Disentangled Knowledge Transfer for OOD Intent Discovery with Unified Contrastive Learning

Yutao Mou[1*], Keqing He[2*], Yanan Wu[1*], Zhiyuan Zeng[1]
Hong Xu[1], Huixing Jiang[2], Wei Wu[2], Weiran Xu[1*]
[1]Pattern Recognition & Intelligent System Laboratory
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Meituan Group, Beijing, China
{myt,yanan.wu,zengzhiyuan,xuhong,xuweiran}@bupt.edu.cn
{hekeqing,jianghuixing,wuwei30}@meituan.com

Abstract

Discovering Out-of-Domain(OOD) intents is essential for developing new skills in a task-oriented dialogue system. The key challenge is how to transfer prior IND knowledge to OOD clustering. Different from existing work based on shared intent representation, we propose a novel disentangled knowledge transfer method via a unified multi-head contrastive learning framework. We aim to bridge the gap between IND pre-training and OOD clustering. Experiments and analysis on two benchmark datasets show the effectiveness of our method. [1]

1 Introduction

Out-of-domain (OOD) intent discovery aims to group new unknown intents into different clusters, which helps improve the dialogue system for future development. Compared to existing text clustering tasks, OOD discovery considers how to leverage the prior knowledge of known in-domain (IND) intents to enhance discovering unknown OOD intents, which makes it challenging to directly apply existing clustering algorithms (MacQueen, 1967; Xie et al., 2016; Chang et al., 2017; Caron et al., 2018) to the OOD discovery task.

Previous unsupervised OOD discovery models (Hakkani-Tür et al., 2015; Padmasundari and Bangalore, 2018; Shi et al., 2018) only model OOD data but ignore prior knowledge of in-domain data thus suffer from poor performance. Therefore, recent work (Lin et al., 2020; Zhang et al., 2021) focus more on the semi-supervised setting where they firstly pre-train an in-domain intent classifier then perform clustering algorithms on extracted OOD intent representations by the pre-trained IND intent classifier. For example, Lin et al. (2020) firstly pre-trains a BERT-based (Devlin et al., 2019) IND intent classifier then uses intent representations to perform a pairwise clustering algorithm (Chang et al., 2017). Further, Zhang et al. (2021) proposes an iterative clustering method, DeepAligned, to obtain pseudo supervised signals using K-means (MacQueen, 1967). However, all of these methods ignore the matching between IND pre-training stage and OOD clustering stage because they formulate IND pre-training as the classification task while OOD clustering as the text clustering task. The different learning objectives make it hard to transfer prior IND knowledge to OOD. Besides, previous work only transfer a single intent representation from the pre-trained IND classifier to OOD clustering. Considering the entanglement of the intent representation, simply transferring IND features may harm OOD clustering. For example, there exist two levels of intent features, instance-level and class-level knowledge in the pre-trained IND classifier. Decoupling different levels of intent features helps better knowledge transferability.

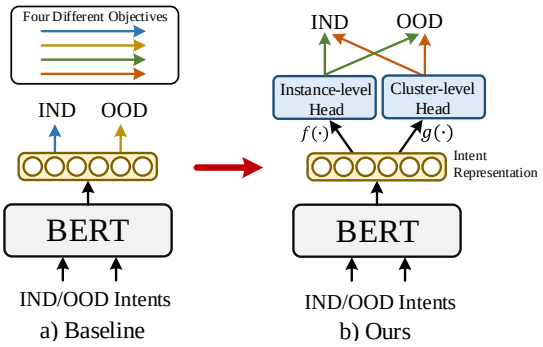


Figure 1: Comparison between baselines and our proposed DKT model.

To solve the issues, we propose a novel **D**isentangled **K**nowledge **T**ransfer method (**DKT**) via a unified multi-head contrastive learning framework to transfer disentangled IND intent representations to OOD clustering. The main intuition is how to perform better knowledge transfer. As shown in Fig 1, we decouple the pre-trained intent representations into two independent subspaces, instance-level and class(cluster)-level using a uni-

*The first three authors contribute equally. Weiran Xu is the corresponding author.

