CoCoID: Learning Contrastive Representations and Compact Clusters for Semi-Supervised Intent Discovery

Qian Cao[1] , **Deyi Xiong**[2,*], **Qinlong Wang**[1] and **Peng Xia**[1]
[1]Leyan Tech, Shanghai, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
caoqian0905@gmail.com; dyxiong@tju.edu.cn;
qinlong.wang@leyantech.com; pengxia24@163.com

Abstract

Intent discovery is to mine new intents from user utterances, which are not present in the set of manually predefined intents. Previous approaches to intent discovery usually automatically cluster novel intents with prior knowledge from intent-labeled data in a semi-supervised way. In this paper, we focus on the discriminative user utterance representation learning and the compactness of the learned intent clusters. We propose a novel semi-supervised intent discovery framework CoCoID with two essential components: contrastive user utterance representation learning and intra-cluster knowledge distillation. The former attempts to detect similar and dissimilar intents from a minibatch-wise perspective. The latter regularizes the predictive distribution of the model over samples in a cluster-wise way. We conduct experiments on both real-life challenging datasets (i.e., CLINC and BANKING) that are curated to emulate the true environment of commercial/production systems and traditional datasets (i.e., StackOverflow and DBPedia) to evaluate the proposed CoCoID. Experiment results demonstrate that our model substantially outperforms state-of-the-art intent discovery models (12 baselines) by over 1.4 ACC and ARI points and 1.1 NMI points across the four datasets. Further analyses suggest that CoCoID is able to learn contrastive representations and compact clusters for intent discovery.

1 Introduction

Intent discovery is to pinpoint novel intents from user utterances, which are not present in the set of predefined intents. Discovering novel intents for goal-oriented dialogue has recently attracted growing attention and interest (Lin et al., 2020; Zhang et al., 2021), not only because it is difficult to manually define all potential user intents for conversational agents deployed in a wide range of real-world scenarios, but also due to the cost of curating intent-labelled data.

Model	NMI ↑	ARI ↑
sup-simcse-bert-**large**-uncased	**80.25**	**66.93**
sup-simcse-bert-**base**-uncased	75.09	59.52

Table 1: Comparison results of the DeepAligned model with supervised SimCSE-BERT_base vs. SimCSE-BERT_large on the test set of DBPedia.[1]Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) are widely-used clustering evaluation metrics (see Section 5.1 for more details).

Approaches to intent discovery usually include unsupervised clustering methods (Kathuria et al., 2010; Cheung and Li, 2012; Padmasundari and Bangalore, 2018) and semi-supervised methods (Basu et al., 2004; Hsu et al., 2018; Han et al., 2019). The former performs unsupervised clustering algorithms, e.g., K-means clustering, to group user utterances into clusters according to their underlying intents. The latter attempts to inject weakly supervised signals (Haponchyk et al., 2018; Caron et al., 2018), or external knowledge (Lin et al., 2020; Zhang et al., 2021) into the procedure of clustering. Compared with the unsupervised clustering methods, models that explore supervised signals or external knowledge are capable of discovering better user intents (Zhang et al., 2021). In this paper, we follow the semi-supervised research philosophy for intent discovery.

However, different from previous works (Lin et al., 2020; Zhang et al., 2021), our focus lies on two aspects: user utterance representation and the compactness of clusters. For utterance representation, we have conducted a simple investigation in our preliminary experiments. With the DeepAligned model (Zhang et al., 2021), we

[*] Corresponding author

[1] We run three times of K-means clustering algorithm and report the average results to avoid the unstable issue in K-means clustering.

Proceedings of EMNLP 2022 Industry Track, pages 236–246
December 9–11, 2020. ©2022 Association for Computational Linguistics

have compared two different settings for user utterance representation learning: supervised SimCSE-BERT_base[2] vs. SimCSE-BERT_large[3] (Gao et al., 2021). The results are shown in Table 1, which clearly demonstrate that better user representations lead to more accurate intent discovery. For the compactness of clusters, intuitively, the more compact a cluster is, the more possible user utterances in the cluster share the same intent.

