UNICON: Unsupervised Intent Discovery via Semantic-level Contrastive Learning

Anonymous ACL submission

Abstract

Discovering new intents is crucial for expanding domains in dialogue systems or natural language understanding (NLU) systems. A typical approach is to leverage unsupervised and semi-supervised learning to train a neural encoder to produce representations of utterances that are adequate for clustering then perform clustering on the representations to detect unseen clusters of intents. Recently, instance-level contrastive learning has been proposed to improve representation quality for better clustering. However, the proposed method suffers from semantic distortion in text augmentation and even from representation inadequacy due to limitations of using representations of pre-trained language models, typically BERT. Neural encoders can be powerful representation learners, but the initial parameters of pre-trained language models do not reliably produce representations that are suitable for capturing semantic distances. To eliminate the necessity of data augmentation and reduce the negative impact of pre-trained language models as encoders, we propose UNICON, a novel contrastive learning method that utilizes auxiliary external representations to provide powerful guidance for the encoder. The proposed method produces clusters that facilitates intent discovery, achieving state-of-the-art on intent detection benchmarks by a large margin in both unsupervised and semi-supervised settings.

1 Introduction

Intent discovery refers to the problem of finding new intent classes in natural language understanding (NLU) tasks from unlabeled user utterances. The ability to discover new intents is fundamentally important for dialogue systems in industrial practice, because users can be creative in interacting with the system and the user population's interest may change over time with varying degrees depending on the applications. Proactively designing new intents is a labor-intensive process, hence

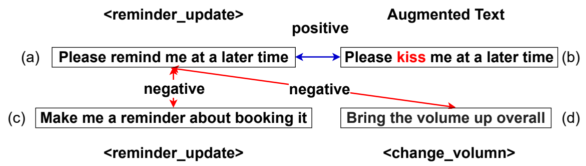

Figure 1: Intance-level contrastive learning concept. (a) is an original text, (b) is augmented from the original text. (c) and (d) are other instances in the same mini-batch. The instance-level contrastive learning keeps the positive sample close and the negative samples away.

a data-driven intent discovery system could drastically reduce the continual intent-designing cost and help keep the user experience more engaging and satisfactory.

Typically, intent discovery is achieved by (1) training a powerful neural encoder, preferrably a pre-trained neural language model such as BERT (Devlin et al., 2018), (2) and performing clustering on the representations produced by the encoder from an unlabeled dataset to detect unseen intent clusters. Training encoders without supervision belongs to the unsupervised clustering family (Hakkani-Tür et al., 2013, 2015; Padmasundari and Bangalore, 2018; Haponchyk et al., 2018; Shi et al., 2018), while semi-supervised clustering utilizes a small amount of intent-labeled data (Lin et al., 2020; Zhang et al., 2021b).

Recent methods leverage deep neural encoders to produce robust and rich representations that can be tailored to produce meaningful clusters via self-supervised learning. Various architectures and training algorithms have been proposed in this regard, namely feature assembly using autoencoders (Shi et al., 2018), pairwise binary classification using instance similarity (Lin et al., 2020), and self-supervised learning with aligned psuedo-labels (Zhang et al., 2021b).

Recently, an instance-level contrastive learning method has attracted much attention. A popular set-up for contrastive learning is the instance-level approach, which trains the encoder to keep the rep-

Intent	Original Text	BERT Augmented Text	RoBERTa Augmented Text
update_playlist	Add this song to shared playlist	Introducing this song to shared playlist	Add this song to shared messages
current_location	My current location	My target location	My current shoes
change_accent	Let's change your accent	Let's change your luck	Let's change your email
cancel	Can you please cancel	Can you please out	Can you please send

Table 1: On the CLINC dataset, we utilize *Contextual Augmenter* (Kobayashi, 2018) which finds the most appropriate words for augmentation by feeding surrounding words to BERT and RoBERTa models. Then, we perform augmentation by inserting them or replacing original words with them. This table shows that augmented text may not preserve the original intent since certain keywords may be changed.

resentations of hard positive samples generated via data augmentation closer to each other in contrast to other negative samples (Chen et al., 2020a; Wu et al., 2020; Giorgi et al., 2020; Grill et al., 2020; Gao et al., 2021; Yan et al., 2021; Kim et al., 2021). Some works proposed to integrate clustering during instance-level contrastive learning to further improve the clustering results. For example, Li et al. (2021) conducts cluster-level contrastive learning on augmented images on top of the instance-level contrastive learning. Zhang et al. (2021a) proposed optimizing both the clustering loss based on KL-divergence and the contrastive learning loss from augmentations.

