Discovering New Intents with Deep Aligned Clustering

Hanlei Zhang,[1, 2] **Hua Xu,**[1, 2*] **Ting-En Lin**[3], **Rui Lv**[4]
[1]State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China,
[2] Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, China
[3] DAMO Academy, Alibaba Group, [4] Beijing University of Posts and Telecommunications University, Beijing 100876, China
zhang-hl20@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn, ting-en.lte@alibaba-inc.com, lvrui2017@bupt.edu.cn

Abstract

Discovering new intents is a crucial task in dialogue system. Most existing methods are limited in transferring the prior knowledge from known intents to new intents. These methods also have difficulties in providing high-quality supervised signals to learn clustering-friendly features for grouping unlabeled intents. In this work, we propose an effective method (Deep Aligned Clustering) to discover new intents with the aid of limited known intent data. Firstly, we leverage a few labeled known intent samples as prior knowledge to pre-train the model. Then, we perform k-means to produce cluster assignments as pseudo-labels. Moreover, we propose an alignment strategy to tackle the label inconsistency during clustering assignments. Finally, we learn the intent representations under the supervision of the aligned pseudo-labels. With an unknown number of new intents, we predict the number of intent categories by eliminating low-confidence intent-wise clusters. Extensive experiments on two benchmark datasets show that our method is more robust and achieves substantial improvements over the state-of-the-art methods.[1]

1 Introduction

Discovering novel user intents is crucial for improving the service quality in dialogue system. By analyzing the discovered new intents, we may find underlying user interests to provide more business opportunities and guide the improvement direction for the dialogue system (Lin and Xu 2019).

Intent discovery has attracted much attention in recent years (Perkins and Yang 2019; Min et al. 2020; Vedula et al. 2020). Many researchers regard it as an unsupervised clustering problem and manage to incorporate some weak supervised signals to guide the clustering process. For example, Hakkani-Tür et al. (2013) propose a hierarchical semantic clustering model and collect web page clicked information as implicit supervision for intent discovery. Hakkani-Tür et al. (2015) utilize a semantic parsing graph as extra knowledge to mine novel intents during clustering. Padmasundari and Bangalore (2018) benefit from the consensus predictions of multiple clustering techniques to discover similar semantic intent-wise clusters. Haponchyk et al. (2018) cluster questions into user intent categories under the supervision of structured outputs. Shi et al. (2018) extract intent features with an autoencoder and automatically label the intents with a hierarchical clustering method.

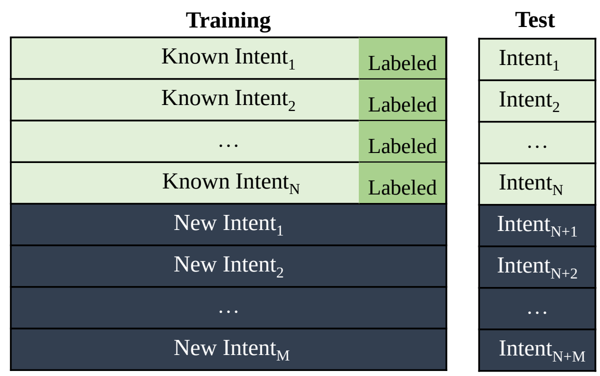

Figure 1: An example for our task. We use limited known intent labeled data as a guide to discover new intents.

However, all of the above methods fail to leverage the prior knowledge of known intents. These methods assume that the unlabeled samples are only composed of undiscovered new intents. A more common case that some labeled data of known intents are accessible and the unlabeled data are mixed with both known and new intents. As illustrated in Figure 1, we may have a few labeled samples (e.g., with a labeled proportion of 10%) of known intents in advance. The remaining known and new intent samples are all unlabeled. Our goal is to find known intents and discover new intents with the prior knowledge of limited labeled data. Our previous work CDAC+ (Lin, Xu, and Zhang 2020) directly tackles this problem. But it uses pairwise similarities as weak supervised signals, which are ambiguous to distinguish a mixture of unlabeled known and new intents. Thus, the performance drops with more new intents.

To summarize, there are two main difficulties in our task. On the one hand, it is challenging to effectively transfer the prior knowledge from known intents to new intents with lim-

*Hua Xu is the corresponding author.