To learn better user utterance representations for clustering and to improve the compactness of learned clusters, we propose CoCoID that learns **Co**ntrastive user utterance representations and **Co**mpact clusters for semi-supervised **I**ntent **D**iscovery. To learn contrastive representations, we define a user utterance to be clustered as the anchor utterance and feed it to the feature extractor twice with different dropout masks. In doing so, we obtain two different representations for the anchor utterance and use them as positive pairs. Other utterances in the same minibatch serve as negative samples. We then perform contrastive learning to improve utterance representations so that utterances with similar underlying intents tend to be close to each other while utterances with dissimilar intents are separated from each other.

To improve the compactness of clusters, we propose an intra-cluster knowledge distillation (ICKD) method. We randomly sample utterances from the cluster where the anchor utterance locates. The intent label of the anchor utterance is used to guide knowledge distillation from the anchor utterance to the sampled utterances. The motivation behind ICKD is to help shorten the distance between utterances in the cluster, which can be also considered as a regularization strategy for the clustering model.

In summary, our contributions are twofold.

1) We propose a novel semi-supervised intent discovery framework CoCoID that learns contrastive user utterance representations and presents an intra-cluster knowledge distillation to improve the compactness of learned intent clusters.

2) We have conducted extensive experiments on four challenging datasets, including CLINC, BANKING, StackOverflow and DBPedia, to examine the effectiveness of the proposed CoCoID. We have achieved significant improvements over state-of-the-art clustering methods by over 1.4 ACC and ARI and 1.1 NMI points. Further analyses demonstrate that the proposed model can indeed learn contrastive representations and compact clusters.

2 Related Work

Our work is related to intent discovery, contrastive learning and knowledge distillation. We briefly review these topics within the constraint of space.

2.1 Intent Discovery

User intent detection is an essential component in dialogue systems. A wide range of approaches, including unsupervised (Padmasundari and Bangalore, 2018), supervised (Hakkani-Tür et al., 2013) and semi-supervised (Basu et al., 2004; Hakkani-Tür et al., 2015; Han et al., 2019) methods, have been explored for intent discovery. In the unsupervised research strand, Kathuria et al. (2010) propose to exploit K-means clustering to understand user intents. In addition to unsupervised clustering approaches, Haponchyk et al. (2018) solve intent discovery by using powerful semantic classifiers to categorize user questions into intents with structured outputs in a supervised way. Caron et al. (2018) propose to produce clustering assignments as pseudo labels and then train a pseudo classifier. For semi-supervised methods, Zhang et al. (2021) investigate the label inconsistent issue and propose a deep alignment strategy. Other semi-supervised studies approach intent discovery by guiding the clustering process with pairwise constraints, such as KCL (Hsu et al., 2018) and CDAC+ (Lin et al., 2020). Our model is also semi-supervised. The significant differences from previous semi-supervised intent discovery lie in the contrastive learning of user utterance representations and intra-cluster knowledge distillation for improving the compactness of clusters.

2.2 Contrastive Learning

The main idea behind contrastive learning is to force semantic representations of similar objects to be close to each other and those of dissimilar objects to be far away from each other, which is widely used in unsupervised visual representation learning. Recent years have witnessed that contrastive learning has also been explored in textual representation learning. Fang and Xie (2020) pro-

