New Intent Discovery with Pre-training and Contrastive Learning

Yuwei Zhang[2*] **Haode Zhang**[1] **Li-Ming Zhan**[1] **Albert Y.S. Lam**[3]
Xiao-Ming Wu[1†]
Department of Computing, The Hong Kong Polytechnic University, Hong Kong S.A.R.[1]
University of California, San Diego[2]
Fano Labs, Hong Kong S.A.R.[3]
zhangyuwei.work@gmail.com
{haode.zhang, lmzhan.zhan}@connect.polyu.edu.hk
csxmwu@comp.polyu.edu.hk, albert@fano.ai

Abstract

New intent discovery aims to uncover novel intent categories from user utterances to expand the set of supported intent classes. It is a critical task for the development and service expansion of a practical dialogue system. Despite its importance, this problem remains under-explored in the literature. Existing approaches typically rely on a large amount of labeled utterances and employ pseudo-labeling methods for representation learning and clustering, which are label-intensive, inefficient, and inaccurate. In this paper, we provide new solutions to two important research questions for new intent discovery: (1) how to learn semantic utterance representations and (2) how to better cluster utterances. Particularly, we first propose a multi-task pre-training strategy to leverage rich unlabeled data along with external labeled data for representation learning. Then, we design a new contrastive loss to exploit self-supervisory signals in unlabeled data for clustering. Extensive experiments on three intent recognition benchmarks demonstrate the high effectiveness of our proposed method, which outperforms state-of-the-art methods by a large margin in both unsupervised and semi-supervised scenarios. The source code will be available at https://github.com/zhang-yu-wei/MTP-CLNN.

1 Introduction

Why Study New Intent Discovery (NID)? Recent years have witnessed the rapid growth of conversational AI applications. To design a natural language understanding system, a set of expected customer intentions are collected beforehand to train an intent recognition model. However, the pre-defined intents cannot fully meet customer needs. This implies the necessity of expanding the intent recognition model by repeatedly integrating new intents discovered from unlabeled user utterances (Fig. 1).

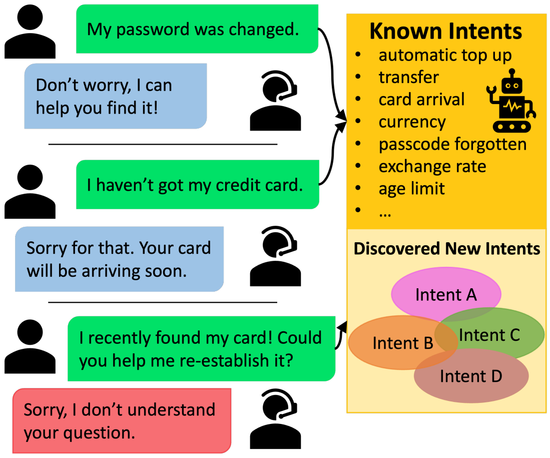

My password was changed.
Don't worry, I can help you find it!
I haven't got my credit card.
Sorry for that. Your card will be arriving soon.
I recently found my card! Could you help me re-establish it?
Sorry, I don't understand your question.
Known Intents
• automatic top up
• transfer
• card arrival
• currency
• passcode forgotten
• exchange rate
• age limit
• ...
Discovered New Intents
Intent A
Intent B
Intent C
Intent D

Figure 1: New Intent Discovery.

To reduce the effort in manually identifying unknown intents from a mass of utterances, previous works commonly employ clustering algorithms to group utterances of similar intents (Cheung and Li, 2012; Hakkani-Tür et al., 2015; Padmasundari, 2018). The cluster assignments thereafter can either be directly used as new intent labels or as heuristics for faster annotations.

Research Questions (RQ) and Challenges. Current study of NID centers around two basic research questions: *1) How to learn semantic utterance representations to provide proper cues for clustering? 2) How to better cluster the utterances?* The study of the two questions are often interwoven in existing research. Utterances can be represented according to different aspects such as the style of language, the related topics, or even the length of sentences. It is important to learn semantic utterance representations to provide proper cues for clustering. Simply applying a vanilla pre-trained language model (PLM) to generate utterance representations is not a viable solution, which leads to poor performance on NID as shown by the experimental results in Section 4.2. Some recent works proposed to use labeled utterances of known intents

[*] Work done while the author was with HK PolyU.
[†] Corresponding author.

for representation learning (Forman et al., 2015; Haponchyk et al., 2018; Lin et al., 2020; Zhang et al., 2021c; Haponchyk and Moschitti, 2021), but they require a substantial amount of known intents and sufficient labeled utterances of each intent, which are not always available especially at the early development stage of a dialogue system. Further, pseudo-labeling approaches are often exploited to generate supervision signals for representation learning and clustering. For example, Lin et al. (2020) finetune a PLM with an utterance similarity prediction task on labeled utterances to guide the training of unlabeled data with pseudo-labels. Zhang et al. (2021c) adopt a deep clustering method (Caron et al., 2018) that uses k-means clustering to produce pseudo-labels. However, pseudo-labels are often noisy and can lead to error propagation.

Our Solutions. In this work, we propose a simple yet effective solution for each research question. **Solution to RQ 1: multi-task pre-training.** We propose a multi-task pre-training strategy that takes advantage of both external data and internal data for representation learning. Specifically, we leverage publicly available, high-quality intent detection datasets, following Zhang et al. (2021d), as well as the provided labeled and unlabeled utterances in the current domain, to fine-tune a pre-trained PLM to learn task-specific utterance representations for NID. The multi-task learning strategy enables knowledge transfer from general intent detection tasks and adaptation to a specific application domain. **Solution to RQ 2: contrastive learning with nearest neighbors.** We propose to use a contrastive loss to produce compact clusters, which is motivated by the recent success of contrastive learning in both computer vision (Bachman et al., 2019; He et al., 2019; Chen et al., 2020; Khosla et al., 2020) and natural language processing (Gunel et al., 2021; Gao et al., 2021; Yan et al., 2021). Contrastive learning usually maximizes the agreement between different views of the same example and minimize that between different examples. However, the commonly used instance discrimination task may push away false negatives and hurts the clustering performance. Inspired by a recent work in computer vision (Van Gansbeke et al., 2020), we introduce neighborhood relationship to customize the contrastive loss for clustering in both unsupervised (i.e., without any labeled utterances of known intents) and semi-supervised scenarios. Intuitively, in a semantic feature space, neighboring utterances should have a similar intent, and pulling together neighboring samples makes clusters more compact.

Our main contributions are three-fold.

- We show that our proposed multi-task pre-training method already leads to large performance gains over state-of-the-art models for both unsupervised and semi-supervised NID.
- We propose a self-supervised clustering method for NID by incorporating neighborhood relationship into the contrastive learning objective, which further boosts performance.
- We conduct extensive experiments and ablation studies on three benchmark datasets to verify the effectiveness of our methods.

