Enhancing New Intent Discovery via Robust Neighbor-based Contrastive Learning

Zhenhe Wu[1,2,3], *Xiaoguang Yu*[3], *Meng Chen*[3*], *Liangqing Wu*[3], *Jiahao Ji*[1], *Zhoujun Li*[1,2*]

[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]State Key Lab of Software Development Environment, Beihang University, Beijing, China
[3]JD AI Research, Beijing, China

{wuzhenhe,jiahaoji,lizj}@buaa.edu.cn
{cdyuxiaoguang,chenmeng20,wuliangqing}@jd.com

Abstract

New intent discovery (NID) has become a hot topic for dialogue system, which aims to discover the Out-Of-Domain intents from conversation corpus and classify these utterances correctly. Existing methods usually focus on learning compact representations of utterances, and leverage the clustering algorithm to generate new intents. Inspired by the recent progress of contrastive learning, in this work, we propose a novel neighbor-based contrastive learning (NCL) model to obtain clustering-friendly representations for utterances. Specifically, to enhance the robustness of NCL, on the one hand, we pick out diverse samples as positive pairs by considering both the anchor neighborhood and nearby neighborhood. On the other hand, we also devise a boundary distance constraint to avoid introducing noisy samples when extending the positives via neighbors. Extensive experiments are conducted on three public NID datasets and the results demonstrate the competitiveness and effectiveness of our proposed approach.

Index Terms: new intent discovery, contrastive learning, clustering

1. Introduction

With the development of conversation AI applications in recent years, a large number of researchers employ user dialogues and partial known intentions to train an intent recognition model, for the reason to design an intelligent natural language understanding system. However, the intentions of user utterances are rich and diverse, and may expand continuously over time, so we can't obtain the universal set of intents and label all the utterances in advance. We raise two examples in Figure 1(a) to illustrate the problem above. In the first sample, intent recognization model analyses the intent of user utterance from a known intent set, and detects the user wants to book flight. Then, the chatbot gives out appropriate response. While in the second sample, the model searches the known intent set and can't obtain the matching intent, which means the intent of this sample belongs to the unknown intent set. Therefore, new intent discovery (NID) for intent recognition has become an important problem. To solve this issue, we need to discover new intents from unlabeled utterances, and group these unlabeled utterances into known and newly discovered intents.

In recent years, researchers paid much attention to NID and similar tasks. Early methods conducted representation learning and clustering algorithm (i.e k-means) to discover new intents [1, 2, 3, 4, 5]. Some researchers learned representations from unlabeled data [1, 2], while some works provided some known intents to support the discovery of unknown intents [3, 4, 5].

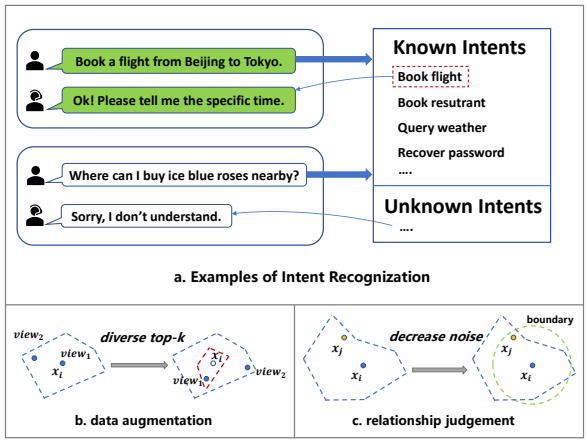

Figure 1: *Task illustration and comparison of different neighbor-based contrastive learning methods.*

Recent works used contrastive learning (CL) to learn sentence embeddings better [6, 7, 8, 9, 10]. Some researchers proposed to pre-train a universal sentence encoder by contrasting a randomly sampled text segment from nearby sentences [8], some other studies adopted a multi-head contrastive learning framework to perform knowledge transfer [9, 10]. More recently, the use of cluster-based contrastive learning framework achieved further improvement [11, 12]. Some researchers showed that combining a contrastive loss with a clustering objective can improve short text clustering [11], while another study used a kNN-based contrastive learning model with a multi-task pre-training process [12].

Although cluster-based contrastive learning methods have made great progress in NID task, two issues about the lack of robustness still remain: 1) In prior study, researchers introduce one top-k parameter to form initial positive pairs in data augmentation stage, which is not fine-grained enough. 2) Previous studies didn't use independent parameters to judge positive or negative relations of augmented data in minibatch, which would introduce noise into training. To improve the robustness of training, we bring up corresponding methods to handle the issues above. As showed in Figure 1(b), we pick out diverse top-k parameters into contrastive learning, which enhances the robustness in data augmentation stage. Specifically, we design an anchor neighborhood and a nearby neighborhood for each utterance to get two views of augmented data. Then in Figure 1(c), we devise a boundary distance constraint combined with a judging neighborhood to determine the positive or negative relations in each minibatch, which is helpful to decrease noise.

*Corresponding authors.