[1]Code available at https://github.com/hanleizhang/DeepAligned-Clustering

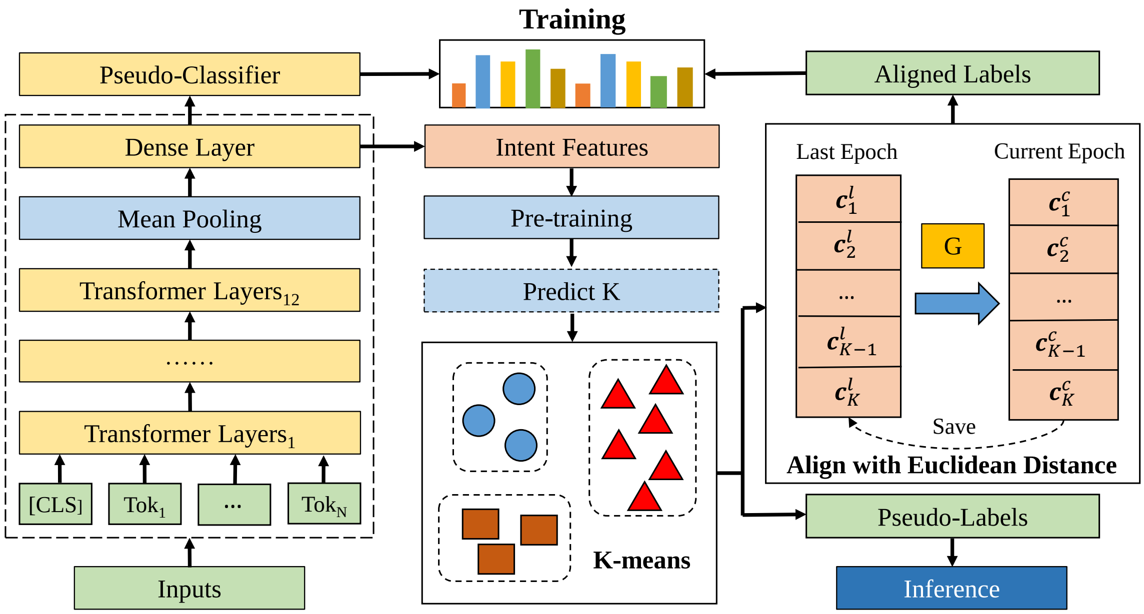

Figure 2: The model architecture of our approach. Firstly, we extract intent features with BERT. We pre-train the model under the supervision of few labeled samples, and predict the cluster number K if we do not know in advance. Then, we perform k-means to produce cluster centroids and use cluster assignments as pseudo-labels. Next, we align the obtained centroids in the current training epoch $\{c_i^c\}_{i=1}^K$ with the saved centroids in the last epoch $\{c_i^l\}_{i=1}^K$, and produce the alignment projection G. Finally, we use G on the pseudo-labels to produce the aligned labels for self-supervised learning.

ited labeled data. On the other hand, it is hard to construct high-quality supervised signals to learn friendly representations for clustering both unlabeled known and new intents.

To solve these problems, we propose an effective method to leverage the limited prior knowledge of intents and provide high-quality supervised signals for feature learning. As illustrated in Figure 2, we firstly use the pre-trained BERT model (Devlin et al. 2019) to extract deep intent features. Then, we pre-train the model with the limited labeled data under the supervision of the softmax loss. We retain the pre-trained parameters and use the learning information to obtain well-initialized intent representations. Next, we perform clustering on the extracted intent features and estimate the cluster number K (unknown beforehand) by eliminating the low-confidence clusters.

As most of the training samples are unlabeled, we propose an original alignment strategy to construct high-quality pseudo-labels as supervised signals for learning discriminative intent features. For each training epoch, we firstly perform k-means on the extracted intent features, and then use the produced cluster assignments as pseudo-labels for training the neural network. However, the inconsistent assigned labels cannot be directly used as supervised signals, so we use the cluster centroids as the targets to obtain the alignment mapping between pseudo-labels in consequent epochs. Finally, we perform k-means again for inference. Benefit from the relatively consistent aligned targets, our method can inherit the history learning information and boost the clustering performance.

We summarize our contributions as follows. Firstly, we propose a simple method that successfully generalizes to mass of new intents and estimate the number of novel classes with limited prior knowledge of known intents. Secondly, we propose an effective alignment strategy to obtain high-quality self-supervised signals, which learns discriminative features for distinguishing both known and new intents. Finally, extensive experiments on two benchmark datasets show our approach yields better and more robust results than the state-of-the-art methods.

2 Related Work

2.1 Intent Modeling

Many researchers try modeling user intents in dialogue system in recent years. A line for these works is to enrich the intent information jointly with other tasks, such as sentiment classification (Qin et al. 2020), slot filling (Qin et al. 2019; Goo et al. 2018; Wang, Shen, and Jin 2018). Another line is to leverage hidden semantic information to construct supervised signals for intent representation learning (Shi et al. 2018; Brychcin and Král 2017; Hakkani-Tür et al. 2013). In this work, we follow the second line to model intents.

2.2 Unsupervised Clustering

There are many classical unsupervised clustering methods, such as partition-based methods (MacQueen et al. 1967), hierarchical methods (Gowda and Krishna 1978) and density-based methods (Ester et al. 1996). However, the high-dimensional pattern representations suffer from high computational complexity and poor performance. Though some feature dimensionality reduction (Gowda 1984) and data transformation methods (Wold, Esbensen, and Geladi 1987) have been proposed, these methods still can not capture high-level semantics of intent features (Lin and Xu 2019).