2 Related Works

New Intent Discovery. The study of NID is still in an early stage. Pioneering works focus on unsupervised clustering methods. Shi et al. (2018) leveraged auto-encoder to extract features. Perkins and Yang (2019) considered the context of an utterance in a conversation. Chatterjee and Sengupta (2020) proposed to improve density-based models. Some recent works (Haponchyk et al., 2018; Haponchyk and Moschitti, 2021) studied supervised clustering algorithms for intent labeling, yet it can not handle new intents. Another line of works (Forman et al., 2015; Lin et al., 2020; Zhang et al., 2021c) investigated a more practical case where some known intents are provided to support the discovery of unknown intents, which is often referred to as semi-supervised NID.

To tackle semi-supervised NID, Lin et al. (2020) proposed to first perform supervised training on known intents with a sentence similarity task and then use pseudo labeling on unlabeled utterances to learn a better embedding space. Zhang et al. (2021c) proposed to first pre-train on known intents and then perform k-means clustering to assign pseudo labels on unlabeled data for representation learning following Deep Clustering (Caron et al., 2018). They also proposed to align clusters to accelerate the learning of top layers. Another approach is to first classify the utterances as known and unknown and then uncover new intents with the unknown utterances (Vedula et al., 2020; Zhang et al., 2021b). Hence, it relies on accurate classification in the first stage.

In this work, we address NID by proposing a multi-task pre-training method for representation learning and a contrastive learning method for clustering. In contrast to previous methods that rely on ample annotated data in the current domain for pre-training, our method can be used in an unsupervised setting and work well in data-scarce scenarios (Section 4.3).

Pre-training for Intent Recognition. Despite the effectiveness of large-scale pre-trained language models (Radford and Narasimhan, 2018; Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020), the inherent mismatch in linguistic behavior between the pre-training datasets and dialogues encourages the research of continual pre-training on dialogue corpus. Most previous works proposed to pre-train on open domain dialogues in a self-supervised manner (Mehri et al., 2020; Wu et al., 2020; Henderson et al., 2020; Hosseini-Asl et al., 2020). Recently, several works pointed out that pre-training with relavant tasks can be effective for intent recognition. For example, Zhang et al. (2020) formulated intent recognition as a sentence similarity task and pre-trained on natural language inference (NLI) datasets. Vulić et al. (2021); Zhang et al. (2021e) pre-trained with a contrastive loss on intent detection tasks. Our multi-task pre-training method is inspired from Zhang et al. (2021d) which leverages publicly available intent datasets and unlabeled data in the current domain for pre-training to improve the performance of few-shot intent detection. However, we argue that the method is more suitable to be applied for NID due to the natural existence of unlabeled utterances.

Contrastive Representation Learning. Contrastive learning has shown promising results in computer vision (Bachman et al., 2019; Chen et al., 2020; He et al., 2019; Khosla et al., 2020) and gained popularity in natural language processing. Some recent works used unsupervised contrastive learning to learn sentence embeddings (Gao et al., 2021; Yan et al., 2021; Kim et al., 2021; Giorgi et al., 2021). Specifically, Gao et al. (2021); Yan et al. (2021) showed that contrastive loss can avoid an anisotropic embedding space. Kim et al. (2021) proposed a self-guided contrastive training to improve the quality of BERT representations. Giorgi et al. (2021) proposed to pre-train a universal sentence encoder by contrasting a randomly sampled text segment from nearby sentences. Zhang et al. (2021e) demonstrated that self-supervised contrastive pre-training and supervised contrastive fine-tuning can benefit few-shot intent recognition. Zhang et al. (2021a) showed that combining a contrastive loss with a clustering objective can improve short text clustering. Our proposed contrastive loss is tailored for clustering, which encourages utterances with similar semantics to group together and avoids pushing away false negatives as in the conventional contrastive loss.

3 Method

Problem Statement. To develop an intent recognition model, we usually prepare a set of expected intents $\mathcal{C}_k$ along with a few annotated utterances $\mathcal{D}^{\text{labeled}}_{\text{known}} = \{(x_i, y_i)|y_i \in \mathcal{C}_k\}$ for each intent. After deployed, the system will encounter utterances $\mathcal{D}^{\text{unlabeled}} = \{x_i|y_i \in \{\mathcal{C}_k, \mathcal{C}_u\}\}$ from both predefined (known) intents $\mathcal{C}_k$ and unknown intents $\mathcal{C}_u$. The aim of new intent discovery (NID) is to identify the emerging intents $\mathcal{C}_u$ in $\mathcal{D}^{\text{unlabeled}}$. NID can be viewed as a direct extension of out-of-distribution (OOD) detection, where we not only need to identify OOD examples but also discover the underlying clusters. NID is also different from zero-shot learning in that we do not presume access to any kind of class information during training. In this work, we consider both unsupervised and semi-supervised NID, which are distinguished by the existence of $\mathcal{D}^{\text{labeled}}_{\text{known}}$, following Zhang et al. (2021c).

Overview of Our Approach. As shown in Fig. 2, we propose a two-stage framework that addresses the research questions mentioned in Sec. 1. In the first stage, we perform multi-task pre-training (MTP) that jointly optimizes a cross-entropy loss on external labeled data and a self-supervised loss on target unlabeled data (Sec. 3.1). In the second stage, we first mine top-K nearest neighbors of each training instance in the embedding space and then perform contrastive learning with nearest neighbors (CLNN) (Sec. 3.2). After training, we employ a simple non-parametric clustering algorithm to obtain clustering results.

3.1 Stage 1: Multi-task Pre-training (MTP)

We propose a multi-task pre-training objective that combines a classification task on external data from publicly available intent detection datasets and a self-supervised learning task on internal data from the current domain. Different from previous works (Lin et al., 2020; Zhang et al., 2021c), our pre-training method does not rely on annotated data ($\mathcal{D}^{\text{labeled}}_{\text{known}}$) from the current domain and hence can

