A Clustering Framework for Unsupervised and Semi-supervised New Intent Discovery

Hanlei Zhang, Hua Xu, *Member, IEEE,* Xin Wang, Fei Long, Kai Gao

Abstract—New intent discovery is of great value to natural language processing, allowing for a better understanding of user needs and providing friendly services. However, most existing methods struggle to capture the complicated semantics of discrete text representations when limited or no prior knowledge of labeled data is available. To tackle this problem, we propose a novel clustering framework, USNID, for unsupervised and semi-supervised new intent discovery, which has three key technologies. First, it fully utilizes unsupervised or semi-supervised data to mine shallow semantic similarity relations and provide well-initialized representations for clustering. Second, it designs a centroid-guided clustering mechanism to address the issue of cluster allocation inconsistency and provide high-quality self-supervised targets for representation learning. Third, it captures high-level semantics in unsupervised or semi-supervised data to discover fine-grained intent-wise clusters by optimizing both cluster-level and instance-level objectives. We also propose an effective method for estimating the cluster number in open-world scenarios without knowing the number of new intents beforehand. USNID performs exceptionally well on several benchmark intent datasets, achieving new state-of-the-art results in unsupervised and semi-supervised new intent discovery and demonstrating robust performance with different cluster numbers.

Index Terms—new intent discovery, clustering, representation learning, semi-supervised learning, deep neural networks.

1 Introduction

Discovering new intents is an important aspect of natural language processing, as it has numerous applications in dialogue and user-modeling systems [1], [2]. These newly discovered intents can help to enrich the intent taxonomy and improve the natural language understanding capabilities of dialogue systems in interacting with users [3]. In addition, they can be used to improve user profiles and analyze user interests and preferences, leading to more personalized services [4].

Typical intent understanding tasks use annotated intent corpora to train a supervised classification model [5], with the goal of accurately predicting the corresponding intent category for each text utterance. However, in widespread real-world applications, there are two main difficulties. First, pre-defined intent categories may not be sufficient to capture the complexity and diversity of user needs, requiring the effective mining of potential clusters of user demands and the formation of new intents. Second, in practice, there is often a large amount of unlabeled data, making it labor-intensive and time-consuming to annotate a sufficient quantity of high-quality intent data. Therefore, it is of great significance to find ways to make full use of unlabeled data or semi-supervised data with a limited amount of labels.

To address these issues, we consider the new intent discovery task, which is a clustering problem. For semi-supervised new intent discovery, we randomly select a portion of intent classes as known and the rest as new intents. Considering the scarcity of labeled data in real applications, we mask most labels with known intents (i.e., 90%). The masked known-intent samples and new-intent samples constitute the unlabeled data. The goal is to use limited labeled and a large amount of unlabeled data to find known and discover new intent groups. For unsupervised new intent discovery, it aims to discover new intent groups without any prior knowledge of labeled data.

Novel category discovery (NCD) [6], [7] is similar to our task in computer vision (CV). The main difference is that it assumes unlabeled data only come from novel classes, which is inapplicable in real-world scenarios as unlabeled data usually contain a mix of known and novel categories. While GCD [8] proposed a generalized setting to address this issue, it still requires a larger proportion of labeled data (e.g., 50% v.s. 10%) and does not provide a solution for the unsupervised setting. Additionally, experiments show that the methods used in NCD [9] and GCD [8] have limitations when applied to our task due to their difficulty in learning the complex semantics of discrete text representations.

The study of new intent discovery has gained attention in recent years. We made the first trials on this task [3], [10], and subsequent works have made further progress in improving performance [11], [12], [13]. It has also been successfully applied in real applications to discover user consumption intents [2]. There are three main challenges

- *H. Zhang, H. Xu, X. Wang, and F. Long are with State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: zhang-hl20@mails.tsinghua.edu.cn; xuhua@tsinghua.edu.cn; wx_hebust@163.com; long-f20@mails.tsinghua.edu.cn*
- *X. Wang and K. Gao are with School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China. Email: wx_hebust@163.com; gaokai@hebust.edu.cn*
- *Hua Xu is the corresponding author. Part of the research was completed in cooperation with Samton (Jiangxi) Technology Development Co., Ltd.*
- *This work was supported by the National Natural Science Foundation of China (Grant No. 62173195), the National Science and Technology Major Project towards the new generation of broadband wireless mobile communication networks of Jiangxi Province (03 and 5G Major Project of Jiangxi Province) (Grant No. 20232ABC03402), High-level Scientific and Technological Innovation Talents "Double Hundred Plan" of Nanchang City in 2022 (Grant No. Hongke Zi (2022) 321-16), and Natural Science Foundation of Hebei Province, China (Grant No. F2022208006).*
- *Data and codes are available at https://github.com/thuiar/TEXTOIR.*

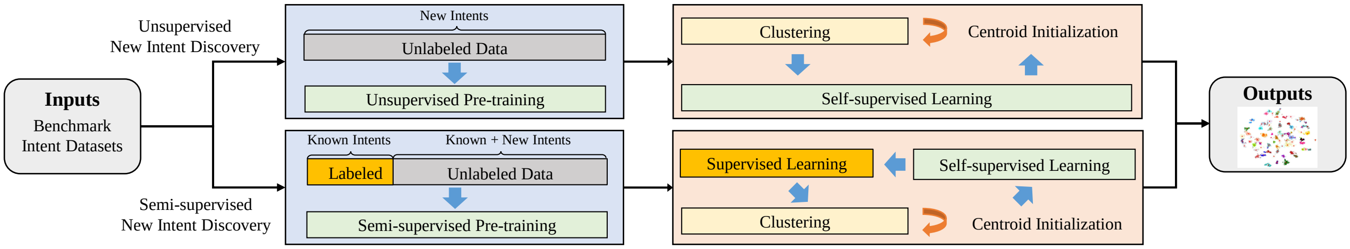

Fig. 1. Overview of the USNID framework for new intent discovery. The unsupervised pipeline first captures primary semantic features from unlabeled data through a pre-training phase and then learns high-level intent representations with two iterative steps. One uses cluster centroids as guidance to obtain consistent targets aligned with the last clustering. The other uses those targets to learn friendly representations for the next clustering. The semi-supervised pipeline further leverages the labeled data as prior knowledge to improve clustering and representation learning.

in this task. Firstly, current methods still heavily rely on labeled data, and their performance suffers significantly in a completely unsupervised setting without any extra knowledge [10], [13]. Secondly, in a semi-supervised scenario, we need to take full advantage of limited labeled data and transfer its knowledge to guide unlabeled data to learn intent representations conducive to clustering. Thirdly, the number of new intents may not be known in advance. In this case, effectively estimating the cluster number is also a crucial factor in determining the final performance.

To tackle these problems, we propose a novel clustering framework called USNID for unsupervised and semi-supervised new intent discovery, as shown in Figure 1. The unsupervised new intent discovery consists of two key steps. The first step is to pre-train the model by applying unsupervised contrastive learning on unlabeled data. We construct positive pairs with each sample and its corresponding strong data augmentation. The second step is to learn high-level intent-wise characteristics through an iterative process of clustering and self-supervised learning.

However, the cluster assignments from the partition-based method (e.g., K-Means [14]) may not be consistent for the same sample across different clustering, making it difficult to use them as pseudo-labels to train a stable classifier for discriminating new intent classes. To overcome this issue, we introduce a centroid-guided clustering mechanism that leverages cluster centroids from adjacent clustering as guidance to obtain aligned targets. One way to achieve this is by minimizing Euclidean distances between the two cluster centroid matrices globally to obtain an alignment projection. Still, each clustering process can still be inefficient and prone to falling into local optima due to the randomness of initial cluster centroid selection. To mitigate this, we propose a centroid initialization strategy that leverages the cluster centroids from the previous iteration's clustering to initialize the current iteration's clustering. This strategy can improve convergence with the prior knowledge of historical clustering information. Moreover, the produced cluster assignments are usually consistent with the results of centroid alignment, which can be directly used as self-supervised signals for representation learning. It is also important to select a suitable self-supervised learning objective to provide friendly representations for the next clustering. The designed objective captures both cluster-level and instance-level information using aligned targets. The former learns a discriminator to distinguish different fine-grained intent classes, while the latter aims to enhance the semantic similarity relationships between instances with intra-class compactness and inter-class separation properties.

The semi-supervised new intent discovery process uses labeled data in two ways to improve performance. First, we optimize the pre-training phase using a combination of semi-supervised contrastive learning and known-intent classification objectives, which utilize limited labeled data to guide the learning of primary semantic features in a large amount of unlabeled data. Second, we incorporate a supervised contrastive learning objective to enhance the memory of the limited labeled data and improve the ability to cluster and learn representations. Nevertheless, the approach still requires the cluster number to be specified in advance, which is not practical in real-world situations. Thus, we propose a simple and effective method for estimating the number of new-intent classes. Our method only requires one clustering operation using a large, pre-defined number of clusters. The main idea is to use the knowledge acquired during the pre-training phase to find high-quality clusters that are denser than a certain threshold. In semi-supervised scenarios, limited labeled data can be used to induce clusters corresponding to known intents and avoid interference with the estimation of the number of new intents.

Our USNID framework is evaluated on several benchmark intent datasets and compared with 15 algorithms that can be used in unsupervised and semi-supervised new intent discovery. It is the first successful attempt at unsupervised new intent discovery, resulting in an absolute increase of 20-30% adjusted rand index (ARI) over the state-of-the-art (SOTA) unsupervised clustering method. In addition, it achieves new SOTA performance in semi-supervised new intent discovery, showing substantial improvements over previous best-performing methods under different known class ratios. The method of estimating the number of intent classes is also evaluated and found to accurately predict the actual number with the lowest errors compared with other estimation methods. Even when the cluster number is varied in a wide range, our approach still achieves robust and the best performance among all methods.

2 Related Works

In this section, we briefly review the most relevant research in areas of unsupervised and constrained clustering, novel class discovery, and new intent discovery.