Deep Clustering With the development of deep learning, researchers adopt deep neural networks (DNNs) to extract friendly features for clustering. The joint unsupervised learning (JULE) (Yang, Parikh, and Batra 2016) combines deep feature learning with hierarchical clustering but needs huge computational and memory cost on large-scale datasets. Deep Embedded Clustering (DEC) (Xie, Girshick, and Farhadi 2016) trains the autoencoder with the reconstruction loss and iteratively refines the cluster centers by optimizing KL-divergence with an auxiliary target distribution. Compared with DEC, Deep Clustering Network (DCN) (Yang et al. 2017) further introduces a k-means loss as the penalty term to reconstruct the clustering loss. Deep Adaptive Image Clustering (DAC) (Chang et al. 2017) utilizes the pairwise similarities as the learning targets and adopts an adaptive learning algorithm to select samples for training. However, all these clustering methods cannot provide specific supervised signals for representation learning.

DeepCluster (Caron et al. 2018) benefits from the structured outputs to boost the discriminative power of the convolutional neural network (CNN). It alternately performs k-means and representation learning. It considers the cluster assignments as pseudo-labels, which are explicit supervised signals for grouping each class. However, it needs to reinitialize the classifier parameters randomly before each training epoch. To deal with this issue, we propose an alignment strategy to produce aligned pseudo-labels for self-supervised learning without reinitialization.

2.3 Constrained Clustering

Although there are various unsupervised clustering methods, the performances of these methods are still limited without prior knowledge for guiding the clustering process. Therefore, researchers perform constrained clustering with the aid of some labeled data.

Classical constrained clustering methods use the pairwise information as constraints for guiding the representation learning and clustering process. COP-KMeans (Wagstaff et al. 2001) uses instance-level constraints (must-link and cannot-link) and modifies k-means to satisfy these constraints. PCK-means (Basu, Banerjee, and Mooney 2004) presents a framework for pairwise constrained clustering, and it further selects informative pairwise constraints with an active learning method. MPCK-means (Bilenko, Basu, and Mooney 2004) incorporates the metric-learning approach into PCK-means and combined the centroid-based methods and metric-based methods into a unified framework. However, these methods need huge computational cost by enumerating pairwise conditions.

KCL (Hsu, Lv, and Kira 2018) uses deep neural networks to perform pairwise constraint clustering. It firstly trains an extra network for binary similarity classification with a labeled auxiliary dataset. Then, it transfers the prior knowledge of pairwise similarity to the target dataset and uses KL-divergence to evaluate the pairwise distance. MCL (Hsu et al. 2019) uses the meta classification likelihood as the criterion to learn pairwise similarities. However, the domain adaptation methods are still limited in our task. CDAC+ (Lin, Xu, and Zhang 2020) is specifically designed for discovering new intents. It uses limited labeled data as a guide to learn pairwise similarities. However, it is limited in providing specific supervised signals and fails to estimate the number of novel classes. DTC (Han, Vedaldi, and Zisserman 2019) is a method for discovering novel classes in computer vision. It improves the DEC algorithm and transfers the knowledge of labeled data to estimate the number of novel classes. However, the amount of the labeled data has a great influence on its performance.

3 Our Approach

In this section, we will describe the proposed method in detail. As shown in Figure 2, we firstly extract intent representations with BERT. Then, we transfer the knowledge from known intents with limited labeled data. Finally, we propose an alignment strategy to provide self-supervised signals for learning clustering-friendly representations.

3.1 Intent Representation

The pre-trained BERT model demonstrates its remarkable effect in NLP tasks (Devlin et al. 2019), so we use it to extract deep intent representations. Firstly, we feed the i^{th} input sentence $\boldsymbol{s}_i$ to BERT, and take all its token embeddings $[CLS, T_1, \cdots, T_M] \in \mathbb{R}^{(M+1)\times H}$ from the last hidden layer. Then, we apply mean-pooling to get the averaged sentence feature representation $\boldsymbol{z}_i \in \mathbb{R}^H$:

$$\boldsymbol{z}_i = \text{mean-pooling}([CLS, T_1, \cdots, T_M]), \tag{1}$$

where CLS is the vector for text classification, M is the sequence length, and H is the hidden size. To further enhance the feature extraction capability, we add a dense layer h to get the intent feature representation $\boldsymbol{I}_i \in \mathbb{R}^D$:

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

where D is the dimension of the intent representation, σ is the Tanh activation function, $W_h \in \mathbb{R}^{H\times D}$ is the weight matrix and $b_h \in \mathbb{R}^D$ is the corresponding bias term.

3.2 Transferring Knowledge from Known Intents

To effectively transfer the knowledge, we use the limited labeled data to pre-train the model and leverage the well-trained intent features to estimate the number of clusters for clustering.

Pre-training We hope to incorporate the limited prior knowledge to obtain a good representation initialization for grouping both known and novel intents. As suggested in (Han, Vedaldi, and Zisserman 2019), we capture such intent feature information by pre-training the model with the labeled data. Specifically, we learn the feature representations under the supervision of the cross-entropy loss. After pre-training, we remove the classifier and use the rest of the network as the feature extractor in the subsequent unsupervised clustering.