However, previous works have three limitations. First, the existing instance-level contrastive learning methods do not consider the semantic similarities among data points and sets up positive and negative samples indiscriminately. As shown in Fig. 1, a typical contrastive learning method uses in-batch samples as the negative samples and augmented text as the positive samples. However, the positive sample may not be truly a positive sample as data augmentation perturbations may cause class-inconsistency, while examples that are considered in the same intent category as the main example may end up being chosen as negative samples. This indiscriminative training procedure may cause harm to the ability of the encoder to learn appropriate representations for producing desired clustering results.

Second, the data augmentation techniques used in previous works (Zhang et al., 2021a; Yan et al., 2021; Wu et al., 2020; Zang et al., 2020) can cause semantic distortion, which results in intent inconsistency in augmented texts. To illustrate semantic distortion, we showcase examples before and after the text augmentation method described in Zhang et al. (2021a) on CLINC dataset. As shown in Table 1, the augmentation may produce perturbed utterances that have different intent classification from the original utterance. The tendency to produce intent-inconsistency samples of text augmentation techniques can be particularly harmful in short utterance intent classification tasks, as there is a higher chance of substituting intent-sensitive keywords in the utterance.

Finally, the typical choice for deep neural encoding (e.g., BERT) may not adequately produce representations that capture semantic distances, greatly increasing the risk of falling into local optima. This phenomenon has been observed in previous studies (Kim et al., 2021; Hu et al., 2020), especially when the `[CLS]` embedding is used as the representation for the entire text or utterance. Our ablation studies (Table 4) also support the idea that naive adoption of BERT as the feature extractor has a detrimental effect in learning clustering-friendly representations, scoring merely 2.82 in the ARI evaluation measure for CLINC.

To alleviate aforementioned problems, we propose a novel contrastive learning that (1) does not require an explicit data augmentation technique, (2) improves representation quality through similarity-based contrastive learning, and (3) circumventing the BERT representation issue via external auxiliary similarity measures.

Using similarity-based pseudo positive samples predicted by insufficiently trained model is extremely unstable because the pseudo-labels may not be correctly selected. The noise caused by incorrect selection accumulates as the training progresses. To mitigate this problem, we propose to adopt auxiliary representations that indicate the presence of words regardless of order. We show the effectiveness of the auxiliary representation and describe the details in Section 3.2.

In summary, our main contributions are as follows:

- We propose a novel contrastive learning method for clustering, called UNICON. This method can conduct semantic-level contrastive learning without data augmentation, which does not suffer from semantic distortion. In addition, the intra-cluster distance could be reduced by selecting two different instances inside the batch as a positive pair, which helps generate proper representations for clustering.

- We propose to use auxiliary representations. An insufficiently fine-tuned PLM may extract positive samples overconfidently, which leads to training failure. The auxiliary representations can mitigate this problem by guiding the model to extract appropriate positive samples.

- To show the effectiveness of our model, we conduct experiments on two intent detection datasets (i.e., CLINC, BANKING). The proposed model outperforms the state-of-the-art model by a large margin of 10-12% in unsupervised setting and 2.5-12% in semi-supervised setting.

2 Related Works

2.1 Intent Discovery

In general, intent detection is a task in dialogue system that tries to find the corresponding intents from the user utterances in a supervised manner when intent structure and the annotated data are given. Then the model classifies an user utterance into a predetermined intent structure. In contrast, the intent discovery task means finding or classifying new intent structures by grouping user utterances of similar meaning in an environment without intent structure or annotated data. Many methods (Hakkani-Tür et al., 2013, 2015; Padmasundari and Bangalore, 2018; Haponchyk et al., 2018; Shi et al., 2018; Lin et al., 2020; Zhang et al., 2021b; Perkins and Yang, 2019; Min et al.; Vedula et al., 2020) have been proposed to solve the intent discovery problem, and approaches through unsupervised or semi-supervised clustering have generally been used.

2.2 Deep Clustering

Since mid 1900's, as an attempt to extract meaningful information from the unlabeled data, clustering task has been actively studied (MacQueen et al., 1967; Gowda and Krishna, 1978; Ester et al., 1996). However, traditional clustering methods suffer with the high-dimensional data due to their lack of ability to learn the proper representation of the data.