[2] https://huggingface.co/princeton-nlp/sup-simcse-bert-base-uncased

[3] https://huggingface.co/princeton-nlp/sup-simcse-bert-large-uncased

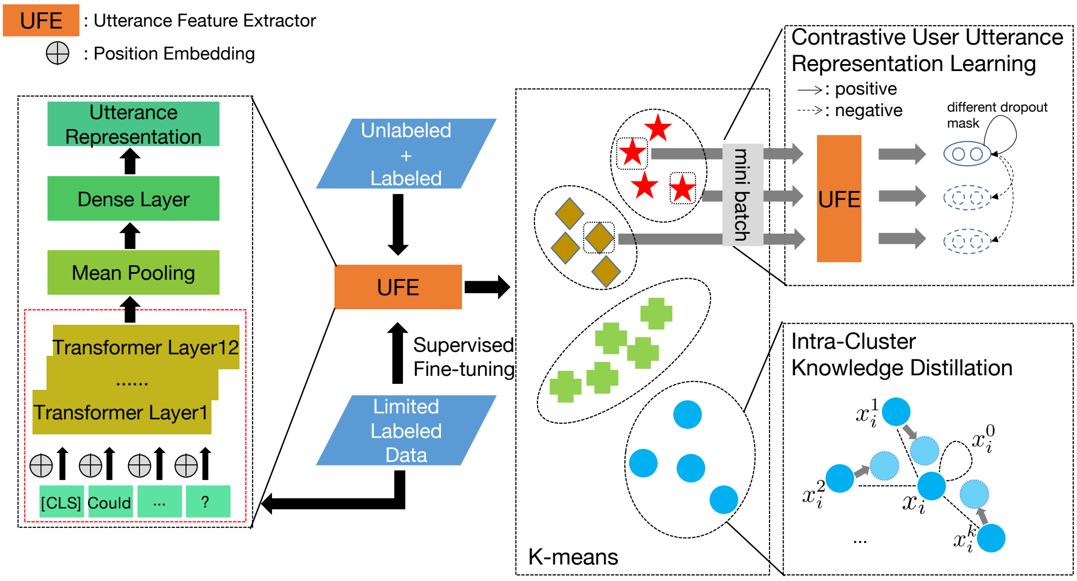


Figure 1: The diagram of CoCoID. We first fine-tune the utterance feature extractor on intent-labeled data. After fine-tuning, we obtain representations for both labeled and unlabeled user utterances. We cluster them via the K-means cluster algorithm. Our contributions lie in the contrastive user utterance representation learning and intra-cluster knowledge distillation used in the semi-supervised clustering procedure.

pose CERT and construct positive and negative utterances through back translation. Gao et al. (2021) find that dropout can act as an efficient data augmentation method for contrastive textual representation learning. Yan et al. (2021) also explore different data augmentation strategies on contrastive learning for sentence representation modeling. In this paper, we follow Gao et al. (2021) to use different dropout masks to create positive samples.

2.3 Knowledge Distillation

Knowledge distillation usually refers to training a large teacher model and distilling its knowledge into a small student model (Hinton et al., 2015). Zhang et al. (2018) propose a strategy where an ensemble of students learn collaboratively and then teach each other. Yuan et al. (2019) find that knowledge distillation is a more general label smoothing regularization and present a teacher-free knowledge distillation framework where a student model learns from itself. Yun et al. (2020) propose a class-wise knowledge distillation method which distills the predictive distribution between different samples of the same label during training, which is similar to our intra-cluster knowledge distillation in the sense of distilling predictive distribution between samples from the same group. The significant difference is that we distill knowledge between user utterances in the same cluster under the semi-supervised setting rather than in the supervised condition for images (Yun et al., 2020).

3 Approach

We elaborate the proposed CoCoID in this section. The diagram of CoCoID is shown in Figure 1. It consists of three major components: semi-supervised clustering backbone, contrastive utterance representation learning and intra-cluster knowledge distillation.

3.1 Semi-supervised Clustering Backbone

The traditional clustering-based approach for intent discovery produces clustering assignments as pseudo labels (Caron et al., 2018), which may result in an inconsistent clustering assignment issue as different labels could be assigned to the same utterance in different training epochs. We therefore use the DeepAligned framework (Zhang et al., 2021) as our semi-supervised clustering backbone, which is composed of three essential components: utterance feature extractor, fine-tuning and DeepAligned clustering (addressing the inconsistent issue). We briefly introduce the first two parts here. More details on the third component can been found in (Zhang et al., 2021).

Utterance Feature Extractor We use BERT (Devlin et al., 2019) to extract features from user utterances. For notational convenience, we define the training corpus as $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ where $\boldsymbol{x}_i$ is the i-th utterance with or without intent label. We assume that $\mathcal{D}$ is composed of $\mathcal{D}_{\text{label}}$ and $\mathcal{D}_{\text{unlabel}}$ that indicate utterances with intent labels and those without manual intent labels, respectively. We feed $\boldsymbol{x}_i$ into BERT and extract the last layer of BERT as $\boldsymbol{H}_i = [\text{CLS}, \boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_m]$ where CLS is a special classification token and m denotes the length of the utterance $\boldsymbol{x}_i$. We then obtain an averaged representation for the corresponding utterance as follows:

$$\boldsymbol{R}_i = f(\boldsymbol{H}_i) \tag{1}$$

where f indicates the meaning-pooling operation. To further enhance the feature extractor, we add a dense layer g:

$$\boldsymbol{h}_i = g(\boldsymbol{R}_i) = \sigma(\boldsymbol{W}_g \boldsymbol{R}_i + \boldsymbol{b}_g) \tag{2}$$

where σ is a ReLU activation function, $\boldsymbol{W}_g$ and $\boldsymbol{b}_g$ are learnable parameters.