Dataset	#Classes (Known + Unknown)	#Training	#Validation	#Test	Vocabulary	Length (max / mean)
CLINC	150 (113 + 37)	18,000	2,250	2,250	7,283	28 / 8.31
BANKING	77 (58 + 19)	9,003	1,000	3,080	5,028	79 / 11.91

Table 1: Statistics of CLINC and BANKING dataset. # indicates the total number of sentences. In each run of the experiment, we randomly select 75% intents as known intents. Taking the CLINC dataset as an example, we randomly select 113 known intents and treat the remaining 37 intents as new intents.

Predict K In real scenarios, we may not always know the number of new intent categories. In this case, we need to determine the number of clusters K before clustering. Therefore, we propose a simple method to estimate K with the aid of the well-initialized intent features.

We assign a big K' as the number of clusters (e.g., two times of the ground truth number of classes) at first. As a good feature initialization is helpful for partition-based methods (e.g., k-means) (Platt et al. 1999), we use the well pre-trained model to extract intent features. Then, we perform k-means with the extracted features. We suppose that real clusters tend to be dense even with K'. And the size of more confident clusters is larger than some threshold t. We drop the low confidence clusters which size smaller than t, and calculate K with:

$$K = \sum_{i=1}^{K'} \delta(|S_i| >= t), \tag{3}$$

where $|S_i|$ is the size of the i^{th} produced cluster, and $\delta(condition)$ is an indicator function. It outputs 1 if $condition$ is satisfied, and outputs 0 if not. Notably, we assign the threshold t as the expected cluster mean size $\frac{N}{K'}$ in this formula.

3.3 Deep Aligned Clustering

After transferring knowledge from known intents, we propose an effective clustering method to find unlabeled known classes and discover novel classes. We firstly perform clustering and obtain cluster assignments and centroids. Then, we propose an original strategy to provide aligned targets for self-supervised learning.

Unsupervised Learning by Clustering As most of the training data are unlabeled, it is important to effectively use a mass of unlabeled samples for discovering novel classes.

Inspired by DeepCluster (Caron et al. 2018), we can benefit from the discriminative power of BERT to produce structured outputs as weak supervised signals. Specifically, we firstly extract features of all training data from the pre-trained model. Then, we use a standard clustering algorithm, k-means, to learn both the optimal cluster centroid matrix $\boldsymbol{C}$ and assignments $\{y_i\}_{i=1}^{N}$:

$$\min_{\boldsymbol{C} \in \mathbb{R}^{K \times D}} \frac{1}{N} \sum_{i=1}^{N} \min_{y_i \in \{1,\ldots,K\}} \| \boldsymbol{I}_i - \boldsymbol{C}_{y_i} \|_2^2, \tag{4}$$

where N is the number of training samples and $\|\cdot\|_2^2$ denotes the squared Euclidean distance. We then utilize the cluster assignments as pseudo-labels for feature learning.

Self-supervised Learning with Aligned Pseudo-labels DeepCluster alternates between clustering and updating network parameters. It performs k-means to produce cluster assignments as pseudo-labels and uses them to train the neural network. However, the indices after k-means are permuted randomly in each training epoch, so the classifier parameters have to be reinitialized before each epoch (Zhan et al. 2020). Thus, we propose an alignment strategy to tackle the assignment inconsistency problem.

We notice that DeepCluster lacks the use of the centroid matrix $\boldsymbol{C}$ in Eq. 4. However, $\boldsymbol{C}$ is a crucial part, which contains the optimal averaged assignment target of clustering. As each embedded sample is assigned to its nearest centroid in Euclidean space, we naturally adopt $\boldsymbol{C}$ as the prior knowledge to adjust the inconsistent cluster assignments in different training epochs. That is, we convert this problem into the centroid alignment. Though the intent representations are updated continually, similar intents are distributed in near locations. The centroid synthesizes all similar intent samples in its cluster, so it is more stable and suitable for guiding the alignment process. We suppose the centroids in contiguous training epochs are relatively consistently distributed in Euclidean space and adopt the Hungarian algorithm (Kuhn 1955) to obtain the optimal mapping G:

$$\boldsymbol{C}^c = G(\boldsymbol{C}^l), \tag{5}$$

where $\boldsymbol{C}^c$ and $\boldsymbol{C}^l$ respectively denote the centroid matrix in the current and last training epoch. Then, we obtain the aligned pseudo-labels y^{align} with $G(\cdot)$:

$$y^{align} = G^{-1}(y^c), \tag{6}$$

where G^{-1} denotes the inverse mapping of G and y^c denotes the pseudo-labels in the current training epoch. Finally, we use the aligned pseudo-labels to perform self-supervised learning under the supervision of the softmax loss $\mathcal{L}_s$:

$$\mathcal{L}_s = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\phi(\boldsymbol{I}_i)^{y_i^{align}})}{\sum_{j=1}^{K} \exp(\phi(\boldsymbol{I}_i)^j)}, \tag{7}$$

where $\phi(\cdot)$ is the pseudo-classifier for self-supervised learning, and $\phi(\cdot)^j$ denotes the output logits of the j^{th} class.