Development of Deep Neural Network (DNN) brought strong representation ability. Especially, pre-trained language models (PLM) such as BERT show impressive representation quality with the general language data. This representation ability of DNN is vigorously utilized and studied in clustering methods as follows: DEC (Xie et al., 2016), DCN (Yang et al., 2017), DAC (Chang et al., 2017) and DeepCluster (Caron et al., 2018).

Moreover, some methods use a small number of labeled data and incorporate weak supervised signal to tackle the intent discovery task. CDAC+ (Lin et al., 2020) uses labeled data to help making binary similarity pseudo-labels. DeepAligned (Zhang et al., 2021b) pretrains the labeled data to better estimate the number of the clusters.

2.3 Contrastive Learning

In addition to PLM, contrastive learning (Becker and Hinton, 1992; Xie et al., 2020; Berthelot et al., 2019), which is a component of self-supervised learning, reports many successes in recent years. Contrastive learning aims to group similar samples closer and separate dissimilar samples far from each other. Especially, augmentation-based instance-level contrastive learning is showing many prominent results in computer vision tasks (He et al., 2020; Chen et al., 2020a,b; Grill et al., 2020) and natural language processing (NLP) tasks (Fang et al., 2020; Wu et al., 2020; Zhang et al., 2021a; Yan et al., 2021; Gao et al., 2021; Li et al., 2021; Kim et al., 2021). In particular, Contrastive Clustering (Li et al., 2021) and SCCL (Zhang et al., 2021a) integrate with the cluster-promoting objective function to generate better representation for clustering.

3 Proposed Method

In this section, we describe how our proposed method works in detail. As shown in Fig. 2, we first encode the data into dense contextual representations while constructing auxiliary representations. Second, we generate similarity matrix, which indicates whether a pair of instances belongs to the same cluster. Finally we select a positive sample from each row of the matrix and train the model with contrastive loss.

3.1 Input Representation

In order to extract the high-level semantic features of data, we use the pre-trained language model (PLM) (e.g., Devlin et al., 2018; Liu et al., 2019). Given N samples, $\{\mathbf{X}_i\}_{i=1}^{N}$, we construct inputs for PLM with the special tokens (e.g., [CLS], [SEP])

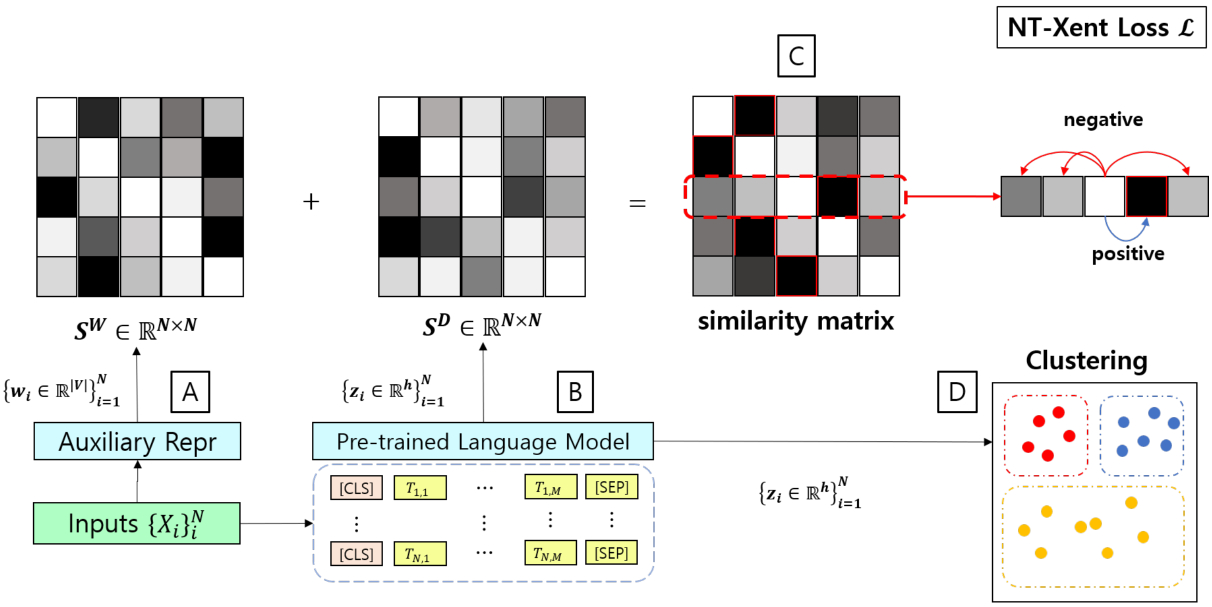

Figure 2: Overview of our proposed method UNICON. (A), (B) Given data, auxiliary representation and contextual representation are extracted from sparse word representation (i.e., TF-IDF) and PLM respectively during the training process. (C) We construct the similarity matrices for each representation and use weighted sum of them as a final matrix. We select a pseudo-positive sample in each row of a final matrix and train the model with the contrastive loss. (D) After the training, we extract representations from the trained PLM and apply various clustering algorithms.