[1]We release our code at https://github.com/myt517/DKT.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 46 - 53

fied contrastive learning framework. Different from existing OOD discovery work, we equip the traditional IND pre-training stage with a similar contrastive objective as the clustering stage. Specifically, we firstly learn intent features using a context encoder like BERT, then add two independent transformation heads (instance-level head f and class-level head g) on top of BERT. In the IND pre-training stage, we use the head f to perform supervised instance-level contrastive learning (Chen et al., 2020; Khosla et al., 2020; Gunel et al., 2021; Zeng et al., 2021) and the head g to compute traditional classification loss like cross-entropy. In the OOD clustering stage, we employ similar objectives for these two heads where f is still used for instance-level contrastive learning and g is used to perform class(cluster)-level contrastive learning (Li et al., 2021). We leave the details in the following Section 2. Using the unified contrastive objectives for pre-training and clustering bridges the gap between the two stages. Besides, the two independent heads decouple the instance- and cluster-level contrastive learning to learn disentangled intent representations for better knowledge transfer. Section 4 demonstrates the effectiveness of multi-head disentanglement.

Our contributions are three-fold: (1) We propose a novel disentangled knowledge transfer method for OOD discovery to better leverage prior IND knowledge. (2) We propose a unified multi-head contrastive learning framework to bridge the gap between IND pre-training and OOD clustering. (3) Experiments and analysis on two benchmark datasets demonstrate the effectiveness of our method for OOD discovery.

2 Approach

Problem Formulation Given a set of labeled in-domain data $(\mathcal{X}_{IND}, \mathcal{Y}_{IND})$ and unlabeled OOD data $(\mathcal{X}_{OOD}, \mathcal{Y}_{OOD})$, OOD discovery aims to cluster OOD groups from unlabeled OOD data using prior knowledge from labeled IND data. Note that IND data has no overlapping with OOD data. Generally, OOD discovery includes two stages, IND pre-training which aims to obtain a decent intent representation via labeled IND data, and OOD clustering which aims to group OOD intents into different clusters.

Overall Architecture Fig 2 shows the overall architecture of our proposed DKT model. We firstly use the same BERT (Devlin et al., 2019) backbone to extract intent representations as the previous work DeepAligned (Zhang et al., 2021). Then we decouple the intent representations into two independent subspaces and use a unified contrastive learning framework to perform both IND pre-training and OOD clustering.

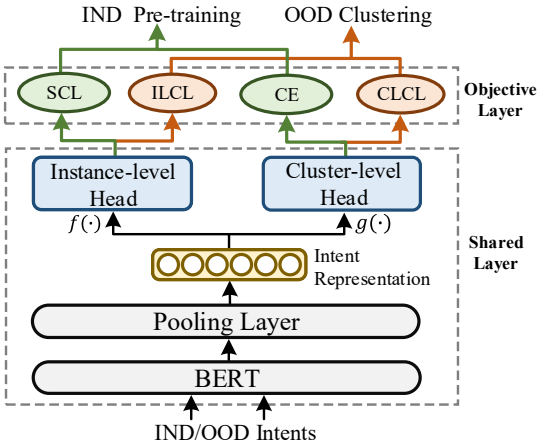


Figure 2: The overall architecture of our DKT.

IND Pre-training Different from existing methods that regard IND pre-training as a single intent classification task, we formulate it as an instance-wise discriminative task and a class-wise classification task via contrastive learning. Given an IND intent example x_i, we firstly obtain its intent representation z_i using a BERT encoder and a pooling layer.[2] Then we use two independent transformation heads f and g to get two disentangled latent vectors $f_i = f(z_i)$ and $g_i = g(z_i)$.[3] On top of the instance-level head f, we perform supervised contrastive learning (SCL) (Khosla et al., 2020; Zeng et al., 2021) as follows:

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \log \frac{\exp\left(f_i \cdot f_j / \tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp\left(f_i \cdot f_k / \tau\right)}$$

where N_{y_i} is the total number of examples in the batch that have the same label as y_i and $\mathbf{1}$ is an indicator function. Following Gao et al. (2021); Yan et al. (2021), we employ simple dropout (Srivastava et al., 2014) as data augmentation. SCL can model instance-wise semantic similarities by pulling together IND intents belonging to the same class while pushing apart samples from different

[2]For a fair comparison, we use the same BERT-based backbone as previous work. We leave the details to Section 3.4.