We use the cluster validity index (CVI) to evaluate the quality of clusters obtained during each training epoch after k-means. Specifically, we adopt an unsupervised metric Silhouette Coefficient (Rousseeuw 1987) for evaluation:

$$SC = \frac{1}{N} \sum_{i=1}^{N} \frac{b(\boldsymbol{I}_i) - a(\boldsymbol{I}_i)}{\max\{a(\boldsymbol{I}_i), b(\boldsymbol{I}_i)\}}, \tag{8}$$

	Method	CLINC			BANKING		
		NMI	ARI	ACC	NMI	ARI	ACC
Unsupervised.	KM	70.89	26.86	45.06	54.57	12.18	29.55
	AG	73.07	27.70	44.03	57.07	13.31	31.58
	SAE-KM	73.13	29.95	46.75	63.79	22.85	38.92
	DEC	74.83	27.46	46.89	67.78	27.21	41.29
	DCN	75.66	31.15	49.29	67.54	26.81	41.99
	DAC	78.40	40.49	55.94	47.35	14.24	27.41
	DeepCluster	65.58	19.11	35.70	41.77	8.95	20.69
Semi-supervised.	PCK-means	68.70	35.40	54.61	48.22	16.24	32.66
	BERT-KCL	86.82	58.79	68.86	75.21	46.72	60.15
	BERT-MCL	87.72	59.92	69.66	75.68	47.43	61.14
	CDAC+	86.65	54.33	69.89	72.25	40.97	53.83
	BERT-DTC	90.54	65.02	74.15	76.55	44.70	56.51
	DeepAligned	**93.89**	**79.75**	**86.49**	**79.56**	**53.64**	**64.90**

Table 2: The clustering results on two datasets. We evaluate both unsupervised and semi-supervised clustering methods.

	Method	CLINC			BANKING		
		NMI	ARI	ACC	NMI	ARI	ACC
Without Pre-training	Reinitialization	57.80	9.63	23.02	34.34	4.49	13.67
	Alignment	62.53	14.10	28.63	36.91	5.23	15.42
With Pre-training	Reinitialization	82.90	45.67	55.80	68.12	31.56	41.32
	Alignment	**93.89**	**79.75**	**86.49**	**79.56**	**53.64**	**64.90**

Table 3: Effectiveness of the pre-training and the alignment strategy on two datasets.

where $a(\boldsymbol{I}_i)$ is the average distance between $\boldsymbol{I}_i$ and all other samples in the i^{th} cluster, which indicates the intra-class compactness. $b(\boldsymbol{I}_i)$ is the smallest distance between $\boldsymbol{I}_i$ and all samples not in the i^{th} cluster, which indicates the inter-class separation. The range of SC is between -1 and 1, and the higher score means the better clustering results.

4 Experiments

4.1 Datasets

We conduct experiments on two challenging benchmark intent datasets. Detailed statistics are shown in Table 1.

CLINC It is an intent classification dataset (Larson et al. 2019), which contains 22,500 queries covering 150 intents across 10 domains.

BANKING It is a fine-grained dataset in the banking domain (Casanueva et al. 2020), which contains 13,083 customer service queries with 77 intents.

4.2 Baselines

Unsupervised We firstly compare with unsupervised clustering methods, including K-means (KM) (MacQueen et al. 1967), agglomerative clustering (AG) (Gowda and Krishna 1978), SAE-KM, DEC (Xie, Girshick, and Farhadi 2016), DCN (Yang et al. 2017), DAC (Chang et al. 2017), and DeepCluster (Caron et al. 2018).

For KM and AG, we represent the sentences with the averaged pre-trained 300-dimensional word embeddings from GloVe (Pennington, Socher, and Manning 2014). For SAE-KM, DEC, and DCN, we encode the sentences with the stacked autoencoder (SAE), which is helpful to capture meaningful semantics on real-world datasets (Xie, Girshick, and Farhadi 2016). As DAC and DeepCluster are unsupervised clustering methods in computer vision, we replace the backbone with the BERT model for extracting text features.

Semi-supervised We also compare our method with semi-supervised clustering methods, including PCK-

	Methods	CLINC (K'=300)		BANKING (K'=154)	
		K (Pred)	Error	K (Pred)	Error
25%	BERT-MCL	38	75.00	19	75.32
	BERT-DTC	94	37.33	37	51.95
	DeepAligned	**122**	**18.67**	**66**	**14.29**
50%	BERT-MCL	75	50.00	38	50.65
	BERT-DTC	**131**	**12.67**	**71**	**7.79**
	DeepAligned	130	13.33	64	16.88
75%	BERT-MCL	112	25.33	58	24.68
	BERT-DTC	195	30.00	110	42.86
	DeepAligned	**129**	**14.00**	**67**	**12.99**

Table 4: The results of predicting K with an unknown number of clusters. We vary ratio of known classes in the range of 25%, 50% and 75%, and set K' as two times of the ground truth number of clusters during clustering.