and provide them to the PLM. PLM outputs the features ($\mathbf{z}_i$) as

$$\begin{aligned} I_i &= \text{[CLS]}\ \ T_{i,1},\ \ \ldots,\ \ T_{i,M}\ \ \text{[SEP]} \\ \mathbf{z}_i &= PLM_{\text{CLS}}(I_i) \in \mathbb{R}^h, \end{aligned} \tag{1}$$

where '[CLS]', '[SEP]' are special tokens that represent the entire sentence and distinguish the sentences, respectively. $\{T_{i,k}\}_{k=1}^M$ denotes the set of tokens of $\mathbf{X}_i$, M is the number of tokens, and $PLM_{\text{CLS}}(\cdot)$ indicates the last hidden state vector corresponding to the '[CLS]' token.

3.2 UNICON

Unlike previous works, we aim for adopting semantic-level contrastive learning method without any data augmentation techniques that can lead to semantic distortion. Let $\{\mathbf{z}_i\}_{i=1}^N$ be the set of dense contextual representations of $\{\mathbf{X}_i\}_{i=1}^N$. We compute the similarity matrix which indicates whether a pair of instances belongs to the same intent (cluster), i.e.,

$$\mathbf{S}_{ij}^D = \begin{cases} -\texttt{inf}, & \textbf{if } i = j \\ \text{sim}(\mathbf{z}_i, \mathbf{z}_i), & \textbf{otherwise}, \end{cases} \tag{2}$$

where `inf` is an infinite number that prevents choosing the same instance as a positive pair, $\mathbf{S}^D$ denotes the similarity matrix that has the $N \times N$ dimensions, and $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ indicates the similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$. In this paper, we use the dot product of representations without the normalization and dimensionality reduction as the similarity function.

Subsequently, the sample most similar to the $\mathbf{X}_i$, except for itself, is denoted as a positive sample and the rest of the samples become negative samples. We use the NT-Xent (the normalized temperature-scaled cross entropy) loss function used in Chen et al. (2020a) as follows:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^N \mathbb{1}_{[k\neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau))}, \tag{3}$$

where $\mathbb{1}_{[k\neq i]} \in \{0, 1\}$ is an indicator function that yields 1 if $k \neq i$, and 0 elsewhere, and τ is a temperature parameter that can help the model to learn from hard negatives.

As a result of Eq. 2 and 3, our method, unlike instance-level contrastive learning, can learn suitable features for clustering by explicitly grouping the data instances that have the same intent.

Auxiliary representation Our method has an advantage over augmentation-based instance-level contrastive learning. Augmentation-based contrastive learning pushes different instances apart regardless of their semantic similarities (Zhang et al., 2021a) while our method groups different instances together, taking semantic similarities into account.

However, extracting correct positive samples from unlabeled data using only similarities between the representations of data that are not fine-tuned is a challenging problem. In the early stage of the training, PLM has not learned enough about the target domain yet and may output the vectors that do not represent instances enough. This is likely to result in the incorrect similarity calculation which leads to the erroneous positive sample selection.

Incorrect selection of positive samples in the early stage can cause noise in the learning, which accumulates as the training progresses. As a result, the model performance can deteriorate.

In order to alleviate this problem, we propose to use auxiliary representations that can complement the dense contextual representations. In this paper, we leverage sparse word representations (e.g., BoW, TF-IDF, etc.), which mainly focus on the presence or absence of words and the importance of words within the dataset, ignoring the order of words.

