual agreement for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 233–239, Online. Association for Computational Linguistics.

Ze Yang, Wei Wu, Jian Yang, Can Xu, and Zhoujun Li. 2019. Low-resource response generation with template prior. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1886–1897. Association for Computational Linguistics.

Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto L. Sangiovanni-Vincentelli. 2021. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13834–13844. Computer Vision Foundation / IEEE.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021a. Discovering new intents with deep aligned clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14365–14373. AAAI Press.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2023a. USNID: A framework for unsupervised and semi-supervised new intent discovery. *CoRR*, abs/2304.07699.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jian-Guo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip S. Yu. 2021b. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1906–1912. Association for Computational Linguistics.

Shun Zhang, Jiaqi Bai, Tongliang Li, Zhao Yan, and Zhoujun Li. 2023b. Modeling intra-class and inter-class constraints for out-of-domain detection. In *Database Systems for Advanced Applications - 28th International Conference, DASFAA 2023, Tianjin, China, April 17-20, 2023, Proceedings, Part IV*, volume 13946, pages 142–158. Springer.

Shun Zhang, Tongliang Li, Jiaqi Bai, and Zhoujun Li. 2023c. Label-guided contrastive learning for out-of-domain detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2022. New intent discovery with pre-training and contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 256–269. Association for Computational Linguistics.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. 2023a. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4018–4031. Association for Computational Linguistics.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023b. A probabilistic framework for discovering new intents. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3771–3784. Association for Computational Linguistics.