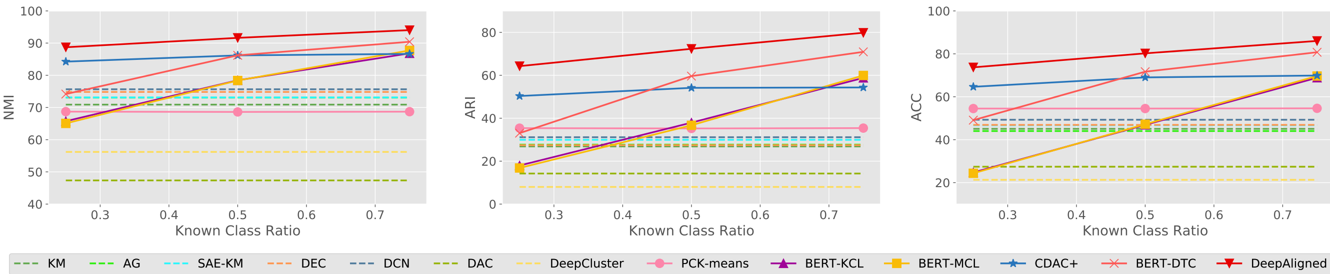

Figure 3: Influence of the known class ratio on CLINC dataset.

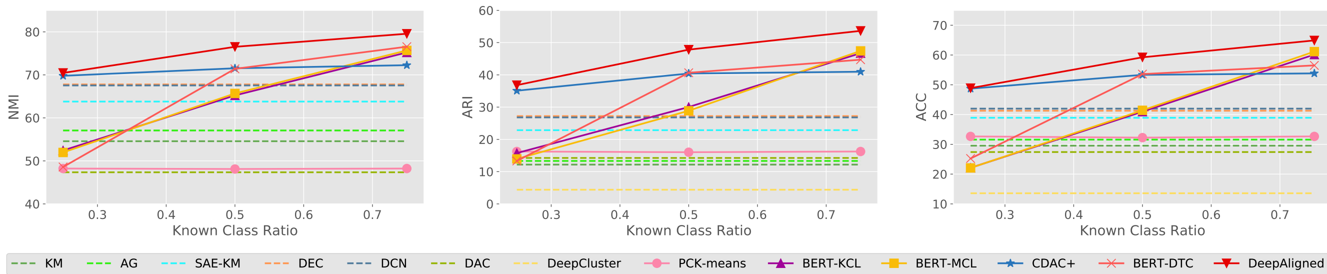

Figure 4: Influence of the known class ratio on BANKING dataset.

means (Basu, Banerjee, and Mooney 2004), BERT-KCL (Hsu, Lv, and Kira 2018), BERT-MCL (Hsu et al. 2019), BERT-DTC (Han, Vedaldi, and Zisserman 2019) and CDAC+ (Lin, Xu, and Zhang 2020). For a fairness comparison, we replace the backbone of these methods with the same BERT model as ours.

4.3 Evaluation Metrics

We adopt three widely used metrics to evaluate the clustering results: Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Accuracy (ACC). To calculate ACC, we use the Hungarian algorithm to obtain the mapping between the predicted classes and ground-truth classes.

4.4 Evaluation Settings

Following the same settings as in (Lin, Xu, and Zhang 2020), we randomly select 10% of training data as labeled and choose 75% of all intents as known. We split datasets into the training, validation, and test sets. The number of intent categories is set as ground-truth. We first use the little labeled data (containing only known intents) for pre-training and tune with the validation set. Then, we use all training data for self-supervised learning and evaluate the cluster performance with Silhouette Coefficient (as mentioned in Eq.8). Finally, we evaluate the performance on the test set and report the averaged results over ten runs of experiments with different random seeds.

4.5 Implementation Details

We use the pre-trained BERT model (bert-uncased, with 12-layer transformer) implemented in PyTorch (Wolf et al. 2019) as our network backbone, and adopt most of its suggested hyper-parameters for optimization. The training batch size is 128, the learning rate is $5e^{-5}$, and the dimension of intent features D is 768. Moreover, we freeze all but the last transformer layer parameters to achieve better performance with BERT backbone, and speed up the training procedure as suggested in (Lin, Xu, and Zhang 2020).

5 Results and Discussion

Table 2 shows the results of all compared methods. We highlight the best results in bold. Compared with baselines, our method consistently achieves the best results and outperforms other baselines by a large margin on all metrics and datasets. It demonstrates the effectiveness of our method to discover new intents with limited known intent data. We also find most semi-supervised methods perform better than unsupervised methods. It indicates that even with limited labeled data as prior knowledge, it is also helpful to improve the performance of unsupervised clustering.

5.1 Effect of the Alignment Strategy

To investigate the contribution of the alignment strategy, we compare our method with the reinitialization strategy (Caron et al. 2018). As shown in Table 3, our method has significant improvements over the reinitialization strategy on both semi-supervised and unsupervised settings. We suppose the reason is that random initialization drops out the well-trained parameters in the classifier in the former epochs. However, our method saves history embedding information by finding the mapping of pseudo-labels between contiguous epochs, which provides stronger supervised signals for representation learning.