These representations explicitly indicate similarity between instances regardless of their semantic meaning by comparing word frequency. Similarity based on the word frequency can guide model to select appropriate positive samples in the early stage of the training. As a result, the auxiliary representations complement our method by reducing noise in the early stage of the training. The auxiliary representations are used as below:

$$\mathbf{w}_i = \texttt{Aux}(\mathbf{X}_i) \in \mathbb{R}^{|V|},$$
$$\mathbf{S}^W_{ij} = \begin{cases} -\texttt{inf}, & \textbf{if } i = j \\ \text{sim}(\mathbf{w}_i, \mathbf{w}_i), & \textbf{otherwise}, \end{cases} \tag{4}$$
$$\mathbf{S}_{ij} = \mathbf{S}^D_{ij} + \gamma^e \lambda S^W_{ij},$$

where $|V|$ is the vocabulary size and γ is a hyper-parameter that reduces the influence of the auxiliary representation every epoch (e). λ adjusts the scale between $\mathbf{S}^D$ and $\mathbf{S}^W$, which is computed as $\lambda = \texttt{std}(\mathbf{S}^D)/\texttt{mean}(\mathbf{S}^W)$.

Clustering Our model learns the features suitable for the clustering with the target of grouping instances that have the same intent together. Then, diverse clustering algorithms can be used. For example, KMeans (Lloyd, 1982) algorithm can be one of the algorithms, which optimizes the following cost function:

$$\min_{\mathbf{W}\in\mathbb{R}^{h\times K}, \{\mathbf{s}_i\in\mathbb{R}^K\}} \sum_{i=1}^{N} ||\mathbf{z}_i - \mathbf{W}\mathbf{s}_i||_2^2$$
$$\text{s.t.} \quad \mathbf{s}_{i,j} \in \{0,1\}, 1^\intercal \mathbf{s}_i = 1 \quad \forall i, j, \tag{5}$$

where K is the predefined number of clusters, $\mathbf{s}_i$ is the assignment vector which has only one non-zero element, $\mathbf{s}_{i,j}$ denotes the jth element of $\mathbf{s}_i$, and kth column of $\mathbf{W}$ indicates the centroid of the kth cluster.

4 Experiments

4.1 Datasets

We conduct experiments on the CLINC and BANKING datasets, which are intent detection benchmark datasets. CLINC (Larson et al., 2019) covers 150 intents over 10 domains. BANKING (Casanueva et al., 2020) is a fine-grained dataset in the banking domain. Detailed information on the datasets is in Table 2.

Dataset	# of intents	Training	Validation	Test
CLINC	150	18,000	2,250	2,250
BANKING	77	9,003	1,000	3,080

Table 2: The statistics for CLINC and BANKING datasets.

4.2 Baselines

We used various unsupervised clustering and semi-supervised clustering algorithms as the baseline. Additionally, we compare UNICON and clustering methods integrating with instance-level contrastive learning.

Unsupervised Clustering The scores of K-Means (KM) (Lloyd, 1982), agglomerative clustering (AG) (Gowda and Krishna, 1978), stacked autoencoder with K-Means (SAE-KM) (Vincent et al., 2010), DEC (Xie et al., 2016), DCN (Yang et al., 2017), and DeepCluster (Caron et al., 2018) are directly reported in DeepAligned (Zhang et al., 2021b).

Semi-supervised Clustering CDAC+ (Lin et al., 2020) and DeepAligned (Zhang et al., 2021b), which mainly focus on intent discovery tasks, were used as the baselines and reproduced using publicly released code.

Contrastive Learning We reproduced the Contrastive Clustering (Li et al., 2021), SimCSE (Gao et al., 2021) and SCCL (Zhang et al., 2021a) by using publicly released code. Since Contrastive Clustering is a clustering model proposed in vision domain, we adapt it appropriately to text domain by replacing backbone model to *bert-base-uncased*, and augmentation method to *Contextual Augmenter* (Kobayashi, 2018), which is an augmentation method applied in SCCL. SimCSE (sup) and SCCL (sbert) leverage labeled NLI datasets for fine-tuning and pre-training, respectively. Otherwise, SimCSE (unsup) and SCCL (bert) are initilized with *bert-base-uncased* for comparing UNICON.

4.3 Evaluation Metric

To compare our model to the baselines, we use three metrics that are mainly used for clustering performance evaluation, i.e., Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Accuracy (ACC). Since the indices of the clusters are randomly allocated, we measure the accuracy using Hungarian algorithm that can align the cluster indices with label indices.

4.4 Implementation Details

We use a pre-trained BERT model (*bert-base-uncased*, with 12-layer transformer and 110M parameters) as a backbone model without any additional layers in a single P40 GPU. In the code, we use Huggingface's Transformers pytorch library[1]. To extract the auxiliary representations, we utilize the unigram TF-IDF. We use training learning rate of $1e^{-4}$, 10% warmup steps and learning rate decay to optimize the parameters. We set temperature τ to 0.5, γ to 0.9, batch size to 1024/450 on the CLINC and BANKING datasets, respectively. The model is trained and evaluated three times. All reported values in figures and tables are the average performance on the test set.