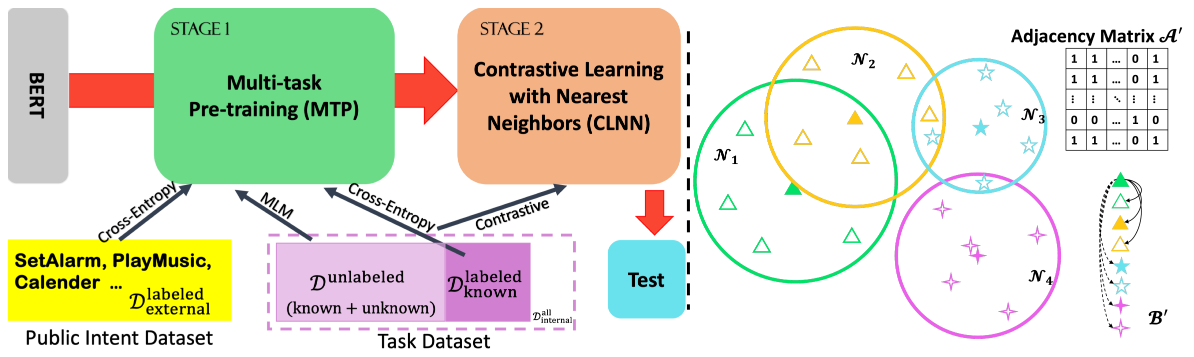

BERT
STAGE 1
Multi-task Pre-training (MTP)
STAGE 2
Contrastive Learning with Nearest Neighbors (CLNN)
Cross-Entropy
MLM
Cross-Entropy
Contrastive
Test
SetAlarm, PlayMusic, Calender ...
$\mathcal{D}^{\text{labeled}}_{\text{external}}$
Public Intent Dataset
$\mathcal{D}^{\text{unlabeled}}$
(known + unknown)
$\mathcal{D}^{\text{labeled}}_{\text{known}}$
$\mathcal{D}^{\text{all}}_{\text{internal}}$
Task Dataset
$\mathcal{N}_1$
$\mathcal{N}_2$
$\mathcal{N}_3$
$\mathcal{N}_4$
Adjacency Matrix $\mathcal{A}'$
1 1 ... 0 1
1 1 ... 0 1
⋮ ⋮ ⋱ ⋮ ⋮
0 0 ... 1 0
1 1 ... 0 1
$\mathcal{B}'$

Figure 2: **The left part shows the overall workflow of our method** where the training order is indicated by the red arrow. The datasets and corresponding loss functions used in each training stage are indicated by the black arrows. **The right part illustrates a simple example of CLNN**. A batch of four training instances $\{x_i\}_{i=1}^4$ (solid markers) and their respective neighborhoods $\{\mathcal{N}_i\}_{i=1}^4$ are plotted (hollow markers within large circles). Since x_2 falls within $\mathcal{N}_1$, x_2 along with its neighbors are taken as positive instance for x_1 (but not vice versa since x_1 is not in $\mathcal{N}_2$). We also show an example of adjacency matrix $\mathbf{A}'$ and augmented batch $\mathcal{B}'$. The pairwise relationships with the first instance in the batch are plotted with solid lines indicating positive pairs and dashed lines indicating negative pairs.

be applied in an unsupervised setting.

Specifically, we first initialize the model with a pre-trained BERT encoder (Devlin et al., 2019). Then, we employ a joint pre-training loss as in Zhang et al. (2021d). The loss consists of a cross-entropy loss on external labeled data and a masked language modelling (MLM) loss on all available data from the current domain:

$$\mathcal{L}_{\text{stg1}} = \underbrace{\mathcal{L}_{\text{ce}}(\mathcal{D}^{\text{labeled}}_{\text{external}};\theta)}_{\text{supervised}} + \underbrace{\mathcal{L}_{\text{mlm}}(\mathcal{D}^{\text{all}}_{\text{internal}};\theta)}_{\text{self-supervised}}, \quad (1)$$

where θ are model parameters. For the supervised classification task, we leverage an external public intent dataset with diverse domains (e.g., CLINC150 (Larson et al., 2019)), denoted as $\mathcal{D}^{\text{labeled}}_{\text{external}}$, following Zhang et al. (2021d). For the self-supervised MLM task, we use all available data (labeled or unlabeled) from the current domain, denoted as $\mathcal{D}^{\text{all}}_{\text{internal}}$.

Intuitively, the classification task aims to learn general knowledge of intent recognition with annotated utterances in external intent datasets, while the self-supervised task learns domain-specific semantics with utterances collected in the current domain. Together, they enable learning semantic utterance representations to provide proper cues for the subsequent clustering task. As will be shown in Sec. 4.3, both tasks are essential for NID.

For semi-supervised NID, we can further utilize the annotated data in the current domain to conduct continual pre-training, by replacing $\mathcal{D}^{\text{labeled}}_{\text{external}}$ in Eq. 1 to $\mathcal{D}^{\text{labeled}}_{\text{known}}$. This step is not included in unsupervised NID.

3.2 Stage 2: Contrastive Learning with Nearest Neighbors (CLNN)

In the second stage, we propose a contrastive learning objective that pulls together neighboring instances and pushes away distant ones in the embedding space to learn compact representations for clustering. Concretely, we first encode the utterances with the pre-trained model from stage 1. Then, for each utterance x_i, we search for its top-K nearest neighbors in the embedding space using inner product as distance metric to form a neighborhood $\mathcal{N}_i$. The utterances in $\mathcal{N}_i$ are supposed to share a similar intent as x_i. During training, we sample a minibatch of utterances $\mathcal{B} = \{x_i\}_{i=1}^M$. For each utterance $x_i \in \mathcal{B}$, we uniformly sample one neighbor x_i' from its neighborhood $\mathcal{N}_i$. We then use data augmentation to generate $\tilde{x}_i$ and $\tilde{x}_i'$ for x_i and x_i' respectively. Here, we treat $\tilde{x}_i$ and $\tilde{x}_i'$ as two views of x_i, which form a positive pair. We then obtain an augmented batch $\mathcal{B}' = \{\tilde{x}_i, \tilde{x}_i'\}_{i=1}^M$ with all the generated samples. To compute contrastive loss, we construct an adjacency matrix $\boldsymbol{A}'$ for $\mathcal{B}'$, which is a $2M \times 2M$ binary matrix where 1 indicates positive relation (either being neighbors or having the same intent label in semi-supervised NID) and 0 indicates negative relation. Hence, we can write the contrastive loss as:

$$l_i = -\frac{1}{|\mathcal{C}_i|}\sum_{j\in\mathcal{C}_i}\log\frac{\exp(\text{sim}(\tilde{h}_i,\tilde{h}_j)/\tau)}{\sum_{k\neq i}^{2M}\exp(\text{sim}(\tilde{h}_i,\tilde{h}_k)/\tau)}, \quad (2)$$

$$\mathcal{L}_{\text{stg2}} = \frac{1}{2M}\sum_{i=1}^{2M} l_i, \quad (3)$$

where $\mathcal{C}_i \equiv \{\boldsymbol{A}'_{i,j} = 1 | j \in \{1, ..., 2M\}\}$ denotes the set of instances having positive relation with $\tilde{x}_i$ and $|\mathcal{C}_i|$ is the cardinality. $\tilde{h}_i$ is the embedding for utterance $\tilde{x}_i$. τ is the temperature parameter. $\text{sim}(\cdot, \cdot)$ is a similarity function (e.g., dot product) on a pair of normalized feature vectors. During training, the neighborhood will be updated every few epochs. We implement the contrastive loss following Khosla et al. (2020).