[3]In the experiments, we use two separate two-layer non-linear MLPs for head f and g. For simplicity, we set both the input dimension and output dim to 768, same as the hidden state dim of BERT-base.

Models		CLINC-10%			CLINC-20%			CLINC-30%			Banking		
		ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
Unsup.	K-means	58.67	43.81	67.77	48.89	30.90	64.68	42.22	23.65	60.55	32.81	8.30	17.30
	DeepCluster	53.15	37.80	62.31	47.73	34.55	65.91	33.96	18.89	56.21	29.81	7.79	17.34
	DeepAligned	62.66	47.60	71.50	48.24	34.49	66.24	39.02	24.50	61.16	36.56	12.57	21.84
	DKT(ours)	**74.22**	**61.37**	**76.67**	**57.56**	**44.94**	**72.40**	**50.07**	**35.53**	**69.81**	**40.00**	**18.20**	**30.10**
Semi-sup.	PTK-means	70.22	50.39	73.92	57.56	37.02	72.71	61.63	40.96	75.90	55.00	36.18	53.75
	DeepCluster	78.13	68.31	82.87	83.42	76.18	89.33	78.09	71.05	88.70	60.59	41.88	55.22
	CDAC+	88.00	75.18	88.33	84.89	75.98	89.96	73.04	64.44	87.90	77.50	60.53	71.14
	DeepAligned	95.11	89.81	94.13	93.80	90.22	95.83	91.56	86.58	94.91	77.78	66.95	76.91
	DKT(ours)	**97.78**	**95.16**	**96.97**	**96.89**	**93.69**	**96.94**	**94.96**	**90.25**	**95.94**	**84.69**	**71.11**	76.92

Table 1: Performance comparison on two datasets. We randomly sample 10%, 20% and 30% of all classes as OOD types for CLINC, 10% for Banking. We evaluate both unsupervised and semi-supervised methods. Unsup DKT denotes DKT w/o IND pre-training. Results are averaged over three random runs. ($p < 0.05$ under t-test)

classes. Therefore, SCL helps maximize inter-class variance and minimize intra-class variance, further improves OOD clustering. On top of the class-level head g, we use a cross-entropy classification loss to learn class(cluster)-wise distinction. Section 4 confirms both the objectives improve the performance and SCL has a larger effect.

OOD Clustering The key challenge of OOD clustering is how to learn intent representations and cluster assignments. Previous state-of-the-art model DeepAligned (Zhang et al., 2021) iteratively repeats the two stages which results in poor clustering efficiency and accuracy. Thus, we propose an end-to-end contrastive clustering method (Li et al., 2021) to jointly learn representations and cluster assignments. Specifically, given an OOD example x_i, we firstly use the pre-trained BERT encoder and transformation heads to get OOD intent latent vectors f_i and g_i. Then, on top of the instance-level head f, we perform instance-level contrastive learning(ILCL) (Chen et al., 2020) as follows:

$$\ell_{i,j}^{ins} = -\log \frac{\exp\left(\text{sim}\left(f_i, f_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp\left(\text{sim}\left(f_i, f_k\right)/\tau\right)}$$

where f_j denotes the dropout-augmented OOD sample and τ denotes temperature [4]. On top of the cluster-level head g, we perform contrastive clustering following Li et al. (2021) . Specifically, given an OOD cluster-level latent vector g_i, we firstly project it to a vector with dimension K which equals to the pre-defined cluster number.[5] Suppose we input a batch of OOD samples so we can get a feature matrix of $N \times K$. Then we regard i-th column of the matrix as the i-th cluster representation y_i and construct cluster-level CL(CLCL) as follows:

$$\ell_{i,j}^{clu} = -\log \frac{\exp\left(\text{sim}\left(y_i, y_j\right)/\tau\right)}{\sum_{k=1}^{2K} 1_{[k\neq i]} \exp\left(\text{sim}\left(y_i, y_k\right)/\tau\right)}$$

where y_j is the dropout-augmented cluster representation of y_i and sim denotes cosine distance. Following Li et al. (2021), we also add a regularization item to avoid the trivial solution that most instances are assigned to the single cluster. For training, we simply add the above objectives in the experiments. For inference, we only use the cluster-level contrastive head and compute the argmax to get the cluster results without additional K-means. Generally, the instance-CL focuses on distinguishing different intent samples while the cluster-CL identifies distinct OOD categories. Combining the two stages, our proposed unified contrastive learning framework can effectively bridge the gap between IND pre-training and OOD clustering.