5 Results and Analysis

Table 3 shows the results comparing our method with the baselines. Our method consistently outperforms the baselines. In terms of accuracy, we achieve a new state-of-the-art performance by a large margin of approximately 10-12% over the closest competitors, i.e. SimCSE (sup) and SCCL (sbert) even though the closest competitors utilized additional resources such as labeled data. The reason for relatively low performance on BANKING dataset is that CLINC dataset consists of a balanced

[1]https://huggingface.co/transformers/index.html

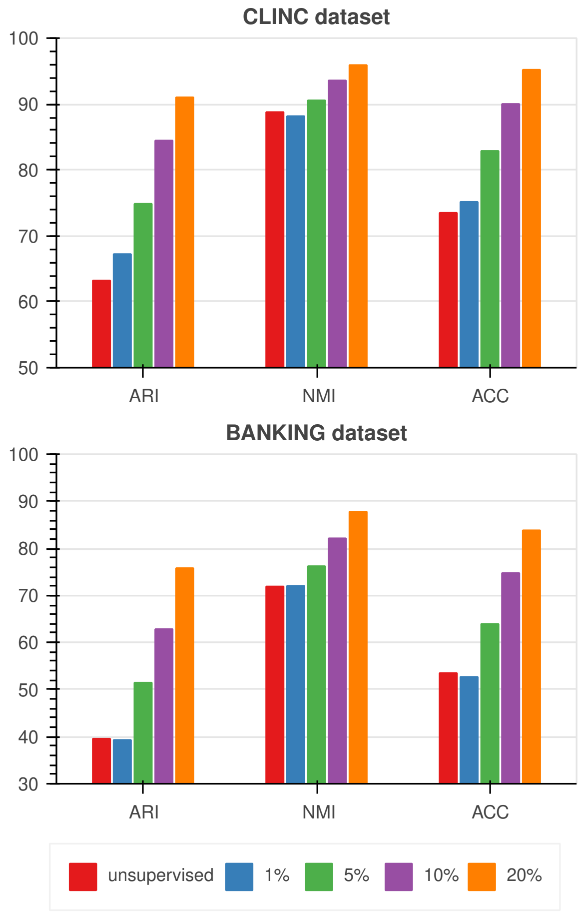

Figure 3: Influence of labeled data ratio on CLINC (first row) and BANKING (second row) datasets. Using only 5% labeled data can improve performance by about 8%, and finally using 20% labeled data improve performance by about 24%.

number of data for each intents, while BANKING does not.

5.1 Semi-supervised Clustering

In this study, we conduct experiments to see the effect that assistance of a few labeled data brings. For the fair comparison, all semi-supervised methods use 10% of labeled data and we assume that all classes are known. Table 3 shows the comparison results. When compared with the baselines, UNICON outperforms competitors by 12% on BANKING dataset and 2.5% on CLINC dataset. UNICON shows relatively lower performance improvement in semi-supervised setting. We speculate the decrease in the effect of auxiliary representation as a reason. Since labeled data already gives enough guidance for the positive selection, auxiliary representation does not help the model as much as in unsupervised setting.

Furthermore, we study how the performance

Setting	Method	CLINC			BANKING		
		NMI	ARI	ACC	NMI	ARI	ACC
Unsupervised	KM‡	70.89	26.86	45.96	54.57	12.18	29.55
	AG‡	73.07	27.70	44.03	57.07	13.13	31.58
	SAE-KM‡	73.13	29.95	46.75	63.79	22.85	38.92
	DEC‡	74.83	27.46	46.89	<u>67.78</u>	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
	SimCSE (unsup)	78.27	39.61	56.27	56.36	20.53	34.84
	SimCSE (sup)	<u>81.84</u>	<u>47.84</u>	<u>61.16</u>	61.61	24.89	38.90
	Contrastive Clustering	71.76	26.04	38.67	34.47	4.48	14.03
	SCCL (bert)	72.69	29.3	45.1	50.22	14.97	28.7
	SCCL (sbert)	81.61	46.74	60.3	64.2	<u>29.33</u>	<u>43.9</u>
	UNICON (ours)	**88.78**	**63.23**	**73.49**	**71.90**	**39.57**	**53.51**
Semi-supervised (ratio=10%)	CDAC+	86.65	54.33	69.89	72.25	40.97	53.83
	DeepAligned	**94.65**	<u>82.16</u>	<u>88.53</u>	<u>78.96</u>	<u>51.66</u>	<u>62.50</u>
	UNICON (ours)	<u>93.58</u>	**84.46**	**91.01**	**82.13**	**62.83**	**74.75**