There are numerous classic unsupervised clustering technologies in the literature [14], [15]. K-Means is a particularly attractive partitioning method among them due to its simplicity and relatively low time complexity [16]. However, it can suffer from poor performance due to the arbitrary sampling of centroids. Therefore, several variants of K-Means have been proposed to address this issue [17], [18]. K-Means++ [18] is selected in this work due to its superior convergence and speed. Deep neural networks (DNNs) have gained popularity in recent years due to their proficiency in handling high-dimensional data and capture complex underlying semantics [19]. As a result, deep clustering methods have been widely studied [20]. For example, Deep embedded clustering (DEC) [21] utilizes a stacked autoencoder (SAE) [22] for low-dimensional feature learning and cluster assignment optimization. Deep clustering network (DCN) [23] also uses an SAE, optimizing both reconstruction loss and K-Means-like regularization. Deep adaptive clustering (DAC) [24] learns pairwise similarities from confident samples, and DeepCluster [25] alternates between clustering and feature learning. Unsupervised contrastive learning [26] is a rising approach, as seen in Contrastive clustering (CC) that performs instance and cluster-level contrastive learning, and SCCL [26], which optimizes instance-level contrastive loss and clustering loss for superior text clustering.

Constrained clustering [27], introduced to enhance unsupervised clustering through extra supervised signals (e.g., labeled data), includes methods like COP-KMeans [28] which incorporates hard pairwise constraints. PCK-Means [29] handles constraint violations with penalty terms while MPCK-Means [30] adds a distance metric learning objective. DNNs are employed for powerful constrained clustering representations in methods like KCL [31] and MCL [32]. KCL trains a DNN with pairwise similarity data and generates weak-supervised signals for unlabeled data. MCL uses categorical distribution similarities as weak pairwise constraints. ASFS [33] optimizes relationships in diverse data types using unlabeled data and semantic regression, and it uses a graph-based constraint for accurate label prediction. ASTCA [34] introduces an adaptive model applied successfully in UAV tracking, providing a unique perspective on unsupervised clustering.

Novel class discovery [6] in computer vision aims to identify new visual classes using labeled data. Approaches include DTC [9], which pre-trains a model with labeled data and incorporates temporal ensemble predictions into DEC loss, and introduces a method to estimate the number of new classes. RankStats [35] uses unlabeled, self-augmented data in pre-training and calculates pairwise similarities via ranking statistics. UNO [7] optimizes with a unified cross-entropy loss by swapping pseudo-labels of concatenated neural classifier outputs from labeled and unlabeled data. However, these methods assume unlabeled data includes only new classes, which might not be appicable in the real world. GCD [8] rectifies this by accommodating both known and new classes in unlabeled data, combining supervised and unsupervised contrastive losses for representation learning and semi-supervised k-means for inference. Despite its superior performance in this task, it struggles with discrete text representations.

The research of new intent discovery is still in its infancy. Prior studies typically focus on known intent classification within closed-world scenarios, utilizing typical intent benchmark datasets [36], [37]. More recently, attention has shifted to a related area known as open intent detection [38], [39], which seeks to detect the unknown class during testing but lacks the ability to discern fine-grained new classes. We have conducted a pilot study using the CDAC+ [3] algorithm, which first captures pairwise sentence relationships with the guidance of labeled data and then refines cluster assignments with the DEC loss. Another of our works, DeepAligned [10], initializes intent representations under the supervision of labeled data and then iteratively performs clustering and representation learning, aligning cluster centroids between adjacent iterations to obtain consistent self-supervised signals. DCSC [12] improves the pre-training stage by applying contrastive losses to both labeled and unlabeled data. It mainly uses the SwAV [40] algorithm for unsupervised learning, which requires each sample to predict the swapped view and uses Sinkhorn-Knopp [41] to produce soft cluster assignments. MTP-CLNN [13] is the current SOTA method, which has two key features. First, it enhances representations by incorporating strong prior knowledge from a network pre-trained on external data in the intention domain (i.e., CLINC dataset [42]) and adding a masked language modeling (MLM) task. Second, it adapts the SCAN algorithm [43] to the semi-supervised setting, creating positive pairs with K-nearest neighbors (KNNs) or samples with the same labels for contrastive learning. However, this method relies heavily on the selected external data, and its performance drops dramatically in a purely unsupervised scenario [13].

3 Problem Formulation

Unsupervised setting: We are given an intent dataset $\mathcal{D}_{\text{un}} = \{\boldsymbol{x}_i | y_i \in \mathcal{I}, i = 1, ..., N\}$, where $\boldsymbol{x}_i$ is the i^{th} utterance, y_i is the ground-truth label (unseen during training), N is the number of all utterances. $\mathcal{I} = \{\mathcal{I}_i\}_{i=1}^{K}$ is the set of intent labels, where K is the number of intent classes. The goal of unsupervised new intent discovery is to cluster $\{\boldsymbol{x}_i\}_{i=1}^{N}$ into K intent groups.

Semi-supervised setting: We are given an intent dataset $\mathcal{D}_{\text{semi}} = \{\mathcal{D}_{\text{semi}}^{l}, \mathcal{D}_{\text{semi}}^{u}\}$, where $\mathcal{D}_{\text{semi}}^{l}$ and $\mathcal{D}_{\text{semi}}^{u}$ are subsets with limited labeled data (e.g., the labeled ratio $\frac{|\mathcal{D}_{\text{semi}}^{l}|}{|\mathcal{D}_{\text{semi}}|} < 10\%$) and unlabeled data, respectively.

Specifically, $\mathcal{D}_{\text{semi}}^{l} = \{(\boldsymbol{x}_i, y_i) | y_i \in \mathcal{I}^{\text{known}}, i = 1, ..., M\}$, where M is the number of labeled utterances, $\mathcal{I}^{\text{known}} = \{\mathcal{I}_i\}_{i=1}^{K^{\text{known}}}$ is the set of known intent labels. K^{known} is the number of known intent classes which is smaller than K (e.g., the known class ratio $\frac{K^{\text{known}}}{K}$ is varied among 25%, 50%, and 75% in this task).

$\mathcal{D}_{\text{semi}}^{u} = \{\boldsymbol{x}_i | y_i \in \mathcal{I}, i = M+1, ..., N\}$, where $\mathcal{I} = \{\mathcal{I}^{\text{known}}, \mathcal{I}^{\text{new}}\}$, and $\mathcal{I}^{\text{new}} = \{\mathcal{I}_i\}_{i=K^{\text{known}}+1}^{K}$ is the set of new intent labels. Note that $\mathcal{D}_{\text{semi}}^{u}$ also contains samples from $\mathcal{I}^{\text{known}}$, which is closer to real-world applications with a mixture of both known and new classes for unlabeled data. In comparison, $\mathcal{D}_{\text{semi}}^{u}$ only contains samples from $\mathcal{I}^{\text{new}}$ in the similar new class discovery task [6]. The goal of semi-supervised new intent discovery is to use $\mathcal{D}_{\text{semi}}^{l}$ as prior knowledge to help learn clustering-friendly representations and find known and discover new intent groups.

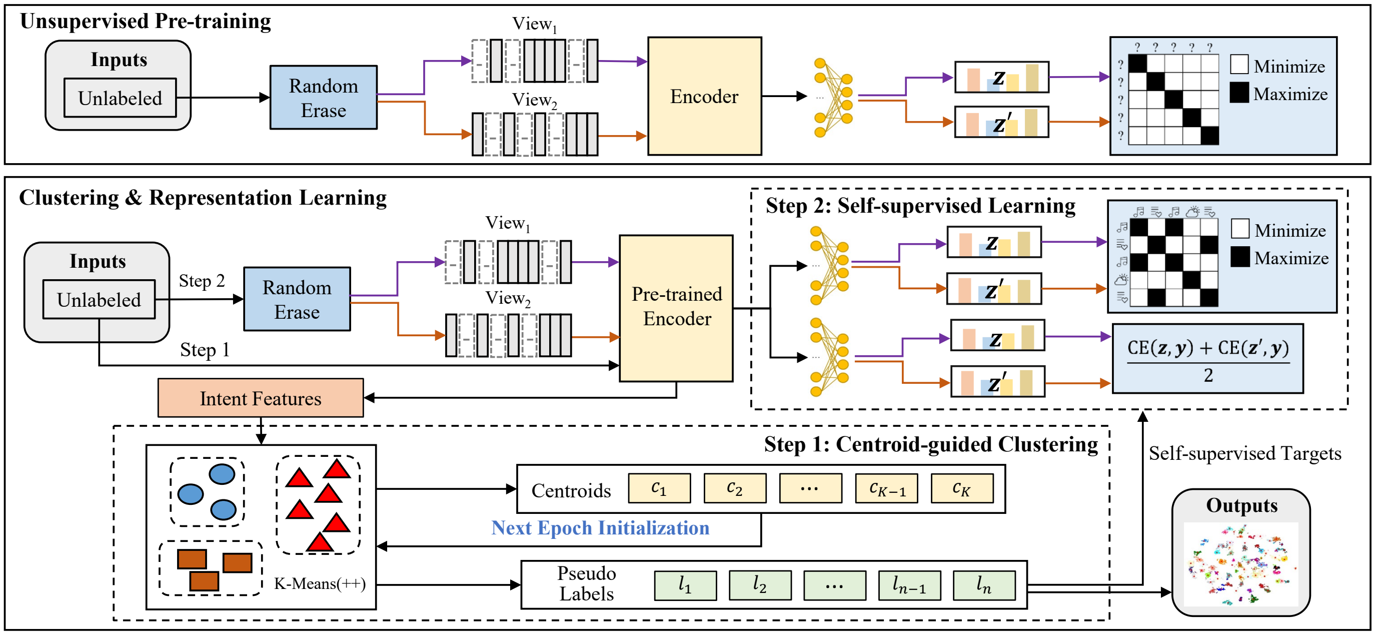

Fig. 2. The pipeline of unsupervised new intent discovery. It first pre-trains the model by applying unsupervised contrastive learning with strong augmented samples. Then, it alternatively performs clustering and representation learning. On the one hand, an efficient centroid-guided clustering algorithm is introduced to produce aligned cluster assignments between adjacent clustering, which can converge well and be used as high-quality self-supervised signals. On the other hand, we learn cluster-level and instance-level information to obtain clustering-friendly intent representations.

4 Methodologies

In this section, we introduce a new clustering framework, USNID. The pipelines of unsupervised and semi-supervised new intent discovery are presented in Figure 2 and Figure 3.

4.1 Intent Representation

The pre-trained BERT language model shows excellent performance in a wide range of NLP tasks [44]. Thus, it is adopted to extract deep intent representations.