Notice that the main difference between Eq. 2 and conventional contrastive loss is how we construct the set of positive instances $\mathcal{C}_i$. Conventional contrastive loss can be regarded as a special case of Eq. 2 with neighborhood size $K = 0$ and the same instance is augmented twice to form a positive pair (Chen et al., 2020). After contrastive learning, a non-parametric clustering algorithm such as k-means can be applied to obtain cluster assignments.

Data Augmentation. Strong data augmentation has been shown to be beneficial in contrastive learning (Chen et al., 2020). We find that it is inefficient to directly apply existing data augmentation methods such as EDA (Wei and Zou, 2019), which are designed for general sentence embedding. We observe that the intent of an utterance can be expressed by only a small subset of words such as "suggest restaurant" or "book a flight". While it is hard to identify the keywords for an unlabeled utterance, randomly replacing a small amount of tokens in it with some random tokens from the library will not affect intent semantics much. This approach works well in our experiments (See Table 5 RTR).

Advantages of CLNN. By introducing the notion of neighborhood relationship in contrastive learning, CLNN can 1) pull together similar instances and push away dissimilar ones to obtain more compact clusters; 2) utilize proximity in the embedding space rather than assigning noisy pseudo-labels (Van Gansbeke et al., 2020); 3) directly optimize in the feature space rather than clustering logits as in Van Gansbeke et al. (2020), which has been proven to be more effective by Rebuffi et al. (2020); and 4) naturally incorporate known intents with the adjacency matrix.

4 Experiment

4.1 Experimental Details

Datasets. We evaluate our proposed method on three popular intent recognition benchmarks. **BANKING** (Casanueva et al., 2020) is a fine-grained dataset with 77 intents collected from banking dialogues, **StackOverflow** (Xu et al., 2015) is a large scale dataset collected from online queries, **M-CID** (Arora et al., 2020) is a smaller dataset collected for Covid-19 services. We choose **CLINC150** (Larson et al., 2019) as our external public intent dataset in stage 1 due to its high-quality annotations and coverage of diverse domains. The dataset statistics are summarized in Table 1. We use the same splits of BANKING and StackOverflow as in Zhang et al. (2021b). Details about dataset splitting are provided in the Appendix.

Dataset	domain	#Intents	#Utterances
CLINC150	general	120	18,000
BANKING	banking	77	13,083
StackOverflow	questions	20	20,000
M-CID	covid-19	16	1,745

Table 1: Dataset statistics.

Experimental Setup. We evaluate our proposed method on both unsupervised and semi-supervised NID. *Notice that in unsupervised NID, no labeled utterances from the current domain are provided.* For clarity, we define two variables. The proportion of known intents is defined as $|\mathcal{C}_k|/(|\mathcal{C}_k|+|\mathcal{C}_u|)$ and referred to as "**known class ratio (KCR)**", and the proportion of labeled examples for each known intent is denoted as "**labeled ratio (LAR)**". The labeled data are randomly sampled from the original training set. Notice that, KCR $= 0$ means unsupervised NID, and KCR > 0 means semi-supervised NID. In the following sections, we provide experimental results for both unsupervised NID and semi-supervised NID with KCR $= \{25\%, 50\%, 75\%\}$ and LAR $= \{10\%, 50\%\}$.

Evaluation Metrics. We adopt three popular evaluation metrics for clustering: normalized mutual information (NMI), adjusted rand index (ARI), and accuracy (ACC).

Baselines and Model Variants. We summarize the baselines compared in our experiments for both unsupervised and semi-supervised NID. Our implementation is based on Zhang et al. (2021b).[1]

- **Unsupervised baselines.** (1) GloVe-KM and (2) GloVe-AG are based on GloVe (Pennington et al., 2014) embeddings and then

[1] For fair comparison, the baselines are re-run with TEXTOIR: `https://github.com/thuiar/TEXTOIR`, and hence some results are different from those reported in Lin et al. (2020); Zhang et al. (2021c).

	Methods	BANKING NMI	 ARI	 ACC	StackOverflow NMI	 ARI	 ACC	M-CID NMI	 ARI	 ACC
unsupervised	GloVe-KM	48.75	12.74	27.92	21.79	4.54	24.26	46.40	35.57	46.99
	GloVe-AG	52.76	14.41	31.18	23.45	4.85	24.48	51.23	32.57	42.35
	SAE-KM	60.12	24.00	37.38	48.72	23.36	37.16	51.03	43.51	52.95
	SAE-DEC	62.92	25.68	39.35	61.32	21.17	57.09	50.69	44.52	53.07
	SAE-DCN	62.94	25.69	39.36	61.34	34.98	57.09	50.69	44.52	53.07
	BERT-KM	36.38	5.38	16.27	11.60	1.60	13.85	37.37	14.02	33.81
	MTP **(Ours)**	77.32	47.33	57.99	63.85	48.71	66.18	72.40	53.04	68.94
	MTP-CLNN **(Ours)**	**81.80**	**55.75**	**65.90**	**78.71**	**67.63**	**81.43**	**79.95**	**66.71**	**79.14**

Table 2: Performance on unsupervised NID. For each dataset, the best results are marked in bold.