Fine-tuning To incorporate prior knowledge from limited intent-labeled data, we follow Zhang et al. (2021) to fine-tune the BERT-based feature extractor on labeled data $\mathcal{D}_{\text{label}}$. Specifically, we stack a simple intent classification layer over the BERT-based feature extractor and fine-tune the extractor with a cross-entropy loss in a supervised fashion. After fine-tuning, we remove the classification layer and keep the rest of the network as feature extractor for later use in the clustering process.

3.2 Contrastive User Utterance Representation Learning

After fine-tuning, we obtain a full-fledged feature extractor. We learn representations of utterances in $\mathcal{D}$ through this feature extractor. Over these learned representations, we perform K-means clustering to categorize utterances into groups. We define groups as $\{G\}_{i=1}^{K}$ where K is the number of groups for clustering.

We define an utterance $\boldsymbol{x}_i$ in question as anchor utterance and its group as $G_{\mu(\boldsymbol{x}_i)}$ where μ is a mapping function that maps $\boldsymbol{x}_i$ to a cluster index $\in [1, K]$. Similar to SimCSE (Gao et al., 2021), we input each anchor utterance $\boldsymbol{x}_i$ to the encoder (as shown in the red dashed box) twice with two different dropout masks d and $d^{'}$. We let $\boldsymbol{h}_i^d$ and $\boldsymbol{h}_i^{d^{'}}$ (calculated according to Eq. 2) denote the representation of $\boldsymbol{x}_i$ with the two different dropout masks. We consider $\boldsymbol{h}_i^{d'}$ as the positive representation to $\boldsymbol{h}_i^d$ for the anchor utterance since they are only different in dropout masks. That is to say, they are semantically similar to each other. Representations of other utterances in the same minibatch are regarded as negative representations to $\boldsymbol{h}_i^d$. In this way, we construct positive and negative representations for contrastive learning. The contrastive learning objective is hence optimized as follows:

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp(\text{sim}(\boldsymbol{h}_i^d, \boldsymbol{h}_i^{d^{'}})/\tau)}{\sum_{j=1}^{N}(\exp(\text{sim}(\boldsymbol{h}_i^d, \boldsymbol{h}_j^{d^{'}})/\tau))} \tag{3}$$

where N is the number of utterances in the minibatch and τ is a temperature hyperparameter. sim is a similarity measurement function, which can be computed as the cosine similarity as follows:

$$\text{sim}(a, b) = \frac{a^\top b}{\|a\| \cdot \|b\|} \tag{4}$$

3.3 Intra-cluster Knowledge Distillation

Dark knowledge can not only be distilled from a large teacher model to a small student model, but also be distillable among continual instances within the same model via self-learning. The latter self-knowledge distillation is feasible as dark knowledge can be regularized to produce the same prediction pattern for instances in the same class (Yun et al., 2020). This inspires us to perform self-knowledge distillation within the same cluster.

Specifically, for each anchor utterance $\boldsymbol{x}_i$, we randomly sample different utterances in $G_{\mu(\boldsymbol{x}_i)}$ (denoted as $\boldsymbol{x}^{in}$) to form multiple $(\boldsymbol{x}_i, \boldsymbol{x}_i^{in})$ pairs. Let $\boldsymbol{u}_i$ denote logit from $\boldsymbol{x}_i$ and $\boldsymbol{u}_i^{in}$ logit from $\boldsymbol{x}_i^{in}$. We then distill dark knowledge from $\boldsymbol{u}_i$ to each sampled logit. Mean square error (MSE) is used as the intra-cluster distillation objective, which is computed as follows:

$$\mathcal{L}_{\text{ICKD}} = \frac{1}{k+1} \sum_{in=0}^{k} \|\boldsymbol{u}_i - \boldsymbol{u}_i^{in}\|_2^2 \tag{5}$$

where k is the number of sampled utterances, which is a hyperparameter to be tuned. Note that $\boldsymbol{u}_i^0$ is logit from the anchor utterance itself with a different dropout mask.