Specifically, for each utterance $\boldsymbol{x}_i$, we take it as input to BERT in the needed format (i.e., the first token is [CLS]) and obtain its final hidden vectors $[C, T_1, ..., T_L] \in \mathbb{R}^{(L+1)\times H}$ of each token through non-linear projection layers of BERT, where L is the length of the i^{th} utterance, H is the hidden size 768. The sentence representation $\boldsymbol{s}_i \in \mathbb{R}^H$ is first obtained by applying mean-pooling operation on the hidden vectors of these tokens:

$$\boldsymbol{s}_i = \text{mean-pooling}([C, T_1, ..., T_L]). \tag{1}$$

Then, a fully-connected layer h is added to enhance the capability to capture the complex semantics of high-dimensional text data, yielding the intent representation $\boldsymbol{I}_i \in \mathbb{R}^D$:

$$\boldsymbol{I}_i = h(\boldsymbol{s}_i) = W_h \boldsymbol{s}_i + b_h, \tag{2}$$

where D is the feature dimension, $W_h \in \mathbb{R}^{H\times D}$ and $b_h \in \mathbb{R}^D$ are weight matrices and bias vectors, respectively.

4.2 Unsupervised New Intent Discovery

4.2.1 Unsupervised Pre-training

Initially, samples from different intent classes often overlap in the feature space, which can hinder the clustering optimization, leading to suboptimal clusters [45]. Well-initialized intent representations, which should be distributed uniformly in the feature space and effectively reflect the data characteristics, can improve clustering performance and convergence [16], [18]. Therefore, our goal in pre-training the model is to push apart distinct samples while capturing implicit semantic relationships between augmented pairs.

Given that only unlabeled data are available, a common way to construct positive pairs is to use two augmented views of the same sample, as suggested in [26]. In particular, let $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}'_i$ be two views of $\boldsymbol{x}_i$ in a mini-batch of n samples. $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}'_i$ are treated as a positive pair, and they form negative pairs with the remaining $2n-2$ augmented samples. They are first encoded as intent representations $\tilde{\boldsymbol{I}}_i$ and $\tilde{\boldsymbol{I}}'_i$, as introduced in section 4.1. Then, we add a non-linear projection head $f_1^u : \mathbb{R}^D \to \mathbb{R}^K$ to obtain $\boldsymbol{z}_i$ and $\boldsymbol{z}'_i$. The unsupervised contrastive loss $\mathcal{L}_{\text{ucl}}$ is defined as:

$$\mathcal{L}_{\text{ucl}} = -\frac{1}{2n}\sum_{i=1}^{2n} \log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}'_i)/\tau)}{\sum_{j=1}^{2n} \mathbb{I}_{[j\neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}, \tag{3}$$

where $\text{sim}(\boldsymbol{a}, \boldsymbol{b})$ performs dot product on L2-normalized $\boldsymbol{a}$ and $\boldsymbol{b}$, τ is the temperature parameter, and $\mathbb{I}(\cdot)$ is an indicator function outputs 1 iff $j \neq i$ and 0 otherwise.

A simple yet effective method, *random erase*, is used as a strong data augmentation for new intent discovery. Specifically, for an utterance $\boldsymbol{x}_i$ with length L, we randomly select $\lfloor L \times a\% \rfloor$ different words and erase them from $\boldsymbol{x}_i$, where a is the erase ratio in a sentence. The intuition is that this operation explicitly provides hard positive pairs (i.e., different missing sets of words) for contrastive learning, which is beneficial to capture the fine-grained semantic relations between different local word sets in a sentence.

After pre-training, we remove the head used in contrastive learning to avoid any unwanted bias that might interfere with the subsequent steps. The rest of the backbone is saved for clustering and representation learning.

4.2.2 Centroid-guided Clustering

Partitioning clustering methods such as the K-Means algorithm can be used to discover intent-wise clusters. However, its effectiveness can be compromised by the choice of initial centroids. In scenarios where the initial centroids are suboptimal, the algorithm risks falling into a local minima, resulting in unsatisfactory clustering. This shortcoming is addressed by K-Means++ [18], which adopts a probabilistic approach to select new centroids, thereby improving convergence and achieving an optimal solution more quickly than the standard K-Means. Consequently, our work utilizes K-Means++ for clustering purposes.

We found that directly using K-Means++ still performs poorly due to a lack of guidance to help enhance the intent representation capability. Thus, we aim to use the clustering information to construct high-quality self-supervised signals for learning high-level intent representations. A natural way to do this is to use the cluster assignments as pseudo-labels for supervision. This is based on the typical clustering approach proposed by [25], where they alternate between clustering and optimizing the convnets based on predicting the cluster assignments. They demonstrate structured outputs of neural networks used as weakly supervised signals can also benefit the unsupervised representation learning. However, a challenge arises as the same sample could be assigned to different clusters across multiple iterations due to centroid selection randomness. Although Caron et al. [25] propose randomly re-initializing classifier parameters before each training iteration to address this, this strategy fails to make effective use of historical training information [46].

In this work, we introduce a novel centroid-guided mechanism to address inconsistencies in self-supervised targets between training iterations and enhance knowledge retention in the classifier. Noting that while cluster assignments may fluctuate between iterations, cluster centroids remain relatively stable due to their global optimization as averaged features, we propose using these centroids as guidance. This approach aims to provide consistent self-supervised targets across training iterations and preserve the well-trained knowledge of the classifier, enhancing the overall effectiveness of the iterative process.

In particular, the cluster centroids and assignments in the last and current training iterations are denoted as $\boldsymbol{C}^{(t-1)}$, $\boldsymbol{y}^{(t-1)}$, and $\boldsymbol{C}^{(t)}$, $\boldsymbol{y}^{(t)}$, respectively. After the $(t-1)^{\text{th}}$ clustering, $\boldsymbol{y}^{(t-1)}$ is used as supervision for feature learning, which helps capture similarity relationships of the samples close to $\boldsymbol{C}^{(t-1)}$. The updated representations are then used for the $(t)^{\text{th}}$ clustering, which generates $\boldsymbol{C}^{(t)}$. The intuition is that $\boldsymbol{C}^{(t)}$ and $\boldsymbol{C}^{(t-1)}$ have relatively consistent distributions in the feature space, and $\boldsymbol{C}^{(t)}$ is aligned with $\boldsymbol{C}^{(t-1)}$ to obtain the optimal mapping G_{opt} as below:

$$G_{\text{opt}} = \underset{G}{\operatorname{argmin}} \left\{ \sum_{i=1}^{K} \| \boldsymbol{C}_i^{(t)} - \boldsymbol{C}_{g_i}^{(t-1)} \|_2 \right\}, \tag{4}$$

where $G : \{1, ..., K\} \rightarrow \{1, ..., K\}$ is a one-to-one mapping, $g_i = G(i)$ is the centroid index corresponding to i in the last iteration. It can be optimized with the Hungarian algorithm [47] to obtain G_{opt}. Then, $\boldsymbol{C}^{(t)}$ and $\boldsymbol{y}^{(t)}$ are updated by:

$$\boldsymbol{C}_i^{(t)} = \boldsymbol{C}_{g_i'}^{(t-1)}, \text{ s.t. } g_i' = G_{\text{opt}}^{-1}(i), \forall i \in \{1, ..., K\}, \tag{5}$$

$$y_i^{(t)} = G_{\text{opt}}^{-1}(y_i^{(t-1)}), \forall i \in \{1, ..., N\}, \tag{6}$$

where G_{opt}^{-1} is the inverse mapping of G_{opt}. The preliminary results of this centroid-guided alignment strategy have been presented in our previous work [10] and show substantial improvements compared with the re-initialization strategy.

However, this strategy is not efficient due to the high time cost of multiple clustering. The reason is that each clustering (i.e., K-Means++) also needs to initialize the first centroid at random, which may select sub-optimal centroids and lead to a degradation of convergence. Thus, finding the optimal solution will take a lot more time. To solve this problem, we propose a concise centroid-guided initialization strategy, aiming to leverage the historical clustering information to improve convergence. Specifically, K-Means++ is only performed at the first training iteration. Then, the cluster centroids produced in the $(t-1)^{\text{th}}$ training iteration are used to initialize K-Means, yielding $\boldsymbol{y}^{(t)}$ and $\boldsymbol{C}^{(t)}$:

$$\boldsymbol{y}^{(t)}, \boldsymbol{C}^{(t)} = \begin{cases} \text{K-Means++}\left(\boldsymbol{I}^{(t)}\right), & \text{if } t = 0, \\ \text{K-Means}\left(\boldsymbol{I}^{(t)}, \boldsymbol{C}^{(t-1)}\right), & \text{if } t \geq 1. \end{cases}$$
$$\text{s.t. } \boldsymbol{I}^{(t)} = \text{Learn}(\boldsymbol{I}^{(t-1)}, \boldsymbol{y}^{(t-1)}), t \geq 1;\ t \in \mathbb{N}, \tag{7}$$

where $\boldsymbol{I}^{(0)}$ denotes the initial intent representations after pre-training, $\text{Learn}(\boldsymbol{I}, \boldsymbol{y})$ denotes the representation learning process with pseudo-labels $\boldsymbol{y}$ as supervision, which will be introduced in section 4.2.3 in detail. With a random centroid initialization, we need to perform alignment between $\boldsymbol{C}^{(t-1)}$ and $\boldsymbol{C}^{(t)}$ to obtain G_{opt} as in Eq. 16. Interestingly, we found that this strategy does not need the alignment process, as experiments show that $G_{\text{opt}}(i) = i,\ \forall i \in \{1, ..., K\}$ usually works, and $\boldsymbol{y}^{(t)}$ can be directly used as aligned targets, which also converge well. It is reasonable because centroid initialization ensures the stability of cluster allocation targets between adjacent clustering and is beneficial to find the optimal solution with prior knowledge of previous cluster centroids.

The stopping criterion of training is to compare cluster assignments between adjacent clustering $\boldsymbol{y}^{(t)}$ and $\boldsymbol{y}^{(t-1)}$:

$$\delta = \frac{\sum_{i=1}^{N} \mathbb{I}\left\{ y_i^{(t)} \neq y_i^{(t-1)} \right\}}{N}, \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 only if the condition holds. Otherwise, it outputs 0. δ indicates the proportion of the difference between $\boldsymbol{y}^{(t)}$ and $\boldsymbol{y}^{(t-1)}$, which values are in the range of [0, 1]. It can well reflect the convergence of the proposed clustering algorithm. The procedure will be stopped when δ is smaller than some threshold δ_{th}. During the inference phase, we perform another K-Means using the cluster centroids that have been well-trained in the previous clustering as the initialization.