3 Experiment

3.1 Datasets

We show the detailed statistics of CLINC(Larson et al., 2019) and BANKING(Casanueva et al., 2020) datasets in Table 2. CLINC contains 22,500 queries covering 150 intents and Banking contains 13,083 customer service queries with 77 intents. To construct IND/OOD data, we ramdomly divided the two datasets in three ramdom runs, according to the specified OOD ratio(10%, 20%, 30% for CLINC, 10% for Banking), and the rest is IND data. Note that we only use the IND data for pre-training and use OOD data for clustering. To avoid the randomness of splitting IND/OOD, we average results over three random runs. For each run, all the models use the same divided dataset. Different from previous work Zhang et al. (2021), we assume that the unlabeled data only contains OOD data instead of a mixture of IND and OOD, aiming to fairly evaluate the OOD clustering performance.

[4] we set it to 0.5 in the experiments.

[5] In this paper, we focus on the fixed cluster number K setting and leave estimating K to future work.

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

Table 2: Statistics of CLINC and BANKING datasets.

In real scenarios, we can use OOD detection models (Xu et al., 2020; Zeng et al., 2021) to collect high-quality OOD data for OOD intent discovery.

3.2 Baselines

We mainly compare our method with semi-supervised baselines: PTK-means (k-means with IND pre-training), DeepCluster (Caron et al., 2018) and two state-of-the-art OOD discovery methods CDAC+ (Lin et al., 2020) and DeepAligned (Zhang et al., 2021). We also report the unsupervised results (without IND pretraining) of these methods for a comprehensive comparison. For fairness, we use the same BERT backbone as the baselines. We leave the detailed baselines in the appendix A.1.

3.3 Evaluation Metrics

We adopt three widely used metrics to evaluate the clustering results: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). To calculate ACC, we use the Hungarian algorithm (Kuhn, 1955) to obtain the mapping between the predicted classes and ground-truth classes.

3.4 Implementation Details

For a fair comparison with previous work, we use the pre-trained BERT model (bert-base-uncased [6], with 12-layer transformer) as our network backbone, and add a pooling layer to get intent representation(dimension=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 (Zhang et al., 2021). During the pre-training phase, the training batch size is 128, and during the clustering phase, the training batch size is 512 for CLINC-10%, CLINC-30%, Banking-10%, and 400 for CLINC-20%. The learning rate is 5e-5 in the pre-training phase and 0.0003 in the clustering phase. Notably, We use dropout (Gao et al., 2021) to construct augmented examples for contrastive learning with dropout rate 0.1. For the instance-level contrastive head, the dimensionality of the row space is set to 128, and the temperatures of SCL and instance-level CL are 0.5. As for the cluster-level contrastive head, the dimensionality of the column space is naturally set to the number of IND classes/OOD clusters, and the cluster-level temperature parameter τ= 1.0 is used for all datasets. We use SC of validation OOD data (still unlabeled data) to choose the best checkpoint. The pre-training stage of our model lasts about 30 minutes and clustering runs for 10 minutes on CLINC-10%, both using a single Tesla T4 GPU(16 GB of memory).