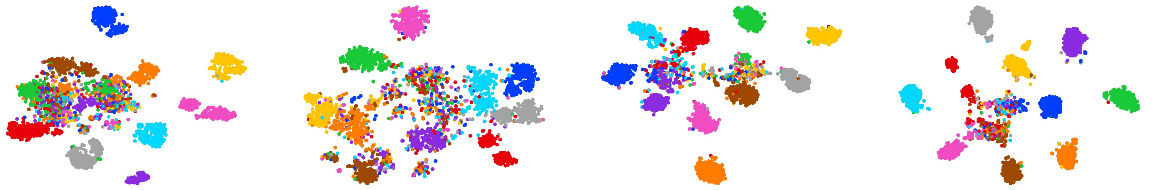
KCR	Methods	BANKING NMI	 ARI	 ACC	StackOverflow NMI	 ARI	 ACC	M-CID NMI	 ARI	 ACC
25%	BERT-DTC	56.05	20.19	32.91	33.38	16.45	30.32	36.00	13.64	27.51
	BERT-KCL	53.85	20.07	28.79	35.47	16.80	32.88	29.35	11.58	24.76
	BERT-MCL	49.46	15.51	24.53	29.44	14.99	31.50	31.16	11.30	26.13
	CDAC+	67.65	34.88	48.79	74.33	39.44	74.30	43.89	19.65	39.37
	DAC	69.85	37.16	49.67	53.97	36.46	53.96	49.83	27.21	43.72
	MTP **(Ours)**	80.00	51.86	62.75	73.75	61.06	75.98	72.40	53.04	68.94
	MTP-DAC **(Comb)**	81.48	55.64	66.12	77.22	61.42	78.60	77.79	62.88	77.02
	MTP-CLNN **(Ours)**	**84.11**	**61.29**	**71.43**	**79.68**	**70.17**	**83.77**	**80.24**	**66.77**	**79.20**
50%	BERT-DTC	69.68	35.98	48.87	53.94	36.79	51.78	51.90	28.94	44.70
	BERT-KCL	62.86	30.16	40.81	57.63	41.90	56.58	42.48	22.83	38.11
	BERT-MCL	62.50	29.80	42.28	49.49	35.96	53.16	41.50	21.46	37.99
	CDAC+	70.62	38.61	51.97	76.18	41.92	76.30	50.47	26.01	46.65
	DAC	76.41	47.28	59.32	70.78	56.44	73.76	63.27	43.52	57.19
	MTP **(Ours)**	82.92	58.46	68.29	77.11	66.45	79.28	72.40	53.04	68.94
	MTP-DAC **(Comb)**	83.43	59.78	70.42	78.91	67.37	81.27	78.17	63.41	77.68
	MTP-CLNN **(Ours)**	**85.62**	**64.93**	**75.23**	**81.03**	**73.02**	**85.64**	**79.48**	**65.71**	**77.85**
75%	BERT-DTC	74.51	44.57	57.34	67.02	55.14	71.14	60.82	38.62	55.42
	BERT-KCL	72.18	44.29	58.70	70.38	57.98	71.50	54.22	34.60	52.15
	BERT-MCL	74.41	48.08	61.57	67.72	55.78	70.82	51.33	31.22	50.77
	CDAC+	71.76	40.68	53.46	76.68	43.97	75.34	55.06	32.52	53.70
	DAC	79.99	54.57	65.87	75.31	60.02	78.84	71.41	54.22	69.11
	MTP **(Ours)**	85.17	64.37	74.20	80.70	71.68	83.74	80.95	69.27	80.92
	MTP-DAC **(Comb)**	85.78	65.28	75.43	80.89	71.17	84.20	80.94	68.27	80.89
	MTP-CLNN **(Ours)**	**87.52**	**70.00**	**79.74**	**82.56**	**75.66**	**87.63**	**83.75**	**73.22**	**84.36**

Table 3: Performance on semi-supervised NID with different known class ratio. The LAR is set to 10%. For each dataset, the best results are marked in bold. **Comb** denotes the baseline method combined with our proposed MTP.

evaluated with k-means (MacQueen et al., 1967) or agglomerative clustering (Gowda, 1984) respectively. (3) BERT-KM applies k-means on BERT embeddings. (4) SAE-KM adopts k-means on embeddings of stacked auto-encoder. (5) Deep Embedding Clustering (SAE-DEC) (Xie et al., 2016) and (6) Deep Clustering Network (SAE-DCN) (Yang et al., 2017) are unsupervised clustering methods based on stacked auto-encoder.

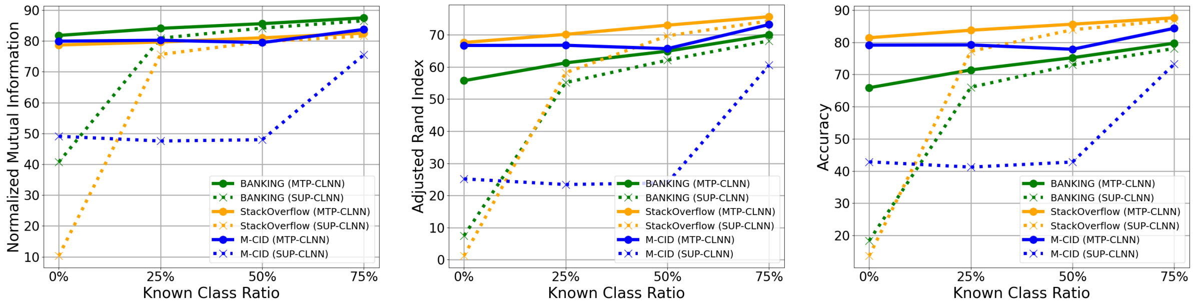
- **Semi-supervised baselines.** (1) BERT-KCL (Hsu et al., 2018) and (2) BERT-MCL (Hsu et al., 2019) employs pairwise similarity task for semi-supervised clustering. (3) BERT-DTC (Han et al., 2019) extends DEC into semi-supervised scenario. (4) CDAC+ (Lin et al., 2020) employs a pseudo-labeling process. (5) Deep Aligned Clustering (DAC) (Zhang et al., 2021c) improves Deep Clustering (Caron et al., 2018) by aligning clusters between iterations.

- Our model variants include MTP and MTP-CLNN, which correspond to applying k-means on utterance representations learned



(a) CDAC+ (b) DAC (c) MTP (Ours) (d) MTP-CLNN (Ours)

Figure 3: Visulization of embeddings on StackOverflow. KCR = 25%, LAR = 10%. Best viewed in color.



(a) (b) (c)

Figure 4: Ablation study on the effectiveness of MTP. The LAR is set to 10%. SUP stands for supervised pre-training on internal labeled data only. The three columns correspond to results in the three metrics respectively.

in stage 1 and stage 2 respectively. Further, we continue to train a DAC model on top of MTP to form a stronger baseline MTP-DAC for semi-supervised NID.

Implementation. We take pre-trained *bert-base-uncased* model from Wolf et al. (2019)[2] as our base model and we use the *[CLS]* token as the BERT representation. For MTP, we first train until convergence on the external dataset, and then when training on $D^{\text{labeled}}_{\text{known}}$, we use a development set to validate early-stopping with a patience of 20 epochs following Zhang et al. (2021c). For contrastive learning, we project a 768-d BERT embedding to an 128-d vector with a two-layer MLP and set the temperature as 0.07. For mining nearest neighbors, we use the inner product method provided by Johnson et al. (2017)[3]. We set neighborhood size $K = 50$ for BANKING and M-CID, and $K = 500$ for StackOverflow, since we empirically find that the optimal K should be roughly half of the average size of the training set for each class (see Section 4.4). The neighborhood is updated every 5 epochs. For data augmentation, the random token replacement probability is set to 0.25. For model optimization, we use the AdamW provided by Wolf et al. (2019). In stage 1, the learning rate is set to $5e^{-5}$. In stage 2, the learning rate is set to $1e^{-5}$ for BANKING and M-CID, and $1e^{-6}$ for StackOverflow. The batch sizes are chosen based on available GPU memory. All the experiments are conducted on a single RTX-3090 and averaged over 10 different seeds. More details are provided in the Appendix.