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2023.3340732

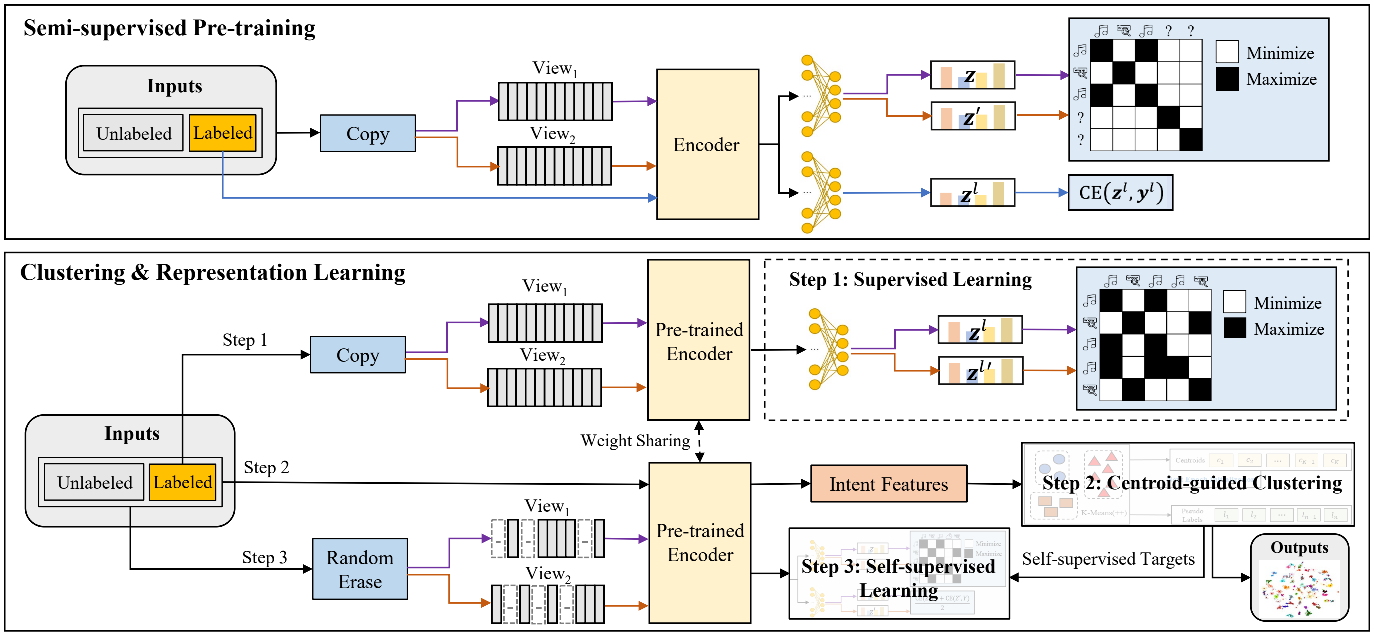

Fig. 3. The pipeline of semi-supervised new intent discovery. On the basis of unsupervised new intent discovery, it enhances the pre-training stage by incorporating labeled data through the use of both semi-supervised contrastive and cross-entropy losses. It also improves the clustering and representation learning stage by adding a supervised contrastive learning step on labeled data to address the issue of *catastrophic forgetting*.

4.2.3 *Self-supervised Learning*

After each clustering, we learn discriminative intent representations to further promote the subsequent clustering. We achieve this through a dual learning strategy that operates at both the instance and cluster levels, drawing inspiration from the methodologies presented in recent studies [40], [43]. Instance-level learning focuses on ensuring that similar instances are allocated to the same class while differentiating them from instances that belong to other classes. This proves especially effective under strong data augmentations [48]. It fosters the development of intra-class compactness and inter-class separability within our model, both of which are fundamental properties that aid in clustering.

Simultaneously, we perform cluster-level learning by updating the model parameters based on the aligned cluster assignments $\boldsymbol{y}^a$. This approach enhances the model's discriminative power to discern and classify intents by predicting cluster assignments. By concurrently learning at both the instance and cluster levels, we gain a more comprehensive and nuanced understanding of the data. This strategy, which focuses on the fine-grained details of individual instances and the overarching characteristics of clusters, significantly enhances the overall performance and effectiveness of our clustering efforts.

In particular, we first perform *random erase* data augmentation and use the pre-trained encoder in section 4.2.1 to extract two views of intent representations $\tilde{\boldsymbol{I}}$ and $\tilde{\boldsymbol{I}}'$. To capture cluster-level information, we perform the classification loss $\mathcal{L}_{\text{cls}}$:

$$\mathcal{L}_{\text{cls}} = \frac{\mathcal{L}_{\text{ce}}(\boldsymbol{z}, \boldsymbol{y}^a) + \mathcal{L}_{\text{ce}}(\boldsymbol{z}', \boldsymbol{y}^a)}{2}, \tag{9}$$

$$\mathcal{L}_{\text{ce}}(\boldsymbol{z}, \boldsymbol{y}^a) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp((\boldsymbol{z}_i)^{y_i^a})}{\sum_{j=1}^{K} \exp((\boldsymbol{z}_i)^j)}, \tag{10}$$

where $\boldsymbol{z}$ and $\boldsymbol{z}'$ are obtained with $f_2^u : \mathbb{R}^D \rightarrow \mathbb{R}^K$, and $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss. As $\boldsymbol{y}^a$ is relatively consistent between adjacent clustering, it can provide stable supervised signals for learning cluster-level patterns. Two augmented views are used as hard examples for classification, which is beneficial to enhance the model's discrimination ability.

To capture instance-level information, we aim to pull samples from the same class close to each other and push samples from different classes away from each other. For this purpose, we apply the supervised contrastive loss as suggested in [49]:

$$\mathcal{L}_{\text{scl}} = -\frac{1}{2n}\sum_{i=1}^{2n}\frac{1}{|P(i)|}\sum_{p\in\mathcal{P}(i)} \log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_p)/\tau)}{\sum_{j=1}^{2n}\mathbb{I}_{[j\neq i]}\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}, \tag{11}$$

where $\boldsymbol{z}_i$ is obtained with $f_3^u : \mathbb{R}^D \rightarrow \mathbb{R}^K$, n is the number of samples in a mini-batch, $\mathcal{P}(i)$ is the set of indices of augmented samples with the same class of $\boldsymbol{z}_i$, and $|\cdot|$ denotes the size of a set.

The overall loss of self-supervised learning can be written as follows:

$$\mathcal{L}_{\text{self-sup}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{scl}}. \tag{12}$$

That is, we joint train $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{scl}}$ to learn both cluster-level and instance-level characteristics, which is helpful to obtain friendly representations for clustering.

While some SOTA unsupervised clustering methods, such as CC and SCCL, also perform representation learning at both the instance and cluster levels, our method stands apart due to three key differentiating factors. First, our method integrates a pre-training strategy to discern distinct instances during the initial stages while concurrently learning latent correlations. This serves as an effective regularization procedure, facilitating more amenable representations for subsequent clustering. Second, instead of relying solely on weak pairwise constraints employed by CC and SCCL, our method generates specific pseudo-labels as self-supervised signals for each sample. Importantly, the cluster-level learning objectives in our model can explicitly differentiate between various intent classes. Third, our approach introduces an innovative perspective on the creation of high-quality self-supervised targets. Particularly, the centroid-guided mechanism enables effectively leveraging historical clustering data to generate aligned instance-level pseudo-labels. This not only greatly enhances clustering performance but also leads to excellent convergence. These distinct features enable our method to outperform current SOTA approaches, achieving substantial improvements of over 30% in standard clustering metrics.

4.3 Semi-supervised New Intent Discovery

4.3.1 Semi-supervised Pre-training

In the pre-training phase, we hope to fully utilize the limited annotated intent data to provide well-initialized representations for clustering. For this purpose, we perform data augmentation on both labeled and unlabeled data and mix them in nearly equal ratios within a mini-batch for contrastive learning. The positive pairs include: (a) samples with the same class in the labeled data, (b) each sample with its augmented view in both labeled and unlabeled data. Thus, we propose the semi-supervised contrastive loss:

$$\begin{aligned}&\mathcal{L}_{\text{semi-scl}} = \\ &-\frac{1}{2n}[\sum_{z_i \in \{\boldsymbol{z}^l\}} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_p)/\tau)}{\sum_{j=1}^{2n} \mathbb{I}_{[j \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)} \\ &+ \sum_{z_i \in \{\boldsymbol{z}^u\}} \log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_i')/\tau)}{\sum_{j=1}^{2n} \mathbb{I}_{[j \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}], \end{aligned} \tag{13}$$

where $\boldsymbol{z}_i$ and $\boldsymbol{z}_i'$ are obtained with $f_1^s : \mathbb{R}^D \rightarrow \mathbb{R}^{K^{\text{known}}}$. $\{\boldsymbol{z}^l\}$ and $\{\boldsymbol{z}^u\}$ are the sets of labeled and unlabeled data, respectively, which satisfy $|\{\boldsymbol{z}^l\}| + |\{\boldsymbol{z}^u\}| = 2n$. Here we use a simple data augmentation method, *dropout* [50], which is efficient and works well. It uses dropout masks in transformers to produce positive pairs with the same sample feed-forward twice in neural networks. Moreover, we add a cross-entropy loss $\mathcal{L}_{\text{ce}}$ with supervised signals of labeled data $\boldsymbol{y}^l$ to enhance the discrimination ability for known classes. The final loss of semi-supervised pre-training is defined as:

$$\mathcal{L}_{\text{semi-pre}} = \mathcal{L}_{\text{semi-scl}} + \mathcal{L}_{\text{ce}}(\boldsymbol{z}^l, \boldsymbol{y}^l), \tag{14}$$

where $\boldsymbol{z}^l$ is obtained with $f_2^s : \mathbb{R}^D \rightarrow \mathbb{R}^{K^{\text{known}}}$. Similar to unsupervised pre-training, we use the pre-trained network without f_1^s and f_2^s in the subsequent steps.

4.3.2 Clustering and Representation Learning

After pre-training, we can directly use all data to perform clustering and representation learning as in unsupervised new intent discovery. However, as the training iteration goes on, we notice that some labeled samples from the same class may be assigned to different clusters when mixed with unlabeled data for clustering, which is also described as the *catastrophic forgetting* phenomenon in [35]. To alleviate this problem, we propose using supervised contrastive learning with labeled data at the beginning of each training iteration. Specifically, we use the *dropout* strategy to generate augmented samples and add a new head $f_3^s : \mathbb{R}^D \rightarrow \mathbb{R}^K$ to perform $\mathcal{L}_{\text{scl}}$ as in Eq. 11. It can not only strengthen the *memory* of supervised similarity relationships but also be beneficial to guide the subsequent clustering process.