[6]https://github.com/google-research/bert

3.5 Main Results

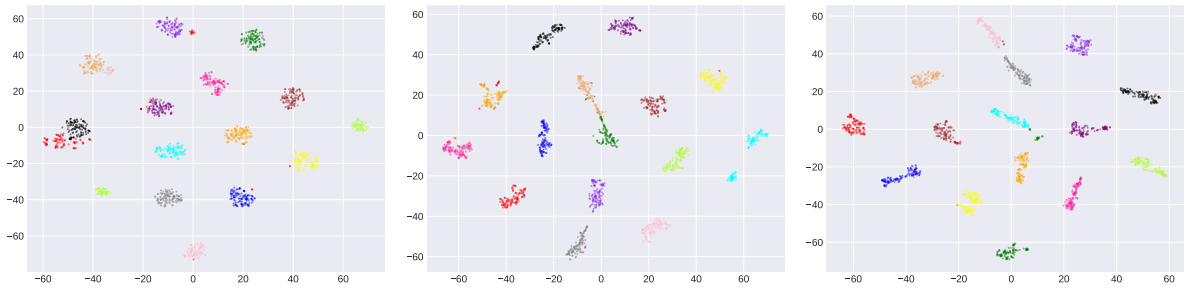
Table 1 shows the performance comparison of different models on two datasets. Under both unsupervised and semi-supervised settings, our proposed DKT consistently outperforms all the baselines. In this paper, we mainly focus on the latter setting. For the Semi-sup setting on CLINC-10%, DKT outperforms the previous state-of-the-art DeepAligned by 2.67%(ACC), 5.35%(ARI), 2.84%(NMI). Similar improvements are observed on other datasets. The results prove the effectiveness of our proposed disentangled knowledge transfer for OOD discovery. Comparing Unsup DKT with Semi-sup DKT, the latter significantly outperforms the former by 23.56%(ACC), 33.79%(ARI), 20.30%(NMI), which demonstrates the effectiveness of IND pre-training(see details in appendix A.2).

4 Qualitative Analysis

Effect of Disentangled Intent Representations Tab 3 shows performance comparison of DKT and KT under two settings. We find Disentangled KT significantly outperforms KT both on two settings, which proves the effectiveness of representation disentanglement for knowledge transfer.

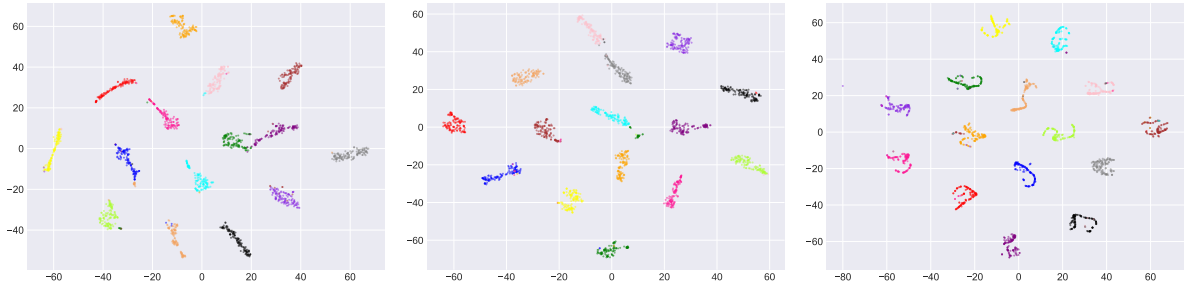
Visualization To confirm the effectiveness of DKT, we perform OOD intent representation visualization of DeepAligned, KT and DKT in Fig 3. Note that we use the same representation following the pooling layer for fair comparison. We find both DeepAligned and KT have some mixed OOD clusters while DKT forms clearly separate decision boundaries between clusters, which shows our proposed DKT obtains discriminative OOD representations for OOD discovery. Besides, Section 4 further explore the effect of different layer and representations after MLP g gets the best performance.

Error Analysis We further analyze the error cases of DeepAligned and DKT in Fig 5. We find that for



(a) DeepAligned (b) KT (c) DKT

Figure 3: Visualization of different methods. KT denotes only using single MLP head.



(a) Instance-level head (b) BERT + pooling layer (c) Cluster-level head

Figure 4: Intent representations at different layers

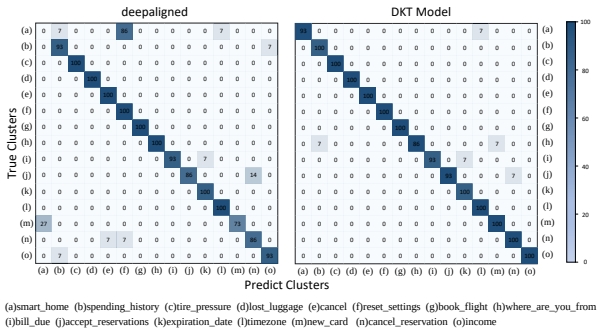


Figure 5: Confusion matrix for the clustering results of DeepAligned and DKT on CLINC-10%. The percentage values along the diagonal represent how many samples are correctly clustered into the corresponding class. The larger the number, the deeper the color.