Table 3: Clustering performance comparison between UNICON and baselines. We evaluate both unsupervised and semi-supervised methods on the test set of CLINC and BANKING datasets. In case of semi-supervised setting, we leverage 10% labeled data. The highest performance is in bold, and the second highest performance is underlined. Methods with ‡ indicate that we directly report the scores from the corresponding paper, and the rest of the methods are reproduced using official code

changes as we use different ratio of labeled data. The experiment results are shown in Fig. 3. Consequently, the performance improves as more labeled data is used. Especially, utilizing 5% of labeled data increases by about 8 points. On the other hand, there is no significant change in performance when we add 1% of labeled data because if 1% of utterances are sampled, it is very unlikely for utterances with the same intent to appear together withing a mini-batch.

5.2 Auxiliary Representation Study

Ablation Study We carry out ablation studies to show the importance and complementarity of each component. First, Fig. 4 shows what the training process looks like when the auxiliary representation is removed. Since the loss of `PLM-only` is very low, it seems like the training is going well. However, we can observe that the actual accuracy decreases as the training progresses. This phenomenon is caused by the accumulation of noise coming from the incorrect positive sample selection. Second, as shown in Table 4, the clustering accuracy is 51.11% when PLM is removed and 15.56% when the auxiliary representation is removed, which is much lower than the accuracy of UNICON. This implies that each model cannot be used for standalone and complements each other. We conjecture that since PLM based representations concentrate on grasping the semantics and the auxiliary representations concentrate on grasping the existence of the specific words, each conveys different information and complements each other.

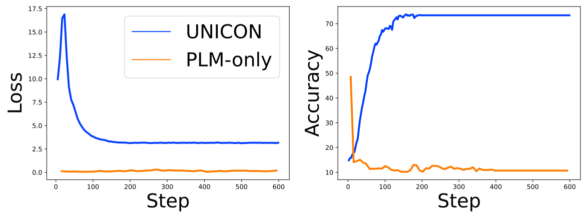

Figure 4: Losses (*left*) and clustering accuracies (*right*) when auxiliary representation is included across the training process and when it is not. We perform experiments on CLINC dataset.

Various Auxiliary Representations We study several representation methods to compensate the noise that comes from the incomplete representation ability of PLM at the early stage of training. We assume that the word representations can complement the contextual representations due to the nature of the intent detection datasets used in dialogue systems. The datasets consist of short utterances and the utterances in the same intent share many keywords with each other. As shown in Table 4, all word representations consistently improve the performance of the model. In particular, TF-IDF method achieves the best performance. The GloVe word embedding model has relatively lower performance than others. This means that the presence or absence of specific keywords has more helpful information, as mentioned above.

Method	NMI	ARI	ACC
PLM + TFIDF (ours)	88.78	63.23	73.49
TFIDF-only	74.49	25.63	51.11
PLM-only	49.89	2.82	15.56
PLM + BoW	79.86	42.62	56.31
PLM + GloVe	78.93	40.02	53.16

Table 4: The experiment results about the auxiliary representations. The experiments are conducted on CLINC dataset.

5.3 Clustering Quality Analysis

We raised the problem of instance-level contrastive learning through data augmentation in Section 1. To show that UNICON can generate more suitable representations for clustering than instance-level contrastive learning-based models, we utilize t-SNE visualization tools on SimCSE, SCCL, Contrastive Clustering and UNICON. As shown in Fig 5, SimCSE and SCCL that utilizes data augmentation does not group data with the same label together nor spread data with the different labels apart. In the case of the Contrastive Clustering, which leverages not only data augmentation but also cluster-promoting objective, clusters data better than SimCSE and SCCL. However, many clusters contain data with various labels which leads to low accuracy. Unlike the other three models, the results of UNICON show that each cluster is well grouped, and the data in each cluster have consistent labels.

Additionally, we measure intra-cluster distance of each model. Intra-cluster distance calculates the euclidean distance between the centroid of the cluster and the data within the cluster, which evaluates how well the model agglomerates the clusters. As depicted in Fig. 6, the intra-cluster distance of UNICON has the lowest average value, followed by Contrastive Clustering and SCCL with clustering-promoting objective, and SimCSE has the worst performance.