Then, we successively carry out centroid-guided clustering (section 4.2.2) and self-supervised learning (section 4.2.3) in the rest of each training iteration. f_3^s and another head $f_4^s : \mathbb{R}^D \rightarrow \mathbb{R}^K$ are used for learning instance-level and cluster-level information, respectively.

4.4 Estimate the Cluster Number K

To deal with an unknown cluster number K, we propose a simple yet effective method for estimating K in unsupervised and semi-supervised new intent discovery. Specifically, the intent representations $\boldsymbol{I}$ after pre-training are first used to perform K-Means++ with a large assigned cluster number K′. Though the pre-training phase lacks explicit supervised signals for distinguishing fine-grained clusters, it can still help capture weak semantic similarity relations using positive augmented samples or limited labeled data.

The assumption is that real clusters tend to be confident of having much more samples that are similar to each other. Specifically, in the unsupervised setting, we remove the low-confidence clusters and estimate K by:

$$K = \sum_{k=1}^{K'} \mathbb{I}\left\{|\mathcal{C}_k| \geq t\right\}, \tag{15}$$

where $\mathcal{C}_k = \{\boldsymbol{x}_i | y_i = k, i = 1, 2, ..., N\}$, t is a threshold defined as the mean cluster size $\frac{\sum_{k=1}^{K'} |\mathcal{C}_k|}{K'}$.

In the semi-supervised setting, we have access to a set of known intent classes with a number K^{known} through limited labeled data. The goal is to estimate the number of new intent classes K^{new}. Since the unlabeled data come from both known and new classes, we need to first distinguish the known intent clusters from clustering results. For this purpose, we propose to use the limited labeled data as prior knowledge for cluster induction. In particular, we perform the Hungarian algorithm to obtain the alignment projection G'_{opt} between the labeled centroids $\boldsymbol{C}^l \in \mathbb{R}^{K^{\text{known}} \times D}$ and the cluster centroids $\boldsymbol{C} \in \mathbb{R}^{K' \times D}$ in the Euclidean space:

$$G'_{\text{opt}} = \underset{G'}{\operatorname{argmin}} \left\{ \sum_{i=1}^{K^{\text{known}}} \|\boldsymbol{C}_i^l - \boldsymbol{C}_{b_i}\|_2 \right\}, \tag{16}$$

where $G' : \{1, ..., K^{\text{known}}\} \rightarrow \{1, ..., K'\}$, $b_i = G'(i)$ and i are the corresponding centroid indices, and $\boldsymbol{C}^l$ is calculated by averaging intent representations of each class in the labeled samples. Then, we can find the set of known intent cluster indices $S = \{G_{opt}^{'}(i)\}_{i=1}^{K^{\text{known}}}$, and K^{new} is calculated by:

$$K^{\text{new}} = \sum_{k=1}^{K'} \mathbb{I}\left\{(|\mathcal{C}_k| \geq t) \wedge (k \notin S)\right\}, \tag{17}$$

TABLE 1
Statistics of BANKING, CLINC150, and StackOverflow datasets. # indicates the total number of sentences. The unsupervised setting only contains new intents. In the semi-supervised setting, we randomly select 25%, 50%, and 75% intents as known and treat the remaining as new intents.

Dataset	#Known Classes + #New Classes	#Training	#Validation	#Testing	Vocabulary	Length (max / mean)
BANKING	0 + 77 / 19 + 58 / 39 + 38 / 58 + 19	9,003	1,000	3,080	5,028	79 / 11.91
CLINC150	0 + 150 / 38 + 112 / 75 + 75 / 113 + 37	18,000	2,250	2,250	7,283	28 / 8.31
StackOverflow	0 + 20 / 5 + 15 / 10 + 10 / 15 + 5	12,000	2,000	6,000	17,182	41 / 9.18

where t is the same as in the unsupervised setting. The total cluster number K is the summation of K^{known} and K^{new}.

5 Experiments

5.1 Datasets

We evaluate the new intent discovery performance with three challenging benchmark datasets: BANKING [36], CLINC150 [42], and StackOverflow [51]. The detailed statistics for these datasets are shown in Table 1.

The BANKING dataset is a collection of customer service queries specifically from the banking domain, comprising 13,083 queries across 77 classes. We follow the data splits in [36] and create a validation set of 1,000 randomly sampled utterances from the original training set.

The CLINC150 dataset is an out-of-scope intent classification dataset with 150 classes across ten domains. Since the out-of-scope utterances lack specific intent annotations for evaluation, we only use the 22,500 in-scope queries in this work. The dataset is split into training, validation, and testing sets by 8:1:1.

The StackOverflow dataset originally contains 3,370,528 technical question titles on Kaggle.com[1]. In this work, we use the curated version of the dataset presented in [51], consisting of 20,000 samples across 20 classes. The dataset is split into training, validation, and testing sets by 6:1:3.

5.2 Baselines

5.2.1 Unsupervised Clustering

The unsupervised clustering baselines include traditional machine learning methods: KM [14], AG [15], SAE-KM, and deep clustering methods: DEC [21], DCN [23], CC [52], SCCL [45].

For KM and AG, the intent representations are extracted with GloVe [53] by averaging the pre-trained 300-dimensional token embeddings in the sentence. For SAE-KM, DEC, and DCN, a stacked autoencoder (SAE) [22] is used to capture semantically meaningful and discriminative representations [21]. Since CC is a method in the field of CV, we adapt it to this task by using BERT to extract intent representations. For SCCL, we use the Sentence transformer [54] as the backbone suggested in [45].

5.2.2 Semi-supervised Clustering

The semi-supervised clustering baselines contain a series of SOTA methods in related fields, including: constrained clustering: KCL [31], MCL [32], novel class discovery: DTC [9], GCD [8], and new intent discovery: CDAC+ [3], DeepAligned [10], DCSC [12], MTP-CLNN [13].

Since KCL, MCL, DTC, and GCD are used for CV tasks, we adapt them to our task by using the BERT backbone. For MTP-CLNN, the parameter of top-K nearest neighbors is set to 50, 60, 300 for BANKING, CLINC150, and StackOverflow, respectively, which is used or calculated as in [13]. For a fair comparison, the external dataset is not used in MTP-CLNN as other baselines.

5.3 Evaluation Metrics

Three widely used metrics are adopted to evaluate the clustering performance, including normalized mutual information (NMI), adjusted rand index (ARI), and accuracy (ACC). The higher values of these metrics indicate better performance. Specifically, NMI is defined as:

$$\text{NMI}(\mathbf{y}^{gt}, \mathbf{y}^{p}) = \frac{MI(\mathbf{y}^{gt}, \mathbf{y}^{p})}{\frac{1}{2}(H(\mathbf{y}^{gt}) + H(\mathbf{y}^{p}))}, \tag{18}$$

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 defined as:

$$\text{ARI} = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - [\sum_i \binom{u_i}{2}\sum_j \binom{v_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{u_i}{2} + \sum_j \binom{v_j}{2}] - [\sum_i \binom{u_i}{2}\sum_j \binom{v_j}{2}]/\binom{n}{2}}, \tag{19}$$

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 i^{th} predicted label and the j^{th} ground-truth label. The values of ARI are in the range of [-1, 1].

ACC is defined as:

$$\text{ACC}(\mathbf{y}^{gt}, \mathbf{y}^{p}) = \max_{m} \frac{\sum_{i=1}^{n} \mathbb{I}\left\{y_i^{gt} = m\left(y_i^{p}\right)\right\}}{n}, \tag{20}$$

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

5.4 Experimental Settings

In the unsupervised setting, we use the data in both training and validation sets for unsupervised learning with the aim of discovering intent-wise clusters in the testing set. In the semi-supervised setting, we randomly select a certain percentage (25%, 50%, and 75%) of known intent classes. In the training set, we keep labels for a limited portion (10%) of the data from these known classes, while the remaining

[1] https://www.kaggle.com/competitions/predict-closed-questions-on-stack-overflow/data

TABLE 2
Results of new intent discovery on the three datasets. KCR denotes the known class ratio, with 0% for unsupervised and 25%, 50%, and 75% for semi-supervised settings. The proposed method USNID is significantly better than others with p-value < 0.05 (†) and p-value < 0.1 (*) using t-test.

KCR	Methods	BANKING			CLINC150			StackOverflow		
		NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
0%	KM	49.30†	13.04†	28.62†	71.05†	27.72†	45.76†	19.87†	5.23†	23.72†
	AG	53.28†	14.64†	31.62†	72.21†	27.05†	44.13†	25.54†	7.12†	28.50†
	SAE-KM	59.80†	23.59†	37.07†	73.77†	31.58†	47.15†	44.96†	28.23†	49.11†
	DEC	62.65†	25.32†	38.60†	74.83†	31.71†	48.77†	58.76†	36.23†	59.49†
	DCN	62.72†	25.36†	38.59†	74.77†	31.68†	48.69†	58.75†	36.23†	59.48†
	CC	44.89†	9.75†	21.51†	65.79†	18.00†	32.69†	19.06†	8.79†	21.01†
	SCCL	63.89†	26.98†	40.54†	79.35†	38.14†	50.44†	69.11†	34.81†	68.15
	USNID	**75.30**	**43.33**	**54.83**	**91.00**	**68.54**	**75.87**	**72.00**	**52.25**	**69.28**
25%	KCL	52.70†	18.58†	26.03†	67.98†	24.30†	29.40†	30.42†	17.66†	30.69†
	MCL	47.88†	14.43†	23.29†	62.76†	18.21†	28.52†	26.68†	17.54†	31.46†
	DTC	55.59†	19.09†	31.75†	79.35†	41.92†	56.90†	29.96†	17.51†	29.54†
	GCD	59.74†	26.04†	38.50†	83.70†	52.23†	64.82†	29.69†	15.48†	34.84†
	CDAC+	66.39†	33.74†	48.00†	84.68†	50.02†	66.24†	46.16†	30.99†	51.61†
	DeepAligned	70.50†	37.62†	49.08†	88.97†	64.63†	74.07†	50.86†	37.96†	54.50†
	DCSC	78.18	49.75	60.15	91.70	72.68	79.89	-	-	-
	MTP-CLNN	80.04†	52.91†	65.06	93.17†	76.20†	**83.26**	73.35	54.80†	74.70
	USNID	**81.94**	**56.53**	**65.85**	**94.17**	**77.95**	83.12	**74.91**	**65.45**	**75.76**
50%	KCL	63.50†	30.36†	40.04†	74.74†	35.28†	45.69†	53.39†	41.74†	56.80†
	MCL	62.71†	29.91†	41.94†	76.94†	39.74†	49.44†	45.17†	36.28†	52.53†
	DTC	69.46†	37.05†	49.85†	83.01†	50.44†	64.39†	49.80†	37.38†	52.92†
	GCD	66.97†	35.07†	48.35†	87.12†	59.86†	70.89†	50.60†	31.98†	55.27†
	CDAC+	67.30†	34.97†	48.55†	86.00†	54.87†	68.01†	46.21†	30.88†	51.79†
	DeepAligned	76.67†	47.95†	59.38†	91.59†	72.56†	80.70†	68.28†	57.62†	74.52†
	DCSC	81.19	56.94	68.30	93.75	78.82	84.57	-	-	-
	MTP-CLNN	83.42†	60.17†	70.97*	94.30†	80.17†	86.18	76.66†	62.24†	80.36
	USNID	**85.05**	**63.77**	**73.27**	**95.45**	**82.87**	**87.22**	**78.77**	**71.63**	**82.06**
75%	KCL	72.75†	45.21†	59.12†	86.00†	58.62†	68.89†	63.98†	54.28†	68.69†
	MCL	74.42†	48.06†	61.56†	87.26†	61.21†	70.27†	63.44†	56.11†	71.71†
	DTC	74.44†	44.68†	57.16†	89.19†	67.15†	77.65†	63.05†	53.83†	71.04†
	GCD	72.48†	43.36†	57.32†	89.42†	65.98†	76.78†	61.99†	43.61†	66.73†
	CDAC+	69.54†	37.78†	51.07†	85.96†	55.17†	67.77†	58.23†	40.95†	64.57†
	DeepAligned	79.39†	53.09†	64.63†	93.92†	79.94†	86.79†	73.28†	60.09†	77.97†
	DCSC	84.65	64.55	75.18	95.28	84.41	89.70	-	-	-
	MTP-CLNN	86.19†	66.98†	77.22	95.45†	84.30†	89.46*	77.12†	69.36†	82.90†
	USNID	**87.41**	**69.54**	**78.36**	**96.42**	**86.77**	**90.36**	**80.13**	**74.90**	**85.66**