Models		ACC	ARI	NMI
Unsup.	KT	68.89	56.33	73.93
	DKT	74.22	61.37	76.67
Semi-sup.	KT	95.11	90.23	94.53
	DKT	**97.78**	**95.16**	**96.97**

Table 3: Effect of disentangled intent representations.

Models	ACC	ARI	NMI
DKT	**97.78**	**95.16**	**96.97**
-w/o SCL	92.26	86.33	92.62
-w/o CE	95.16	90.61	94.80
-w/o ILCL	90.93	85.43	92.07
-w/o CLCL	90.36	82.91	90.55

Table 4: Effect of different learning objectives.

similar OOD intents, DeepAligned is probably confused but our DKT can effectively distinguish them. For example, DeepAligned incorrectly groups *accept_reservation* intents into *cancel_reservation* (14% error rate) vs DKT(7%), which proves DKT helps separate semantically similar OOD intents.

Ablation Study To understand the effect of different objectives of DKT, we perform abalation study in Tab 4 by removing each loss. Results show all the losses contribute to the performance especially SCL, ILCL and CLCL, which confirms the effectiveness of our unified contrastive framework.

Intent Representations at Different Layers In order to further explore the effectiveness of disentangled representation, we visualize the output vectors of instance-level head and cluster-level head and compare them with the output vector after BERT + pooling in Fig 4. We can find that the output obtained by instance-level head forms a narrow and long cluster distribution, while the output obtained by cluster-level head forms a more compact and uniform cluster distribution. We argue that this reflects the effect of decoupling, that is, instance-level head decouples the uniqueness of each sample, and cluster-level head decouples the category characteristics of each sample.

5 Conclusion

In this paper, we propose a novel disentangled knowledge transfer method (DKT) via a unified multi-head contrastive learning framework to transfer disentangled IND intent representations to OOD clustering. Experiments and analysis on two benchmarks demonstrate the effectiveness of DKT for OOD discovery. We hope to explore more self-supervised representation learning methods for OOD discovery in the future.

Acknowledgements

We thank all anonymous reviewers for their helpful comments and suggestions. This work was partially supported by National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, DOCOMO Beijing Communications Laboratories Co., Ltd, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701.

Broader Impact

Task-oriented dialogue systems have demonstrated remarkable performance across a wide range of applications, with the promise of a significant positive impact on human production mode and lifeway. Intent classification is an important component of Task-oriented dialogue system. The existing intent classification models follow a closed set assumption and can only identify a limited number of predefined intent types. However, the real world is open. During the online deployment of dialogue system, out-of-domain (OOD) or unknown intents will appear continually. Recently, out-of-domain intent detection task has been widely studied, which can be used to collect these new intent data. The OOD intent discovery task studied in this paper is to make further use of these new intent data. It aims to cluster these OOD samples according to intents, so as to mine new intent types automatically, guide the future development of the system, and expand the classification ability of intent classification models.

References

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149.

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

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. *ArXiv*, abs/2011.01403.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *ArXiv*, abs/2004.11362.

H. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *2021 AAAI Conference on Artificial Intelligence (AAAI)*.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8360–8367.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations.

Padmasundari and S. Bangalore. 2018. Intent discovery through unsupervised semantic text clustering. In *INTERSPEECH*.

Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng Wang, and Lintao Zhang. 2018. Auto-dialabel: Labeling dialogue data with unsupervised learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 684–689.

Nitish Srivastava, Geoffrey E. Hinton, A. Krizhevsky, Ilya Sutskever, and R. Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. *ArXiv*, abs/1511.06335.

Hong Xu, Keqing He, Yuanmeng Yan, Sihong Liu, Zijun Liu, and Weiran Xu. 2020. A deep generative distance-based classifier for out-of-domain detection with mahalanobis space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1452–1460, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *ACL/IJCNLP*.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–878, Online. Association for Computational Linguistics.