Forcing the predictive distributions (i.e., logit) from sampled utterances to be close to that from the anchor utterance in the same cluster, we want these utterances in the same cluster to be similar to each other. In other words, their distance could be

Model	CLINC			BANKING			StackOverflow			DBPedia		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
KM†	45.06	26.86	70.89	29.55	12.18	54.57	13.55	1.46	8.24	61.00	49.93	67.26
AG†	44.03	27.70	73.07	31.58	13.31	57.07	14.66	2.12	10.62	56.07	43.92	65.63
SAE-KM†	46.75	29.95	73.13	38.92	22.85	63.79	34.44	17.07	32.62	50.29	31.72	59.70
PCK-means†	54.61	35.40	68.70	32.66	16.24	48.22	24.16	5.35	17.26	83.11	71.27	79.76
DEC†	46.89	27.46	74.83	41.29	27.21	67.78	13.09	3.76	10.88	39.60	29.43	53.36
DCN†	49.29	31.15	75.66	41.99	26.81	67.54	34.26	15.45	31.09	47.48	32.31	54.54
DAC†	55.94	40.49	78.40	27.41	14.24	47.35	16.30	2.76	14.71	63.96	56.30	75.37
BERT-KCL†	68.86	58.79	86.82	60.15	46.72	75.21	13.94	7.81	8.84	60.62	61.03	83.16
DeepCluster†	35.70	19.11	65.58	20.69	8.95	41.77	-	-	-	-	-	-
BERT-DTC†	74.15	65.02	90.54	56.51	44.70	76.55	-	-	-	-	-	-
CDAC+†	69.89	54.33	86.65	53.83	40.97	72.25	73.48	52.59	69.84	91.66	89.41	94.74
DeepAligned†	86.49	79.75	93.89	64.90	53.64	79.56	-	-	-	-	-	-
CDAC+‡	70.18	58.68	87.23	54.34	41.86	72.68	74.8	50.75	76.96	91.31	87.47	93.01
DeepAligned‡	86.69	80.19	94.07	65.44	53.76	79.80	78.55	**61.21**	75.97	92.89	89.54	94.04
CoCoID (ours)	**87.51**	**81.35**	94.49	**67.81**	**57.08**	81.32	78.43	60.25	**77.60**	**95.56**	**91.91**	**94.95**
w/o ICKD	86.74	80.31	94.06	66.44	54.83	80.26	**79.10**	60.13	77.46	95.45	91.52	94.72
w/o CL	86.13	80.7	**94.53**	66.52	56.51	**81.75**	78.35	59.04	77.52	93.74	90.34	94.37

Table 2: Clustering results on the four datasets. †: results from Zhang et al. (2021) and Lin et al. (2020). ‡: results that we reproduced.

shortened by the proposed intra-cluster knowledge distillation so that clusters are more compact.

4 Training Objective

Contrastive utterance representation learning attempts to pull utterances with similar semantic intents together and push apart utterances with different intents in a minibatch-wise manner. By contrast, intra-cluster knowledge distillation is to shorten distances of utterances in the same cluster by regularizing predictive distributions in a cluster-wise way. These two approaches are complementary to each other and therefore can be combined in a unified framework CoCoID. The final joint loss of CoCoID with the two components can be formulated as:

$$\mathcal{J} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{ICKD}} \tag{6}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss used in the DeepAligned model (Zhang et al., 2021).

5 Experiments

We conducted a series of experiments on four widely-used datasets to evaluate the proposed CoCoID. More details about the datasets and baselines can be found in the appendix A and B.

5.1 Evaluation Metrics

We used three metrics to evaluate performance in our experiments. Normalized Mutual Information (NMI) is commonly used in measuring the quality of clustering by estimating the similarity of clustering results to the ground-truth results. Adjusted Rand Index (ARI) treats the analysis of clustering as a series of decisions, one for each of the $N(N-1)/2$ pairs of collections. The Rand index measures the percentage of decisions that are correct while ARI is the corrected-for-chance version of it, ensuring that the value for random clustering tends to be 0. In addition to the two metrics, we follow Zhang et al. (2021) to use the Hungarian algorithm to obtain the mapping between the predicted classes and ground-truth classes to estimate clustering Accuracy (ACC). For all metrics, the higher the score is, the better the performance is.