data from known classes and all data from new classes are unlabeled. To simulate real-world scenarios, the validation set only contains labeled data from known classes. The goal is to find known and discover new intent-wise clusters in the testing set.

We use the pre-trained BERT language model with 12 transformer layers as the backbone, which is implemented in [55]. For all experiments, we use AdamW [56] as the optimizer to train the model. The training process consists of 100 epochs, with a batch size of 128, and learning rates searched from {1e-5, 2e-5, 5e-5}. The intent feature dimension D is 768. All the non-linear projection heads $\{f_i^u\}_{i=1}^3$ and $\{f_i^s\}_{i=1}^4$ have the same architecture of $W\sigma(\cdot)+b$, where W and b are the weight matrix and bias term of a single linear layer, and σ is the Tanh activation function.

For the unsupervised setting, the temperature τ and *random erase* ratio a are set to 0.07 and 0.5, respectively. In addition, we fine-tune with the parameters of the last transformer layer as suggested in [3], [10], which can improve training efficiency and maintain good performance. For the semi-supervised setting, τ and a are set to {0.05, 0.4} for BANKING and StackOverflow, and {0.1, 0.3} for CLINC150. The pre-training stage follows the same fine-tuning strategy as in the unsupervised setting, while the clustering and representation learning stage fine-tunes with all the transformer layers as in [13], which can fully explore high-level semantics with the guidance of labeled data. The K-Means++ clustering algorithm is implemented with the Scikit-learn [57] toolkit. The threshold δ_{th} for stopping the training procedure is set to 0.0005. We implement our approach in PyTorch 1.8.1 and run experiments on NVIDIA Geforce RTX 3090 GPUs. For all experiments, we report the averaged results over ten runs with random seeds of 0-9. All the baselines are built upon our TEXTOIR platform [58].

6 Results and Discussion

6.1 Results of New Intent Discovery

The main experimental results of unsupervised and semi-supervised new intent discovery are presented in Table 2. We highlight the best results for each setting (KCR=0%, 25%, 50%, and 75%) in bold and conduct significance t tests between our method (USNID) and the other baselines[2].

In unsupervised new intent discovery (KCR=0%), traditional clustering methods (e.g., KM and AG) show the lowest performance across all datasets, mainly due to their inability to comprehend complicated semantics using static feature-engineering representations. On the contrary, deep clustering methods demonstrate superior performance (over 10% score improvement on ARI on BANKING and StackOverflow datasets) by learning representations end-to-end with deep neural networks during clustering. While CC incorporates instance-level and cluster-level contrastive learning techniques, its limitations in capturing cluster-level relations without learning specific targets lead to performance inferior to even some traditional methods. SCCL, the current SOTA unsupervised method in NLP, improves performance by replacing cluster-level contrastive learning with a technique forcing each sample to learn from high-confidence constructed soft targets via KL-divergence. Yet,

[2] Since DCSC is not open source, we only report the results as in [12].

TABLE 3
Ablation studies of USNID. ”w / o” means removing a component of USNID. Detailed information of each component can be seen in section 6.2.

KCR	Stage 1	Stage 2	BANKING			CLINC150			StackOverflow		
			NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
0%	w/o UCL	Full	66.69	30.17	41.65	84.73	50.76	62.00	36.61	18.06	31.89
	Full	K-Means++	62.16	27.97	40.69	77.56	38.36	53.26	22.08	9.05	23.66
		w/o CGM	64.53	30.03	42.17	82.68	48.42	60.57	20.20	6.66	19.94
		w/o CE	72.52	39.45	50.36	87.67	59.64	68.89	41.79	21.50	39.48
		Full	**75.30**	**43.33**	**54.83**	**91.00**	**68.54**	**75.87**	**72.00**	**52.25**	**69.28**
25%	w/o Semi-SCL	Full	80.06	52.01	61.43	92.99	73.24	78.85	71.04	60.73	71.52
	Full	K-Means++	65.99	33.71	48.68	83.21	52.16	65.13	48.10	33.66	53.93
		w/o Self-Sup	71.39	41.00	55.12	89.75	67.41	76.76	57.44	43.05	63.85
		w/o CGM	72.25	42.32	55.54	90.06	67.66	76.17	54.78	42.46	60.12
		w/o Sup-SCL	80.16	52.20	63.01	93.78	76.63	82.38	72.10	61.56	73.13
		Full	**81.94**	**56.53**	**65.85**	**94.17**	**77.95**	**83.12**	**74.91**	**65.45**	**75.76**
50%	w/o Semi-SCL	Full	83.83	59.90	68.63	94.52	79.14	84.28	76.74	69.80	80.36
	Full	K-Means++	73.60	45.48	58.95	87.33	62.64	73.60	58.08	44.72	64.91
		w/o Self-Sup	77.20	51.28	64.32	92.59	75.72	83.42	68.01	51.64	72.40
		w/o CGM	79.67	55.32	67.18	92.95	76.47	83.41	67.19	58.64	75.02
		w/o Sup-SCL	83.54	59.87	69.39	94.67	80.34	85.64	77.18	68.79	80.00
		Full	**85.05**	**63.77**	**73.27**	**95.48**	**82.99**	**87.28**	**78.77**	**71.63**	**82.06**
75%	w/o Semi-SCL	Full	86.70	67.32	75.91	96.10	85.14	88.99	79.03	73.78	84.17
	Full	K-Means++	78.06	53.89	67.29	90.24	70.05	79.29	68.10	54.93	74.78
		w/o Self-Sup	81.80	60.33	72.60	94.84	82.79	88.41	73.57	57.51	78.00
		w/o CGM	83.65	63.52	74.51	95.07	83.14	88.23	74.78	67.75	81.87
		w/o Sup-SCL	85.91	66.05	75.61	95.70	84.22	88.86	78.58	72.06	83.52
		Full	**87.41**	**69.54**	**78.36**	**96.42**	**86.77**	**90.36**	**80.13**	**74.90**	**85.66**

TABLE 4
Cluster number estimation results in unsupervised and semi-supervised settings on the three datasets.

		BANKING		CLINC150		StackOverflow	
KCR	Methods	K	Error	K	Error	K	Error
0%	USNID	**74.00**	**3.90**	**137.80**	**8.13**	**16.80**	**16.00**
25%	DTC	42.30	45.06	108.20	27.87	9.50	52.50
	DeepAligned	63.50	17.53	122.00	18.67	16.60	17.00
	USNID	**74.30**	**3.51**	**139.60**	**6.93**	**16.80**	**16.00**
50%	DTC	83.40	8.31	157.50	5.00	**18.90**	**5.50**
	DeepAligned	65.10	15.45	125.60	16.27	11.40	43.00
	USNID	**77.50**	**0.65**	**143.20**	**4.53**	18.30	8.50
75%	DTC	112.00	45.45	218.00	45.33	27.10	35.50
	DeepAligned	68.83	10.61	128.60	14.27	16.60	17.00
	USNID	**82.80**	**7.53**	**154.50**	**3.00**	**18.90**	**5.50**

our method, USNID, outperforms SCCL by 16.35%, 30.14%, and 17.44% in ARI scores on the BANKING, CLINC150, and StackOverflow datasets, respectively, and shows improvements of over 10% in all three metrics on the BANKING and CLINC150 datasets.

In semi-supervised new intent discovery (KCR$\neq$0%), various methods (e.g., KCL, MCL, GCD, and CDAC+) construct pairwise similarity relations and use them to learn friendly representations for clustering by pulling similar and repelling dissimilar pairs. DTC constructs the target distribution as in SCCL and extends it by incorporating temporal information and consistent constraints. Despite this, their pairwise constraints have weak correlations, resulting in difficulties in distinguishing complex semantic intent-wise groups. In contrast, DeepAligned and DCSC, which use alignment strategies to generate categorical discrimination pseudo-labels, show significant improvements of over 10% scores in ARI. MTP-CLNN is the existing SOTA method in the semi-supervised setting, intensifies the pairwise constraints by including additional similarity connections in the nearest neighbor space. However, this could lead to unstable clustering targets as the representation learning process progresses. Our method, USNID, uses relatively stable targets guided by cluster centroids and achieves better results. Notably, USNID outperforms MTP-CLNN in ARI by 10.65%, 9.39%, and 5.54% with 25%, 50%, and 75% known classes on the StackOverflow dataset, respectively and shows significant improvements in NMI and ARI by over 1% across nearly all settings.