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

A Appendix

A.1 Baselines

The details of baselines are as follows:

- **PTK-means** A method based on k-means with IND pre-training. And the IND pre-training objectives uses CE + SCL proposed in this paper.
- **DeepCluster** An iterative clustering algorithm proposed by (Caron et al., 2018), in each iteration, firstly, k-means is used to assign pseudo label to the unlabeled samples, and then the cross-entropy objective is used for representation learning. The cluster header parameters need to be reinitialized during each iteration. In the semi-supervised setting, we use the same IND pre- training objective as DeepAligned (Zhang et al., 2021)
- **CDAC+** The first work of new intent discovery proposed by (Lin et al., 2020), and it firstly pre-trains a BERT-based (Devlin et al., 2019) in-domain intent classifier then uses intent representations to calculate the similarity of OOD intent pairs as weak supervised signals.
- **DeepAligned** The second work of new intent discovery proposed by (Zhang et al., 2021).It is an improved version of DeepCluster. It designed a pseudo label alignment strategy to produce aligned cluster assignments for better representation learning.

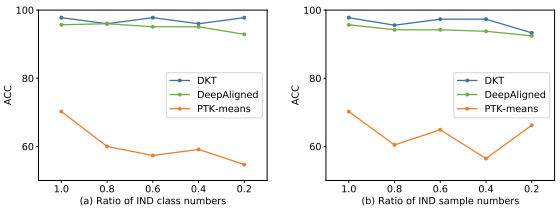


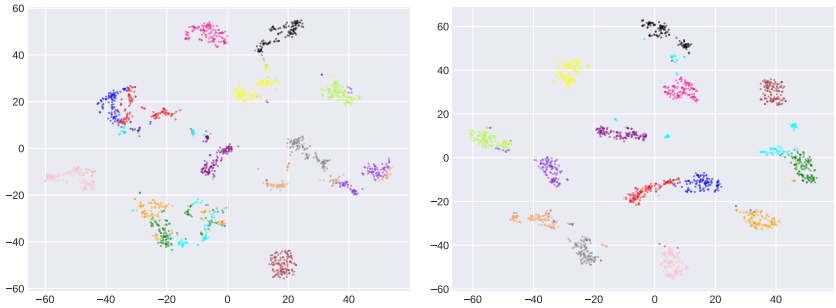
Figure 6: Effect of IND Data.

A.2 Effect of IND Data

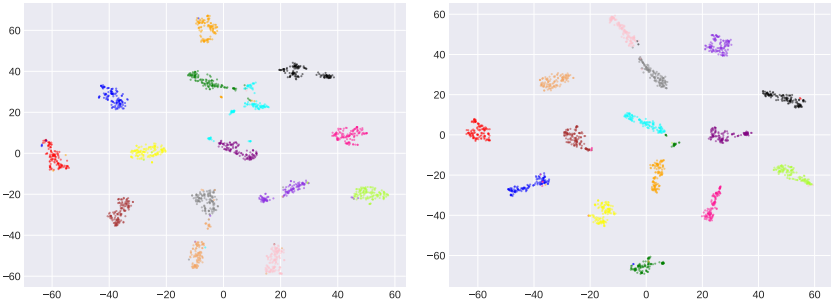
We analyze the effect of IND data for OOD discovery from two perspectives, the number of IND classes and samples per class. Figure 6(a) shows the trend of the number of different IND classes, and Figure 6(b) shows the trend of the number of different samples in each class. Results show DKT outperforms baselines under all settings and gets the smallest varying degrees of performance drop, which proves the robustness and stability of our method.

A.3 Visualization at Different Training Epochs

To see the evolution of our method in the training process, we show a visualization at four different timestamps throughout the training process in Fig 7. Results show representation vector of different intent classes are mixed in the beginning and cluster assignments become increasingly visible and distinct as the training process goes.



(a) Epoch=0 (NMI=19.85) (b) Epoch=4 (NMI=80.29)



(c) Epoch=8 (NMI=92.32) (d) Epoch=14 (NMI=96.97)

Figure 7: OOD intent visualization of different training epochs for our proposed DKT method.