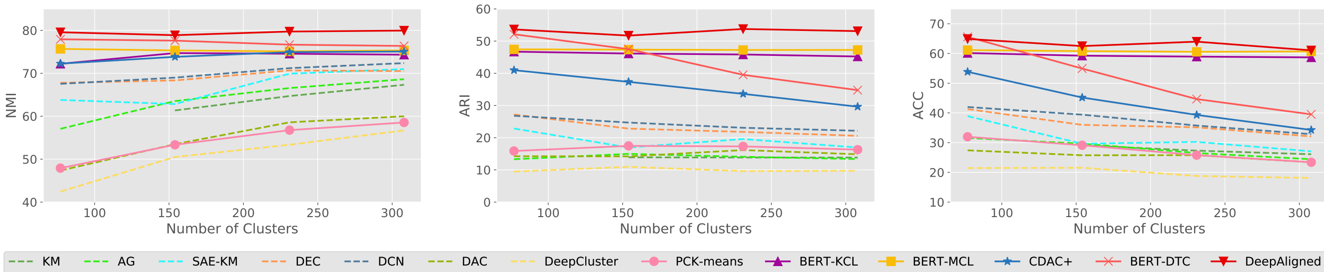

Figure 5: Influence of the number of clusters on BANKING dataset.

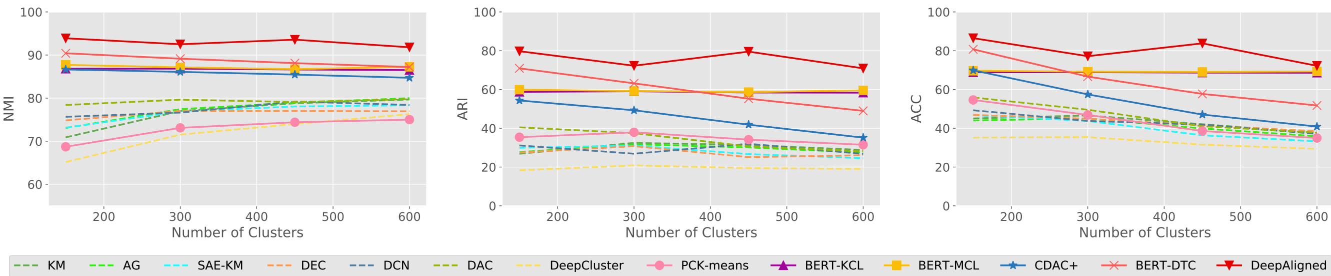

Figure 6: Influence of the number of clusters on CLINC dataset.

5.2 Estimate K

To investigate the effectiveness to predict K, we assign K' as two times of the ground truth number of classes and compare with another two state-of-the-art methods (BERT-MCL and BERT-DTC). We vary the ratio of known classes in the range of 25%, 50%, and 75%, and calculate the error rate (the lower is better) for evaluation. As shown in Table 4, our method achieves the lowest or competitive results with different known class ratios. It shows the reasonability to estimate the cluster number by removing low-confidence clusters. We notice that BERT-DTC is a strong baseline, especially with 50% known classes. It relies on labeled samples to generate the probe set for determining the optimal number of classes. But the performance is unstable as the number of known intents changes. We also find the predicted K of BERT-MCL is close to the number of known classes, which is because it uses known labeled data for classification while clustering. However, the classification information dominates in training and tends to predict known classes during testing.

5.3 Effect of Known Class Ratio

To investigate the influence of the number of known intents, we vary the known class ratio in the range of 25%, 50% and 75%. As shown in Figure 3 and Figure 4, our method still achieves the best results in all settings. All semi-supervised methods are vulnerable to the number of known intents. Particularly, though BERT-MCL and BERT-DTC are competitive with a 75% known class ratio, they drop dramatically as the known class ratio decreases. We suppose the reason is that they largely depend on the prior knowledge of known intents to construct high-quality supervised signals (e.g., the pairwise similarity in BERT-MCL, initialized centroids in BERT-DTC) for clustering. In contrast, our method only needs known intent data for learning intent representations and is free from prior during the self-learning process. So our method performs more robust with fewer known classes.

5.4 Effect of the Number of Classes

To investigate the sensitiveness to the assigned cluster number K', we vary K' from the ground-truth number to four times of it with 75% known classes. As shown in Figure 5 and Figure 6, our method achieves robust results even with large assigned cluster numbers. We suppose it benefits from a relatively accurate estimated cluster number. We also notice that BERT-MCL and BERT-KCL seem to be insensitive to the number of classes. However, their performances largely depend on the prior knowledge of known intents.

6 Conclusion and Future Work

In this work, we have introduced an effective method for discovering new intents. Our method successfully transfers the prior knowledge of limited known intents and estimates the number of intents eliminating low-confidence clusters. Moreover, it provides more stable and concrete supervised signals to guide the clustering process. We conduct extensive experiments on two challenging benchmark datasets to evaluate the performance. Our method achieves significant improvements over the compared methods and obtains more accurate estimated cluster numbers with limited prior knowledge. In the future, we will try different clustering methods besides k-means to produce supervised signals and explore more self-supervised learning methods.