5.2 Settings

To make fair comparisons, we used the same settings that randomly select 10% of the data as the labeled data, and 75% of the intents as the known intents as in previous works (Zhang et al., 2021). The division of the datasets also follows Zhang et al. (2021) and Lin et al. (2020). We first fine-tuned our model on the labeled data and performed contrastive learning and intra-cluster knowledge distillation during clustering. We set the wait patient as 20 to avoid overfitting. We evaluated the cluster performance with Silhouette Coefficient which is an unsupervised metric to evaluate clustering performance. After training, we evaluated the performance on test sets and averaged 10 random-seed results as our final result. Note that we set the num-

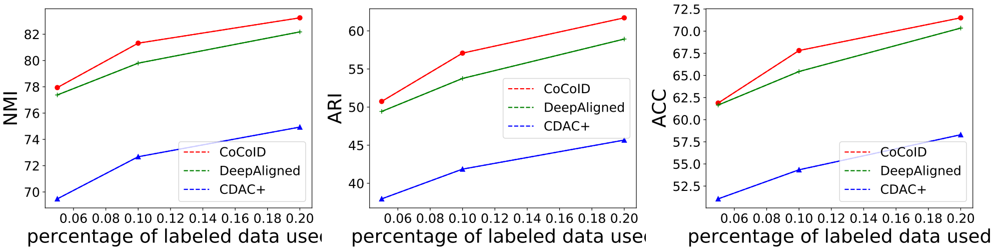


Figure 2: Comparison results along varying labeled data ratios on the BANKING dataset.

ber of intents as ground-truth during clustering, which is consistent with Zhang et al. (2021).

We employed the pre-trained BERT model (base-uncased, with 12-layer Transformer) to build the feature extractor. We set the training batch size as 128, and the learning rate as $5e^{-5}$.The temperature for contrastive utterance representation learning was set to 1.0 for CLINC, StackOverflow and DBPedia dataset. For BANKING dataset, we set the temperature of contrastive learning as 0.05. We also followed Zhang et al. (2021) to freeze all BERT parameters except the last layer to avoid overfitting and speed up the training process. For each anchor utterance, we sampled 3 utterances.

5.3 Main Results

Table 2 shows the results on the four datasets. Our reproduced CDAC+ and DeepAlighed model are better than those reported by Zhang et al. (2021) and Lin et al. (2020) in most cases, indicating strong baselines to be compared. Our proposed CoCoID outperforms the 12 state-of-the-art baselines on the majority of cases on the four datasets. Particularly, we achieve an average improvement of 1.43 ACC, 1.47 ARI and 1.12 NMI over the reproduced DeepAligned (better than the original).

Interestingly, it is worth noting that the length of utterances in datasets seems to have an impact on the performance improvements. The average length of utterances in both CLINC and StackOverflow datasets is relatively shorter than 10, hence the improvement obtained by our proposed model is limited. In contrast, on the BANKING and DBPedia datasets, utterances are longer. The improvements in terms of the three metrics are substantially higher than those on the other two datasets.

5.4 Ablation Study

To examine the effectiveness of the proposed two methods , we further conducted ablation study. In the last two rows of Table 2, we show the results of our model without contrastive learning and intra-cluster distillation. It can be found that the removal of contrastive learning will reduce the ACC of the model to some extent, while the removal of ICKD will negatively affect the NMI and ARI metric. The absense of contrastive learning causes ACC to decrease by 1.14 on average, suggesting that contrastive learning is able to boost intent discovery accuracy. The absence of intra-cluster knowledge distillation results in larger performance drops in CLINC and BANKING than those in StackOverflow and DBPedia. This indicates that ICKD is beneficial to intent discovery with a larger number of novel intents (37/19 vs. 5/4)

6 Analysis

We carried out in-depth analyses to investigate how the proposed methods improve intent discovery.

6.1 Analysis on the Impact of the Percentage of Labeled Data Used

As our CoCoID is a semi-supervised intent discovery approach that utilizes intent-labeled data to fine-tune the feature extractor, we would like to know the impact of the percentages that used intent-labeled data account for among all data on the performance. For this, we conducted experiments on the BANKING dataset by gradually increasing the percentage of labeled data used for fine-tuning from 0.05 to 0.2. Results of CoCoID against the two state-of-the-art models DeepAligned and CDAC+ are illustrated in Figure 2. Under all settings, all the three models benefit from the growing amount of intent-labeled data used. But our CoCoID is always the best among the three models, indicating its strong capacity in exploring intent-labeled data.