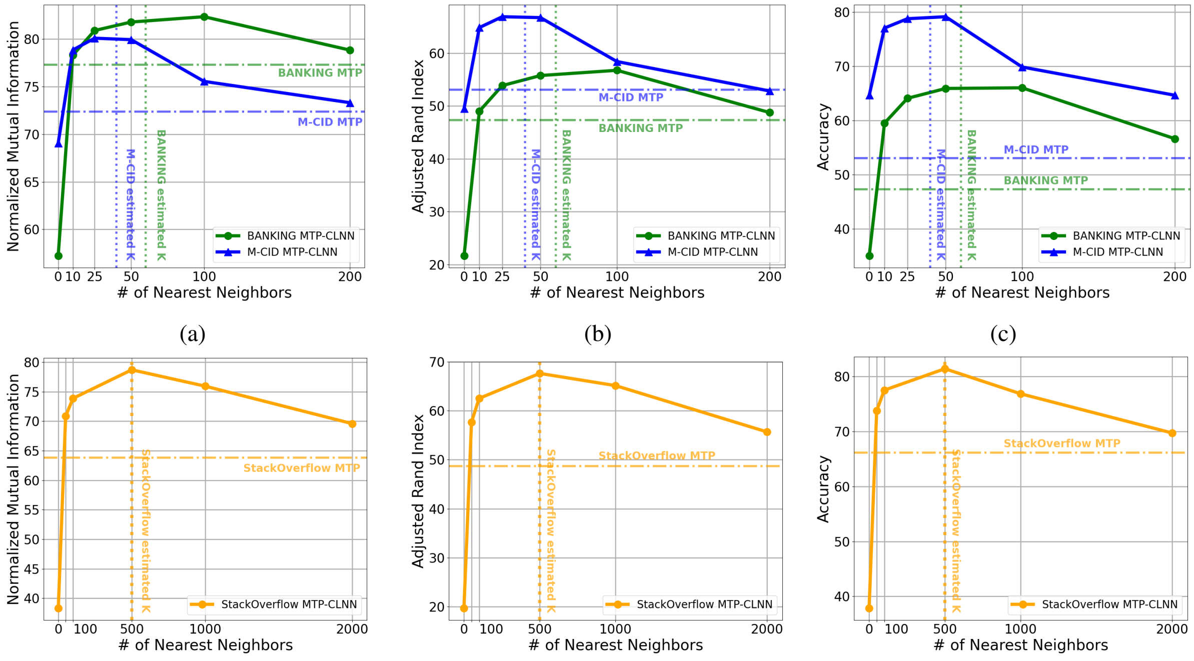
4.2 Result Analysis

Unsupervised NID. We show the results for unsupervised NID in Table 2. First, comparing the performance of BERT-KM with GloVe-KM and SAE-KM, we observe that BERT embedding performs worse on NID even though it achieves better performance on NLP benchmarks such as GLUE, which manifests learning task-specific knowledge is important for NID. Second, our proposed pre-training method MTP improves upon baselines by a large margin. Take the NMI score of BANKING for example, MTP outperforms the strongest baseline SAE-DCN by 14.38%, which demonstrates the effectiveness of exploiting both external public datasets and unlabeled internal utterances. Furthermore, MTP-CLNN improves upon MTP by around 5% in NMI, 10% in ARI, and 10% in ACC across different datasets.

Semi-supervised NID. The results for semi-supervised NID are shown in Table 3. First, MTP

[2] https://github.com/huggingface/transformers

[3] https://github.com/facebookresearch/faiss



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5: Analysis on the number of nearest neighbors in CLNN for unsupervised NID. Vertical dashed lines correspond to our empirical estimations of optima. Horizontal dashed lines represent the results of only training with MTP. When the number of nearest neighbors is 0, we simply augment the same instance twice as in conventional contrastive learning (Chen et al., 2020). The three columns correspond to results in the three metrics respectively.

significantly outperforms the strongest baseline DAC in all settings. For instance, on M-CID, MTP achieves 22.57% improvement over DAC in NMI. Moreover, MTP is less sensitive to the proportion of labeled classes. From KCR $= 75\%$ to KCR $= 25\%$ on M-CID, MTP only drops 8.55% in NMI, as opposed to about 21.58% for DAC. The less performance decrease indicates that our pre-training method is much more label-efficient. Furthermore, with our proposed contrastive learning, MTP-CLNN consistently outperforms MTP and the combined baseline MTP-DAC. Take BANKING with KCR $= 25\%$ for example, MTP-CLNN improves upon MTP by 4.11% in NMI while surpassing MTP-DAC by 2.63%. A similar trend can be observed when LAR $= 50\%$, and we provide the results in the Appendix.

Visualization. In Fig. 3, we show the t-SNE visualization of clusters with embeddings learned by two strongest baselines and our methods. It clearly shows the advantage of our methods, which can produce more compact clusters. Results on other datasets can be found in the Appendix.

4.3 Ablation Study of MTP

To further illustrate the effectiveness of MTP, we conduct two ablation studies in this section. First, we compare MTP with the pre-training method employed in Zhang et al. (2021c), where only internal labeled data are utilized for supervised pre-training (denoted as SUP).[4] In Fig. 4, we show the results of both pre-training methods combined with CLNN with different proportions of known classes. Notice that when KCR $= 0$ there is no pre-training at all for SUP-CLNN. It can be seen that MTP-CLNN consistently outperforms SUP-CLNN. Furthermore, the performance gap increases while KCR decreases, and the largest gap is achieved when KCR $= 0$. This shows the high effectiveness of our method in data-scarce scenarios.

Second, we decompose MTP into two parts: supervised pre-training on external public data (PUB) and self-supervised pre-training on internal unlabeled data (MLM). We report the results of the two pre-training methods combined with CLNN as well as MTP in Table 4. We can easily conclude that either PUB or MLM is indispensable and multi-task

[4] Notice that we make a simple modification to their pre-training to optimize the entire model rather than the last few layers for fair comparison.

	BANKING		StackOverflow		M-CID	
Methods	NMI	ARI	NMI	ARI	NMI	ARI
PUB-CLNN	75.69	44.58	42.22	24.77	73.64	56.94
MLM-CLNN	77.02	47.79	78.62	**68.77**	71.28	53.28
MTP-CLNN	**81.80**	**55.75**	**78.71**	67.63	**79.95**	**66.71**

Table 4: Ablation study of MTP for unsupervised NID.

pre-training is beneficial.