7 Acknowledgments

This paper is founded by National Natural Science Foundation of China (Grant No: 61673235) and National Key RD Program Projects of China (Grant No: 2018YFC1707605). This work is also supported by seed fund of Tsinghua University (Department of Computer Science and Technology)-Siemens Ltd., China Joint Research Center for Industrial Intelligence and Internet of Things.

References

Basu, S.; Banerjee, A.; and Mooney, R. J. 2004. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, 333–344. SIAM.

Bilenko, M.; Basu, S.; and Mooney, R. J. 2004. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 11. ACM.

Brychcin, T.; and Král, P. 2017. Unsupervised Dialogue Act Induction using Gaussian Mixtures. In *EACL*.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149.

Casanueva, I.; Temcinas, T.; Gerz, D.; Henderson, M.; and Vulic, I. 2020. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 5879–5887.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, 226–231.

Goo, C.-W.; Gao, G.; Hsu, Y.-K.; Huo, C.-L.; Chen, T.-C.; Hsu, K.-W.; and Chen, Y.-N. 2018. Slot-Gated Modeling for Joint Slot Filling and Intent Prediction. In *Proceedings of NAACL*, 753–757.

Gowda, K. C. 1984. A feature reduction and unsupervised classification algorithm for multispectral data. *Pattern Recognition* 667–676.

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

Hakkani-Tür, D.; Ju, Y.-C.; Zweig, G.; and Tur, G. 2015. Clustering novel intents in a conversational interaction system with semantic parsing. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Hakkani-Tür, D.; Celikyilmaz, A.; Heck, L.; and Tur, G. 2013. A Weakly-Supervised Approach for Discovering New User Intents from Search Query Logs. In *INTERSPEECH*.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to Discover Novel Visual Categories via Deep Transfer Clustering. In *International Conference on Computer Vision (ICCV)*.

Haponchyk, I.; Uva, A.; Yu, S.; Uryupina, O.; and Moschitti, A. 2018. Supervised Clustering of Questions into Intents for Dialog System Applications. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2310–2321.

Hsu, Y.-C.; Lv, Z.; and Kira, Z. 2018. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations*.

Hsu, Y.-C.; Lv, Z.; Schlosser, J.; Odom, P.; and Kira, Z. 2019. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics* 2(1-2): 83–97.

Larson, S.; Mahendran, A.; Peper, J. J.; Clarke, C.; Lee, A.; Hill, P.; Kummerfeld, J. K.; Leach, K.; Laurenzano, M. A.; Tang, L.; and Mars, J. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of EMNLP-IJCNLP*, 1311–1316.

Lin, T. E.; and Xu, H. 2019. Deep Unknown Intent Detection with Margin Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Lin, T.-E.; Xu, H.; and Zhang, H. 2020. Discovering New Intents via Constrained Deep Adaptive Clustering with Cluster Refinement. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

MacQueen, J.; 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, 281–297. Oakland, CA, USA.

Min, Q.; Qin, L.; Teng, Z.; Liu, X.; and Zhang, Y. 2020. Dialogue State Induction Using Neural Latent Variable Models. In *Proceedings of IJCAI*, 3845–3852.

Padmasundari; and Bangalore, S. 2018. Intent Discovery Through Unsupervised Semantic Text Clustering. In *Proceedings of Interspeech 2018*, 606–610.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP 2014*, volume 14, 1532–1543.

Perkins, H.; and Yang, Y. 2019. Dialog Intent Induction with Deep Multi-View Clustering. In *Proceedings of EMNLP-IJCNLP*, 4016–4025.

Platt, J.; et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10(3): 61–74.

Qin, L.; Che, W.; Li, Y.; Ni, M.; and Liu, T. 2020. DCR-Net: A Deep Co-Interactive Relation Network for Joint Dialog Act Recognition and Sentiment Classification. *Proceedings of AAAI* 34: 8665–8672.

Qin, L.; Che, W.; Li, Y.; Wen, H.; and Liu, T. 2019. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding. In *Proceedings of EMNLP*.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53 – 65.

Shi, C.; Chen, Q.; Sha, L.; Li, S.; Sun, X.; Wang, H.; and Zhang, L. 2018. Auto-Dialabel: Labeling Dialogue Data with Unsupervised Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 684–689.

Vedula, N.; Lipka, N.; Maneriker, P.; and Parthasarathy, S. 2020. Open Intent Extraction from Natural Language Interactions. In *Proceedings of The WWW*, 2009–2020.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schrödl, S. 2001. Constrained K-means Clustering with Background Knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 577–584.

Wang, Y.; Shen, Y.; and Jin, H. 2018. A Bi-Model Based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling. In *Proceedings of NAACL*, 309–314.

Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2(1-3): 37–52.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3861–3870.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5147–5156.

Zhan, X.; Xie, J.; Liu, Z.; Ong, Y.-S.; and Loy, C. C. 2020. Online Deep Clustering for Unsupervised Representation Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.