6.2 Analysis on the Impact of the Ratio of Unknown Intent Classes

Our approach is able to detect both known (predefined) and unknown (novel) intent classes. There-

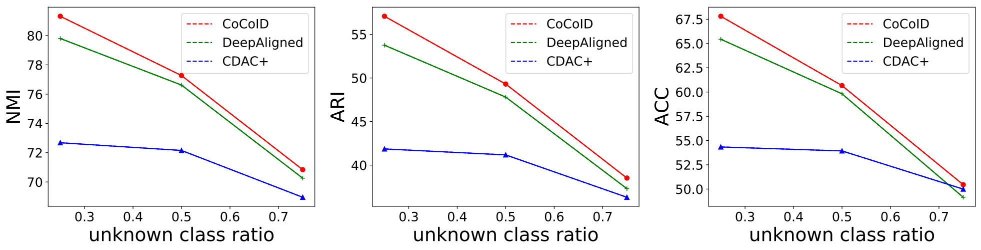
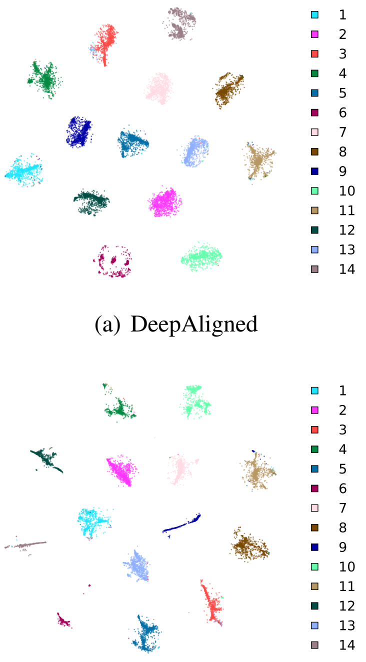


Figure 3: Comparison results along varying known class ratios on the BANKING dataset.



(a) DeepAligned

(b) CoCoID

Figure 4: Visualization of the 14 intents on the DBPedia dataset and 5, 6, 9, 13 are the unknown intents.

fore, we further analyzed the impact of the unknown intent class ratio on the performance. Figure 3 shows the results of CoCoID in comparison to CDAC+ and DeepAligned. The performance of the three models drops when the ratio of unknown intent classes increases, which is reasonable as it is more challenging to detect novel intents without labeled data than predefined intents with labeled data. However, our model achieves smaller performance drops in all three evaluation metrics along growing unknown intent class ratios, suggesting that our model is more capable of detecting novel intents than both DeepAligned and CDAC+.

6.3 Visualization of Intent Distribution

Figure 4 visualizes the distribution of 14 intents in the semantic space of the DBPedia dataset. We merged the training data and test data to generate utterance representations. We then visualized them through t-sne. To obtain a better global structure, we set the perplexity of t-sne to 500. Figure 4(a) visualizes the intent distribution yielded by DeepAligned while figure 4(b) displays results yielded by our CoCoID. Note that the same color across the two sub-figures represents the same intent cluster. From the visualization, we can easily find that the area of intent clusters produced by CoCoID is smaller than that by DeepAlighed. And some clusters are even compacted into strips (e.g., intent 3 and 9 in Figure 4(b)).

In order to explicitly show the degree of compactness of intent clusters, we also calculated the average distance from each utterance to the cluster centroid. The average distance for DeepAligned is 0.34 while only 0.14 for our CoCoID, which strongly suggests that our proposed method does make clusters more compact.

7 Conclusions

In this paper, we have presented a semi-supervised intent discovery framework CoCoID. It consists of two essential components: contrastive user utterance representation learning and intra-cluster knowledge distillation. Extensive experiments and analyses on real-life challenging datasets demonstrate that CoCoID outperforms previous state-of-the-art intent discovery models (12 baselines) and is able to learn contrastive representations and compact clusters.

Acknowledgments

Deyi Xiong was supported by Zhejiang Lab (No. 2022KH0AB01). We would like to thank the anonymous reviewers for their insightful comments.

References

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

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 Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. *CoRR*, abs/2003.04807.