4.4 Analysis of CLNN

Number of Nearest Neighbors. We conduct an ablation study on neighborhood size K in Fig. 5. We can make two main observations. First, although the performance of MTP-CLNN varies with different K, it still significantly outperforms MTP (dashed horizontal line) for a wide range of K. For example, MTP-CLNN is still better than MTP when $K = 50$ on StackOverflow or $K = 200$ on BANKING. Second, despite the difficulty to search for K with only unlabeled data, we empirically find an effective estimation method, i.e. *to choose K as half of the average size of the training set for each class*[5]. It can be seen that the estimated $K \approx 60$ on BANKING and $K \approx 40$ on M-CID (vertical dashed lines) lie in the optimal regions, which shows the effectiveness of our empirical estimation method.

Exploration of Data Augmentation. We compare Random Token Replacement (RTR) used in our experiments with other methods. For instance, dropout is applied on embeddings to provide data augmentation in Gao et al. (2021), randomly shuffling the order of input tokens is proven to be effective in Yan et al. (2021), and EDA (Wei and Zou, 2019) is often applied in text classification. Furthermore, we compare with a Stop-words Replacement (SWR) variant that only replaces the stop-words with other random stop-words so it minimally affects the intents of utterances. The results in Table 5 demonstrate that (1) RTR and SWR consistently outperform others, which verifies our hypothesis in Section 3.2. (2) Surprisingly, RTR and SWR perform on par with each other. For simplicity, we only report the results with RTR in the main experiments.

	BANKING		StackOverflow		M-CID	
Methods	NMI	ARI	NMI	ARI	NMI	ARI
dropout	79.52	50.83	75.60	57.67	79.64	66.14
shuffle	79.02	49.72	75.70	58.95	79.68	66.09
EDA	78.29	49.02	71.50	49.80	79.73	66.39
SWR(Ours)	82.03	56.18	78.48	67.15	79.23	65.74
RTR(Ours)*	81.80	55.75	78.71	67.63	79.95	66.71

Table 5: Ablation study on data augmentation for unsupervised NID. * is the method used in the main results.

5 Conclusion

We have provided simple and effective solutions for two fundamental research questions for new intent discovery (NID): (1) how to learn better utterance representations to provide proper cues for clustering and (2) how to better cluster utterances in the representation space. In the first stage, we use a multi-task pre-training strategy to exploit both external and internal data for representation learning. In the second stage, we perform contrastive learning with mined nearest neighbors to exploit self-supervisory signals in the representation space. Extensive experiments on three intent recognition benchmarks show that our approach can significantly improve the performance of NID in both unsupervised and semi-supervised scenarios.

There are two limitations of this work. (1) We have only evaluated on balanced data. However, in real-world applications, most datasets are highly imbalanced. (2) The discovered clusters lack interpretability. Our clustering method can only assign a cluster label to each unlabeled utterance but cannot generate a valid intent name for each cluster.

6 Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This research was supported by the grants of HK ITF UIM/377 and PolyU DaSAIL project P0030935 funded by RGC.

[5] We presume prior knowledge of the number of clusters. There are some off-the-shelf methods that can be directly applied in the embedding space to determine the optimal number of clusters (Zhang et al., 2021c).

References

Abhinav Arora, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly. 2020. Cross-lingual transfer learning for intent detection of covid-19 utterances.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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.

Ajay Chatterjee and Shubhashis Sengupta. 2020. Intent mining from past conversations for conversational agent. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4140–4152, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

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*, pages 383–392.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Forman, Hila Nachlieli, and Renato Keshet. 2015. Clustering by intent: A semi-supervised method to discover relevant clusters incrementally. In *Machine Learning and Knowledge Discovery in Databases*, pages 20–36, Cham. Springer International Publishing.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

K Chidananda Gowda. 1984. A feature reduction and unsupervised classification algorithm for multispectral data. *Pattern recognition*, 17(6):667–676.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Dilek Hakkani-Tür, Yun-Cheng Ju, Geoffrey Zweig, and Gokhan Tur. 2015. Clustering novel intents in a conversational interaction system with semantic parsing. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Iryna Haponchyk and Alessandro Moschitti. 2021. Supervised neural clustering via latent structured output learning: Application to question intents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3364–3374, Online. Association for Computational Linguistics.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2018. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations (ICLR)*.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *International Conference on Learning Representations (ICLR)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online. 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)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

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.

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*, volume 1, pages 281–297. Oakland, CA, USA.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *ArXiv preprint*, abs/2009.13570.

Srinivas Bangalore Padmasundari. 2018. Intent discovery through unsupervised semantic text clustering. *Proc. Interspeech 2018*, pages 606–610.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. 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 4016–4025, Hong Kong, China. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2020. Lsd-c: Linearly separable deep clusters. *arXiv*.

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.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*.

Nikhita Vedula, Rahul Gupta, Aman Alok, and Mukund Sridhar. 2020. Automatic discovery of novel intents & domains from text utterances.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

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.

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.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. 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*, volume 70 of *Proceedings of Machine Learning Research*, pages 3861–3870. PMLR.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021a. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430, Online. Association for Computational Linguistics.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021b. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 167–174, Online. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021c. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373.

Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert Y. S. Lam. 2021d. Effectiveness of pre-training for few-shot intent classification.

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

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, Online. Association for Computational Linguistics.

A Experimental Details

A.1 Datasets

In this section, we provide more details about the datasets. The development sets are prepared to exclude no unknown intents.

- **BANKING** (Casanueva et al., 2020) is a fine-grained intent detection dataset in which 77 intents are collected for banking dialogue system. The dataset is splitted into 9,003, 1,000 and 3,080 for training, validation, and test sets respectively.

- **StackOverflow** (Xu et al., 2015) is a large scale dataset for online questioning which contains 20 intents with 1,000 examples in each class. We split the dataset into 18,000 for training, 1,000 for validation, and 1,000 for test.

- **M-CID** (Arora et al., 2020) is a small scale dataset for cross-lingual Covid-19 queries. We only use the English subset of this dataset which has 16 intents. We split the dataset into 1,220 for training, 176 for validation, and 349 for test.

- **CLINC150** (Larson et al., 2019) consists of 10 domains across multiple unique services. We use 8 domains [6] and remove the out-of-scope data. We only use this dataset during training stage 1.

A.2 Implementation

The batch size is set to 64 for stage 1 and 128 for stage 2 in all experiments to fully utilize the GPU memory. In stage 1, we first train until convergence on external data and then train with validation on internal data. In stage 2, we train until convergence without early-stopping.