Interestingly, the performance of unsupervised USNID outperforms more than half of the semi-supervised clustering methods with 75% known classes. This is attributed to two main factors: (1) USNID, unlike other methods only using limited labeled data for pre-training, it strongly augments all samples and applies unsupervised contrastive learning, providing superiorly initialized representations and avoiding potential overfitting problem. (2) It uses a centroid-guided mechanism to create specific pseudo-labels rather than ambiguous pairwise relations as clustering targets, which brings categorical information to guide class differentiation.

The clustering performance is notably superior on the CLINC150 dataset compared to the BANKING dataset, primarily due to their inherent differences. The CLINC150 dataset contains utterances from 10 general domains, enabling better distinction of intent classes. In contrast, the BANKING dataset, originates from a singular banking domain, limiting the use of diverse semantic backgrounds for intent detection. Additionally, the BANKING dataset presents overlapping intent categories and contains longer text sequences with complex semantics, thus posing increased challenges to the model’s capability to generalize effectively. The performance improves as the number of known classes increases, indicating the positive influence of labeled data on clustering. Semi-supervised USNID achieves top-tier results across all settings and substantial improvements over all baselines and unsupervised USNID, highlighting the advantage of our method in utilizing labeled data for new intent discovery.

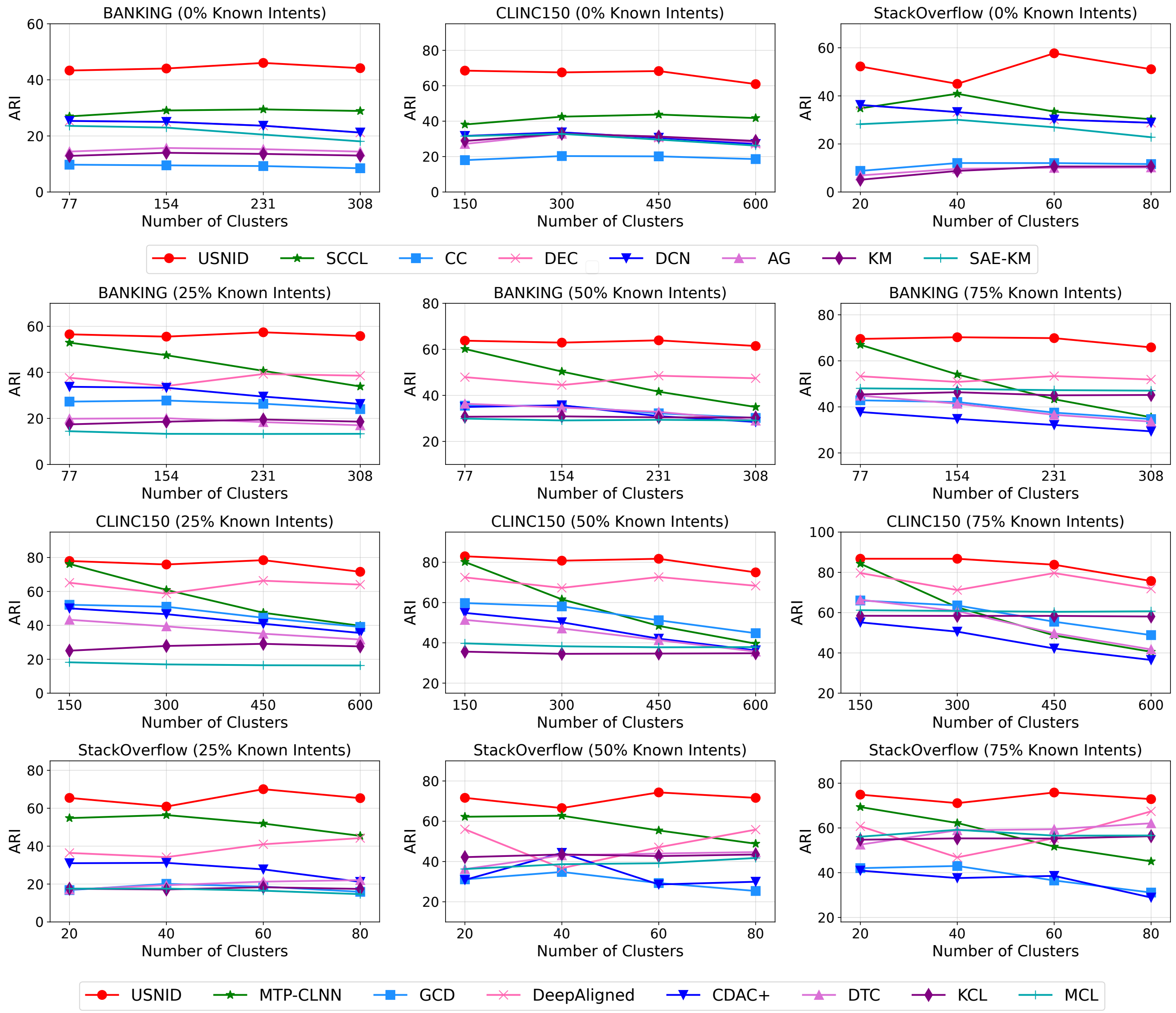

Fig. 4. Unsupervised and semi-supervised new intent discovery results with different cluster numbers on the three datasets.

6.2 Ablation Studies

To validate the effectiveness of the components in USNID, we conduct comprehensive ablation studies and show the results in Table 3. USNID has two stages: pre-training (stage 1) and clustering and representation learning (stage 2). For stage 1, removing the unsupervised contrastive loss (*w/o UCL*) results in a 13-34% drop in ARI, while removing the semi-supervised contrastive loss (*w/o Semi-SCL*) causes a 1-4% decrease across three datasets. These results highlight the importance of pre-training in generating well-initialized representations for clustering. In stage 2, directly performing K-Means++ after stage 1 causes a 15-43% and 15-31% absolute ARI decrease in unsupervised and semi-supervised settings, respectively. This underscores the significance of this stage. Without the centroid-guided mechanism (*w/o CGM*), performing K-Means++ once and using its pseudo-labels as targets for representation learning leads This suggests that using historical centroids as guidance for updating pseudo-labels effectively constructs high-quality self-supervised signals for feature learning. Furthermore, removing the cross-entropy loss (*w/o CE*) in the unsupervised setting leads to a 2-30% decrease across all datasets. In the semi-supervised setting, removing the self-supervised learning loss (*w/o Self-Sup*) results in decreases of 5-15%, 1-10%, and 6-22% in various KCR settings across all datasets. This implies that specific pseudo-labels significantly improve clustering performance over the pairwise constraints of the contrastive loss. Lastly, omitting the additional supervised contrastive loss (*w/o Sup-SCL*) in the semi-supervised setting leads to a 1-4% drop in ARI across all KCR settings of the three datasets. This shows that *Sup-SCL* mitigates the *catastrophic forgetting* problem and better uses labeled data.

6.3 Cluster Number Estimation

In this section, we explore new intent discovery in a more challenging situation where the ground truth number of

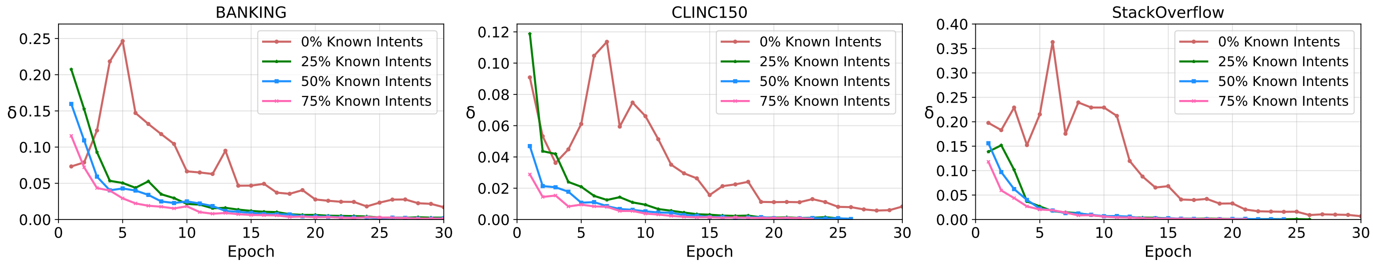

Fig. 5. The convergence curves of USNID in unsupervised and semi-supervised settings on the three datasets.

clusters is not known in advance. As suggested in [10], the initial cluster number is set to large values (i.e., twice the ground truth number) of 154, 300, and 40 for BANKING, CLINC150, and StackOverflow, respectively. We compare our approach (described in section 4.4) with two strong baselines for estimating the number of clusters, as proposed in [9], [10]. To evaluate the accuracy of the estimated cluster numbers, we compute the error between the average of the estimated K (obtained from ten runs of experiments) and the ground truth number, with the lower error being better. The results are shown in Table 4.

Our method consistently achieves the lowest errors on all three datasets in semi-supervised settings, with the exception of the 50% KCR setting on the StackOverflow dataset. Though DTC performs well on the 50% KCR setting, it is unstable with different amounts of labeled data and performs worse than the other two methods in the 25% and 75% settings on all three datasets. DeepAligned is a preliminary version of our method that also predicts K by removing low-confidence clusters. However, it ignores the use of prior knowledge of labeled data to induce the known intent clusters, which may be falsely dropped under the assumption of high-quality cluster selection. As a result, it usually yields lower prediction results with 4-14%, 11-17%, and 1-34% higher errors than our method on the three datasets for all KCR settings. Interestingly, USNID even exhibits competitive performance in the unsupervised setting, suggesting that it can also benefit from the weak semantic similarity relations learned through unsupervised contrastive learning.

6.4 Effect of the Number of Clusters

In this section, we study the impact of the cluster number on the performance of new intent discovery. As suggested in [3], [10], we vary the number of clusters in the range of one to four times the ground truth number. The last three settings correspond to open-world scenarios for discovering new intents, as is often the case in real applications. We use ARI as the metric and show the results in Figure 4.

In unsupervised new intent discovery, USNID consistently demonstrates the best performance with significant improvements over the other methods on all three datasets. It also maintains robust performance with small fluctuations even when the assigned cluster number is large. We observe that most unsupervised baselines are not sensitive to the number of clusters, especially on the BANKING and CLINC150 datasets. However, their performance remains poor compared to the ground truth number, and there is still a significant gap between them and our method.