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

Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 383–392.

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

Hongchao Fang and Pengtao Xie. 2020. CERT: contrastive self-supervised learning for language understanding. *CoRR*, abs/2005.12766.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

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

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

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

Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti. 2018. Supervised clustering of questions into intents for dialog system applications. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2310–2321.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

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

Ashish Kathuria, Bernard J. Jansen, Carolyn Theresa Hafernik, and Amanda Spink. 2010. Classifying the user intent of web queries using k-means clustering. *Internet Res.*, 20(5):563–581.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

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.

Padmasundari and Srinivas Bangalore. 2018. Intent discovery through unsupervised semantic text clustering. In *Proceedings of Interspeech*, pages 606–610.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 478–487.

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

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5065–5075.

Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3861–3870.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2019. Revisit knowledge distillation: a teacher-free framework. *CoRR*, abs/1909.11723.

Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13876–13885.

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.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328.

Dataset	#Intents (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
StackOverflow	20 (15 + 5)	18,000	1,000	1,000	17,182	41/9.18
DBPedia	14 (10 + 4)	12,600	700	700	45,077	54/29.97

Table 3: Statistics on the used datasets. Intents are divided into known (predefined) and unknown (novel intents).

A Datasets

Detailed statistics of the four datasets are shown in Table 3.

CLINC is a multi-domain intent classification dataset (Larson et al., 2019), which contains 150 intents and 23,700 utterances across 10 domains. To compare with previous works, we used the same data division as in (Zhang et al., 2021).

BANKING is a single-domain intent classification dataset (Casanueva et al., 2020), which provides a fine-grained set of intents in the banking domain. It contains 13,083 utterances labeled with 77 intents. In our experiments, the split of training, validation and test set follows Zhang et al. (2021).

StackOverflow is an intent classification dataset collected and processed by Xu et al. (2015) from Kaggle.com.[4] Xu et al. (2015) randomly select 20,000 question titles from 20 different labels. In our experiments, the division of training, validation and test set follows Lin et al. (2020).

DBPedia is a DBpedia ontology dataset which contains 14 non-overlapping classes (Zhang and LeCun, 2015). We follow Lin et al. (2020) on data division to compare with their results fairly.

B Baselines

We compared with 12 different baselines, listed as follows.

K-means (KM): an iterative algorithm of clustering, which first randomly selects K objects as the initial cluster centroids and then assigns each object to the nearest cluster.

Agglomerative Clustering (AG): a bottom-up hierarchical clustering method which calculates the distance/similarity between classes.

SAE-KM: a method similar to K-means. The difference is that the feature extractor is a stacked autoencoder (SAE).

PCK-means (Basu et al., 2004): a clustering framework with paired constraints and a new theoretically motivated method that actively selects good paired constraints for semi-supervised clustering.

DEC (Xie et al., 2016): a method that uses self-training target distribution to iteratively optimize clustering targets according to KL divergence.

DCN (Yang et al., 2017): a method that combines a dimension reduction and K-means clustering approach to maintain the advantages of both tasks.

DAC (Chang et al., 2017): a method that converts the clustering problem to a binary pair classification to determine whether a pair of images belongs to the same cluster.

BERT-KCL (Hsu et al., 2018): an approach using predictive pairwise similarity as the knowledge to be transferred, and formulating a learnable objective function to utilize pairwise information in a manner similar to constrained clustering.

DeepCluster (Caron et al., 2018): a clustering method which jointly learns the parameters of the neural network and the cluster assignments from extracted features.

BERT-DTC (Han et al., 2019): a modified variant of DEC (Xie et al., 2016) which can cluster data during learning data representations. The purpose of this change is to allow clustering to be guided by known classes.

CDAC+ (Lin et al., 2020): an end-to-end clustering method that can naturally use paired constraints as prior knowledge to guide the clustering process.

DeepAligned (Zhang et al., 2021): a semi-supervised intent discovery method based on DeepCluster, which uses an alignment strategy to tackle the label inconsistency issue.

Limitations

For the combination of multi-task losses, we interpolate them with equal weights. We believe that there are better strategies to combine these losses

[4] https://www.kaggle.com/c/predict-closed-questions-onstack-overflow/download/train.zip.

which we leave to our future work. Additionally, the impact of query length on the performance is yet to be investigated.