A.3 More Experimental Results

The results on semi-supervised NID when $LAR = 50\%$ are shown in Table 6. It can be seen that our methods still achieve the best performance in this case. In Fig. 6 and Fig. 7, we show the t-SNE visualization of clusters on BANKING and M-CID with embeddings learned by two strongest baselines and our methods. Again, it shows that our methods can produce more compact clusters.

[6] The domains “Banking” and “Credit Cards” are excluded because they are semantically close to the evaluation data.

		BANKING			StackOverflow			M-CID		
KCR	Methods	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
25%	BERT-DTC	66.66	32.47	45.03	36.66	20.85	35.43	40.52	17.49	32.21
	BERT-KCL	56.88	19.40	23.28	35.26	16.18	30.86	31.81	11.13	23.18
	BERT-MCL	52.34	17.30	24.82	34.30	19.12	34.68	31.28	11.69	26.79
	CDAC+	71.71	39.60	53.25	73.92	39.29	72.76	32.22	12.98	30.49
	DAC	73.89	42.84	55.01	56.80	39.51	55.33	53.72	31.36	47.36
	MTP (**Ours**)	80.94	53.44	63.68	75.30	63.27	76.97	73.77	55.37	69.00
	MTP-DAC (**Comb**)	83.05	58.36	68.17	78.15	64.64	79.25	78.36	64.33	77.36
	MTP-CLNN (**Ours**)	**85.30**	**64.12**	**73.76**	**80.15**	**71.29**	**84.56**	**80.61**	**67.31**	**79.34**
50%	BERT-DTC	76.32	48.04	60.35	55.76	39.46	55.58	57.62	35.66	50.40
	BERT-KCL	70.60	37.47	45.37	55.82	37.29	48.32	53.88	32.02	44.76
	BERT-MCL	69.52	37.04	45.65	53.75	38.58	52.04	48.66	28.38	45.67
	CDAC+	76.15	47.01	59.31	74.87	39.38	74.38	51.58	28.96	48.08
	DAC	79.89	54.09	65.14	72.51	59.12	74.51	68.21	49.59	63.84
	MTP (**Ours**)	85.03	62.97	72.63	79.58	70.49	83.51	78.53	63.93	76.50
	MTP-DAC (**Comb**)	86.78	67.23	76.48	81.36	72.58	85.18	81.42	69.36	81.98
	MTP-CLNN (**Ours**)	**88.09**	**71.07**	**80.38**	**82.84**	**75.54**	**87.21**	**82.46**	**71.21**	**82.75**
75%	BERT-DTC	81.60	58.00	69.47	69.61	57.10	72.56	69.57	51.19	66.39
	BERT-KCL	81.23	57.94	68.33	66.85	53.02	66.14	68.46	50.21	64.67
	BERT-MCL	80.98	57.72	68.46	68.39	54.40	65.98	64.09	45.82	63.21
	CDAC+	77.76	49.59	61.48	76.09	41.37	76.64	65.72	39.95	63.87
	DAC	84.78	64.25	74.09	77.67	65.45	81.45	77.37	63.84	77.05
	MTP (**Ours**)	88.69	72.65	81.24	83.27	76.19	87.12	83.53	73.45	83.87
	MTP-DAC (**Comb**)	89.52	74.57	82.97	83.97	77.24	87.91	84.89	75.84	86.59
	MTP-CLNN (**Ours**)	**90.51**	**77.55**	**85.18**	**84.46**	**78.39**	**88.98**	**85.78**	**77.40**	**87.74**

Table 6: Performance on semi-supervised NID with different known class ratio. The LAR is set to 50%. For each dataset, the best results are marked in bold. **Comb** denotes the baseline method combined with our proposed MTP.

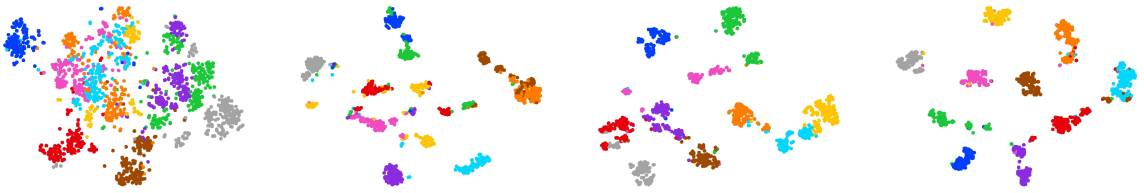

(a) CDAC+ (b) DAC (c) MTP (Ours) (d) MTP-CLNN (Ours)

Figure 6: t-SNE visulization of embeddings on BANKING. KCR $= 25\%$, LAR $= 10\%$. Best viewed in color.

		BANKING			StackOverflow			M-CID		
KCR	Methods	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
0%	SUP-CLNN	40.74	7.50	18.39	10.35	1.27	13.70	49.08	25.16	42.87
	MTP-CLNN	81.80	55.75	65.90	78.71	67.63	81.43	79.95	66.71	79.14
25%	SUP-CLNN	80.91	55.15	66.05	75.65	58.23	77.22	47.56	23.42	41.23
	MTP-CLNN	84.11	61.29	71.43	79.68	70.17	83.77	80.24	66.77	79.20
50%	SUP-CLNN	84.19	62.13	73.01	79.77	69.70	83.91	47.97	23.91	42.81
	MTP-CLNN	85.62	64.93	75.23	81.03	73.02	85.64	79.48	77.65	77.85
75%	SUP-CLNN	86.56	68.15	78.10	81.61	74.35	86.94	75.55	60.56	73.24
	MTP-CLNN	87.52	70.00	79.74	82.56	75.66	87.63	83.75	73.22	84.36

Table 7: Ablation study on the effectiveness of MTP. The LAR is set to 10%. SUP stands for supervised pre-training on internal labeled data only.

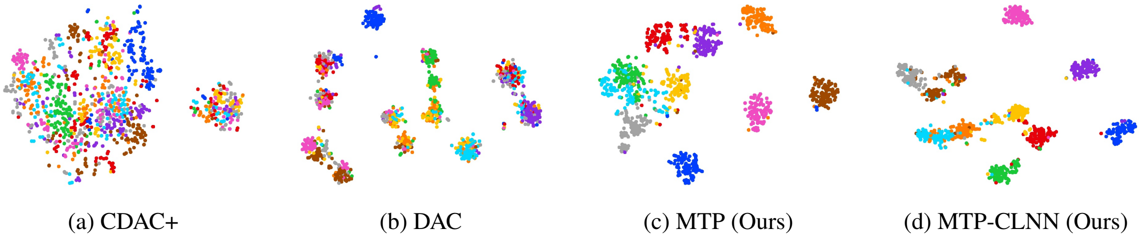


Figure 7: t-SNE visulization of embeddings on M-CID. KCR $= 25\%$, LAR $= 10\%$. Best viewed in color.