In semi-supervised new intent discovery, USNID also outperforms all other baselines on the three datasets. It is particularly more robust than other methods, and its performance is only slightly affected by the number of clusters. In contrast, while MTP-CLNN performs well when using the ground truth number, it is extremely sensitive to the cluster number. Its performance drops dramatically as the cluster number increases, resulting in much lower performance than USNID. DeepAligned is relatively more robust among these baselines as it uses a similar strategy to estimate the cluster number as our method.

6.5 Convergence Analysis

In this section, we analyze the convergence of USNID by tracking the variation of the cluster allocation difference δ (described in section 4.2.2) over the number of training epochs. The results are depicted in Figure 5.

It can be observed that, when using labeled data as prior knowledge, our method stably and efficiently converges to a small threshold (i.e., 0.0005) within a few epochs (around 25) on all three datasets. This demonstrates the usefulness of labeled data in guiding the clustering performance. Furthermore, increasing the amount of labeled data from 25% to 75% of known intents leads to a lower δ and generally faster convergence time. In the more challenging unsupervised setting, although there are fluctuations in the first few epochs, our method is still able to converge gradually to a small value. We suggest that this is because the provided aligned targets, although not guided by any prior knowledge, can still produce high-quality, consistent targets despite potentially introducing some noise.

7 Conclusions and Future Work

In this paper, we address the problem of new intent discovery in both unsupervised and semi-supervised settings, which is relevant to real-world scenarios. We propose USNID, a novel clustering framework that utilizes several key techniques to tackle this problem. First, USNID captures elementary semantic features through a pre-training stage by learning similarity information with self-augmented samples or limited labeled data, which has been shown to enhance the subsequent clustering. Second, we introduce a novel centroid-guided clustering mechanism to address the issue of inconsistent cluster assignments in partition-based methods during multiple clustering. This method obtains aligned targets by initializing the current clustering with the cluster centroids from the previous clustering, which demonstrates efficient convergence. Third, USNID learns fine-grained intent-specific group characteristics by jointly learning cluster-level and instance-level information with the targets of aligned pseudo-labels from the previous iteration's clustering, which significantly improves performance with clustering-friendly representations. Incorporating high-quality prior knowledge from labeled data has also been shown to bring additional benefits. When evaluated on several intent benchmarks, USNID outperforms all other methods in unsupervised and semi-supervised settings by a significant margin. Furthermore, we propose an effective method for estimating the number of clusters, which helps maintain robust performance in realistic scenarios without prior knowledge of the number of new classes.

In this study, our approach still depends on the specification of a large cluster count, an aspect that is somewhat reliant on empirical experience. In our future research, we intend to investigate the possibility of automatically determining the number of clusters with minimal assumptions. Additionally, our proposed framework is currently offline and necessitates the completion of clustering across all data. We foresee the potential for future research to extend our methodology to an online strategy, which could prove more efficient and applicable to larger datasets. Finally, while we currently employ K-Means++, there might be potential for exploring more efficient and effective centroid-based clustering methods. By addressing these areas for improvement, we aspire to expand the capabilities of our framework, ultimately enhancing both its applicability and efficiency.

References

[1] L. Qin, T. Xie, W. Che, and T. Liu, "A survey on spoken language understanding: Recent advances and new frontiers," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 4577–4584.

[2] Y. Li, C. Gao, X. Du, H. Wei, H. Luo, D. Jin, and Y. Li, "Automatically discovering user consumption intents in meituan," in *Proc. 28th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2022, pp. 3259–3269.

[3] T.-E. Lin, H. Xu, and H. Zhang, "Discovering new intents via constrained deep adaptive clustering with cluster refinement," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 8360–8367.

[4] H. Li, X. Wang, Z. Zhang, J. Ma, P. Cui, and W. Zhu, "Intention-aware sequential recommendation with structured intent transition," *IEEE Trans. Knowl. Data Eng.*, pp. 5403–5414, 2022.

[5] J. Schuurmans and F. Frasincar, "Intent classification for dialogue utterances," *IEEE Trans. Intell. Transp. Syst.*, pp. 82–88, 2019.

[6] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, "Autonovel: Automatically discovering and learning novel visual categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 6767–6781, 2021.

[7] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, and E. Ricci, "A unified objective for novel class discovery," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9284–9292.

[8] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Generalized category discovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7492–7501.

[9] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8400–8408.

[10] H. Zhang, H. Xu, T.-E. Lin, and R. Lyu, "Discovering new intents with deep aligned clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 14 365–14 373.

[11] R. Kumar, M. Patidar, V. Varshney, L. Vig, and G. Shroff, "Intent detection and discovery from user logs via deep semi-supervised contrastive clustering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 1836–1853.

[12] F. Wei, Z. Chen, Z. Hao, F. Yang, H. Wei, B. Han, and S. Guo, "Semi-supervised clustering with contrastive learning for discovering new intents," *arXiv: 2201.07604*, 2022.

[13] Y. Zhang, H. Zhang, L.-M. Zhan, X.-M. Wu, and A. Lam, "New intent discovery with pre-training and contrastive learning," in *Proc. 60th Assoc. Comput. Linguistics*, 2022, pp. 256–269.

[14] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability.*, 1967, pp. 281–297.

[15] K. C. Gowda and G. Krishna, "Agglomerative clustering using the concept of mutual nearest neighbourhood," *Pattern Recognit.*, pp. 105–112, 1978.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, pp. 264–323, 1999.

[17] E. H. Ruspini, "A new approach to clustering," *Inf. Technol. Control.*, pp. 22–32, 1969.

[18] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proc. SIAM Int. Conf. Data Mining*, 2007, pp. 1027–1035.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, pp. 436–444, 2015.

[20] Y. Ren, J. Pu, Z. Yang, J. Xu, G. Li, X. Pu, P. S. Yu, and L. He, "Deep clustering: A comprehensive survey," *arXiv: 2210.04142*, 2022.

[21] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *J. Mach. Learn. Res.*, pp. 3371–3408, 2010.

[23] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3861–3870.

[24] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5879–5887.

[25] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 132–149.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[27] S. Basu, I. Davidson, and K. Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. USA:CRC Press, 2008.

[28] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 577–584.

[29] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proc. SIAM Int. Conf. Data Mining*. SIAM, 2004, pp. 333–344.

[30] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 81–88.

[31] Y.-C. Hsu, Z. Lv, and Z. Kira, "Learning to cluster in order to transfer across domains and tasks," in *Proc. Int. Conf. Learn. Representations*, 2018.

[32] Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, and Z. Kira, "Multi-class classification without multi-class labels," in *Proc. Int. Conf. Learn. Representations*, 2019.

[33] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multim.*, vol. 21, no. 5, pp. 1276–1288, 2019.

[34] D. Yuan, X. Chang, Z. Li, and Z. He, "Learning adaptive spatial-temporal context-aware correlation filters for uav tracking," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 3, 2022.

[35] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, "Automatically discovering and learning new visual categories with ranking statistics," in *Proc. Int. Conf. Learn. Representations*, 2020.

[36] I. Casanueva, T. Temcinas, D. Gerz, M. Henderson, and I. Vulic, "Efficient intent detection with dual sentence encoders," *arXiv: 2003.04807*, 2020.

[37] H. Zhang, H. Xu, X. Wang, Q. Zhou, S. Zhao, and J. Teng, "Mintrec: A new dataset for multimodal intent recognition," in *Proc. of the 30th ACM Int. Conf. on Multimedia*, 2022, pp. 1688–1697.

[38] H. Zhang, H. Xu, and T.-E. Lin, "Deep open intent classification with adaptive decision boundary," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 14 374–14 382.

[39] H. Zhang, H. Xu, S. Zhao, and Q. Zhou, "Learning discriminative representations and decision boundaries for open intent detection," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 31, pp. 1611–1623, 2023.

[40] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Advances. Neural Inf. Proces. Syst.*, 2020, pp. 9912–9924.

[41] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Proc. Advances Neural Inf. Process. Syst.*, pp. 2292–2300, 2013.

[42] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2019, pp. 1311–1316.

[43] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "Scan: Learning to classify images without labels," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 268–285.

[44] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[45] D. Zhang, F. Nan, X. Wei, S.-W. Li, H. Zhu, K. R. McKeown, R. Nallapati, A. O. Arnold, and B. Xiang, "Supporting clustering with contrastive learning," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 5419–5430.

[46] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6688–6697.

[47] H. W. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, pp. 83–97, 2010.

[48] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, and X. Peng, "Twin contrastive learning for online clustering," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2205–2221, 2022.

[49] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2020, pp. 18 661–18 673.

[50] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 6894–6910.

[51] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, and H. Hao, "Short text clustering via convolutional neural networks," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2015.

[52] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8547–8555.

[53] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[54] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2019, pp. 3982–3992.

[55] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 38–45.

[56] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[58] H. Zhang, X. Li, H. Xu, P. Zhang, K. Zhao, and K. Gao, "TEXTOIR: An integrated and visualized platform for text open intent recognition," in *Proc. 59th Assoc. Comput. Linguistics.*, 2021, pp. 167–174.

Hanlei Zhang received the B.S. degree from the Department of Computer Science and Technology, Beijing Jiaotong University, Beijing, China, in 2020. He is currently working toward the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He has authored or coauthored six papers in top-tier international conferences and journals, including AAAI, ACM MM, ACL, and IEEE/ACM Transactions on Audio, Speech and Language Processing. His research interests include intent analysis, open world classification, clustering, multimodal language understanding, and natural language processing.

Hua Xu received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1998, and the M.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2000 and 2003, respectively. He is a Tenured Associate Professor with the Department of Computer Science and Technology, Tsinghua University. He has authored or coauthored more than 130 peer-reviewed papers in top-tier international journals and conferences. His research interests include multi-modal intelligent information processing for natural interaction of service robots, evolutionary learning, and intelligent optimization. Prof. Xu was the recipient of the Second Prize from the National Science and Technology Progress of China, First Prize from Beijing Science and Technology, and Third Prize from Chongqing Science and Technology.

Xin Wang received the B.S. degree in 2020 from the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, China, where he is currently working toward the M.S. degree with the School of Information Science and Engineering. He has authored or coauthored one paper in the ACM MM international conference. His research interests include unsupervised learning, semi-supervised learning, and natural language processing.

Fei Long is working toward the undergraduate degree in the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include natural language processing, clustering, and machine learning.

Kai Gao received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China. He is a Professor in the School of Information Science and Engineering, Hebei University of Science and Technology. He has authored or coauthored over 60 academic papers in international conferences and journals. His research interests include natural language processing, knowledge discovery, and multimodal intelligent information processing.