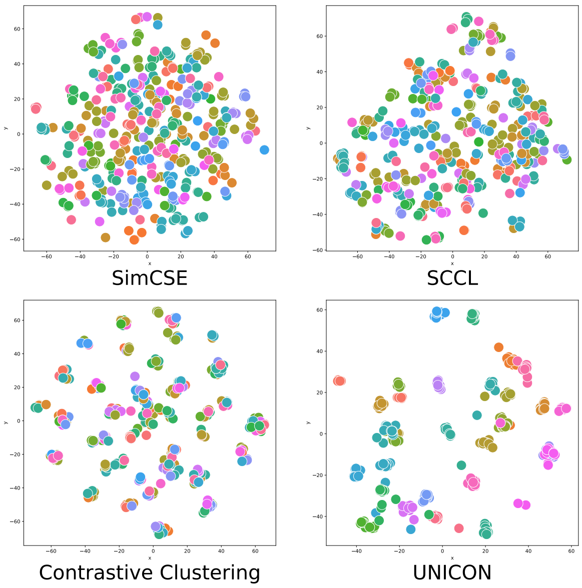

Figure 5: We compare UNICON and other baselines with the contrastive learning through t-SNE visualization. We randomly sample 30 intents from CLINC dataset.

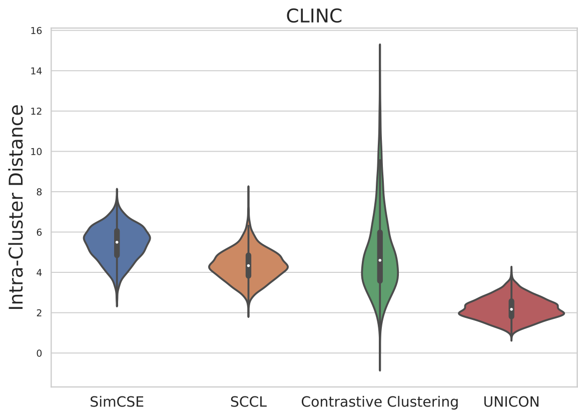

Figure 6: Intra-cluster distance distribution of each model on CLINC dataset.

6 Conclusion

In this work, we propose a clustering method that utilizes power of contrastive learning. To avoid the semantic distortion problem in language data augmentation, we propose to pair an instance with another instance based on the similarity measure. Additionally, we introduce auxiliary representation which guides the model to select appropriate positive pair at the early stage of the training. Extensive experiments on two challenging benchmark datasets report significant improvement in the both unsupervised and semi-supervised clustering performance compared to the baselines. In the future, we plan to study methods to select more robust positive samples with various datasets.

References

Suzanna Becker and Geoffrey E Hinton. 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

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

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proc. the 2nd Workshop on Natural Language Processing for Conversational AI*.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2017. Deep adaptive image clustering. In *Proc. the IEEE international conference on computer vision (CVPR)*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. the Second International Conference on Knowledge Discovery and Data Mining (KDD)*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *Proc. the Conference on Empirical Methods in Natural Language (EMNLP)*.

John M Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

K Chidananda Gowda and G Krishna. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Dilek Hakkani-Tür, Asli Celikyilmaz, Larry Heck, and Gokhan Tur. 2013. A weakly-supervised approach for discovering new user intents from search query logs. In *Proc. the Conference of the International Speech Communication Association (INTERSPEECH)*.

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

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. In *Proc. Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proc. the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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 *Proc. the 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 *Proc. the 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 *Proc. the AAAI Conference on Artificial Intelligence, (AAAI)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. the fifth Berkeley symposium on mathematical statistics and probability*.

Qingkai Min, Libo Qin, Zhiyang Teng, Xiao Liu, and Yue Zhang. Dialogue state induction using neural latent variable models. In *Proc. the International Joint Conference on Artificial Intelligence, IJCAI-20*.

Padmasundari and Srinivas Bangalore. 2018. Intent discovery through unsupervised semantic text clustering. In *Proc. the Conference of the International Speech Communication Association (INTERSPEECH)*.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. In *Proc. the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

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 *Proc. the Conference on Empirical Methods in Natural Language (EMNLP)*.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open intent extraction from natural language interactions. In *Proc. the Web Conference*.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research (JMLR)*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proc. the International Conference on Machine Learning (ICML)*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

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

Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proc. the International Conference on Machine Learning (ICML)*.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proc. the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shangwen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew Arnold, and Bing Xiang. 2021a. Supporting clustering with contrastive learning. In *Proc. the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. In *Proc. the AAAI Conference on Artificial Intelligence, (AAAI)*.