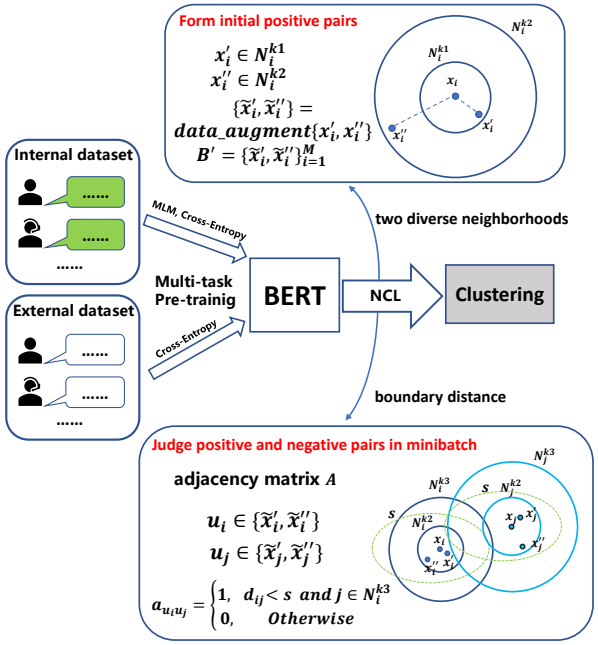

Figure 2: *Framework of the proposed framework.*

To summarize, we propose a novel neighbor-based contrastive learning (NCL) model to obtain clustering-friendly representations for utterances, then we apply k-means for clustering. Extensive experiments are conducted on three public NID datasets and the results indicate that our proposed model obtains the new state-of-the-art performance.

2. Methods

2.1. Problem Definition

To develop an intent recognition model, we define the set of all intents as $I = \{I_k, I_u\}$, which consists of known intents I_k and unknown intents I_u, corresponding to the dataset $D_{internal}^{all} = \{D_k, D_u\}$. In the real scenario, we usually have only a few labeled utterances for each known intents, so we have $D_k = \{D_k^l, D_k^u\}$, which presents labeled and unlabeled part of D_k. The goal of our task is to identify the unknown intents I_u in D_u, and perform correct clustering. Previous works defined "known class ratio" (KCR), while KCL=0 means unsupervised NID, and KCR>0 means semi-supervised NID [12]. In this work, we consider both unsupervised and semi-supervised scenarios. Figure 2 shows the overall framework of our model.

2.2. Pretraining

The multi-task learning strategy enables knowledge transfer from general intent detection tasks[13, 14]. Inspired by this, we adopt Multi-task Pre-training (MTP) approach in our model [12]. Following previous study [15], we use a pre-trained BERT encoder [16] with joint pre-training losses. The losses consist of a cross-entropy loss on external labeled data and a masked language modelling (MLM) loss on all the internal data:

$$\mathcal{L}_{pre1} = \mathcal{L}_{\text{ce}}\left(D_{external}^{labeled}; \theta\right) + \mathcal{L}_{\text{mlm}}\left(D_{internal}^{all}; \theta\right) \quad (1)$$

where θ are model parameters. We choose CLINC150 [17] as our external dataset for its diverse domains. For semi-supervised scenaio, we can further pre-train the model by replacing the external data to labeled indomain data :

$$\mathcal{L}_{pre2} = \mathcal{L}_{\text{ce}}\left(D_k^l; \theta\right) + \mathcal{L}_{\text{mlm}}\left(D_{internal}^{all}; \theta\right) \quad (2)$$

2.3. NCL

We propose a novel contrastive learning (NCL) model, which enhances the robustness of training from two stages. Firstly, we introduce diverse top-k paramenters to form initial positive pairs in data augmentation stage, then we devise a boundary distance constraint to judge positive and negative pairs in minibatch. Inspired by CLNN [12], we firstly encode each utterances x_i in the pre-training stage and search for its top-k nearest neighbors N_i in the embedding space. The utterances in N_i are supposed likely to have the same intent as x_i.

2.3.1. Picking Diverse Neighborhoods

As one top-k parameter would cause the lack of robustness, we introduce two diverse neighborhoods by using k_1 and k_2. We firstly construct an anchor neighborhood $N_i^{k_1}$ and a nearby neighborhood $N_i^{k_2}$ (k_1 is much smaller than k_2) for each x_i. Then, for each utterance x_i in minibatch B, we randomly choose x_i' from $N_i^{k_1}$ and x_i'' from $N_i^{k_2}$. After using data augmentation to generate $\tilde{x}_i'$ and $\tilde{x}_i''$ from x_i' and x_i'', we get two views of x_i, which forms a positive pair. Next, we construct an adjacency matrix A for the augmented batch $B' = \{\tilde{x}_i', \tilde{x}_i''\}_{i=1}^{M}$, which is a $2M \times 2M$ binary matrix where 1 indicates positive relation and 0 indicates negative relation. We show the loss function as follow [18]:

$$loss_i = -\frac{1}{|C_i|}\sum_{j \in C_i} log\frac{exp(sim(\tilde{h}_i, \tilde{h}_j)/\tau)}{\sum_{k \neq i}^{2M} exp(sim(\tilde{h}_i, \tilde{h}_k)/\tau)} \quad (3)$$

$$\mathcal{L} = \frac{1}{2M}\sum_{i=1}^{2M} l_i \quad (4)$$

where C_i indicates the sum of instances having positive relation with $\tilde{x}_i$, $\tilde{h}_i$ is the embedding of $\tilde{x}_i$, τ is the temperature parameter.

2.3.2. Boundary Distance Constraint

In previous work [12], researchers used the same parameter top-k to control both the range of choosing initial positive pair and the range of judging positive pair in adjacency matrix A, which we think should be considered separately. So, besides anchor neighborhood $N_i^{k_1}$ and nearby neiborhood $N_i^{k_2}$, we use a much larger k_3 to form a judging neighborhood $N_i^{k_3}$, combining with a distance boundary constraint to determine the binary value in matrix A. Hence, we can write the formula as:

$$a_{\chi_i \chi_j} = \begin{cases} 1, & d_{ij} < s \quad and \quad j \in N_i^{k_3} \\ 0, & \text{Otherwise} \end{cases} \quad (5)$$

where $\chi_i \in \{\tilde{x}_i', \tilde{x}_i''\}$, $\chi_j \in \{\tilde{x}_j', \tilde{x}_j''\}$. $a_{\chi_i \chi_j}$ is the element of adjacency matrix A, which represents the positive or negative relationship between χ_i and χ_j. So we can infer the relationship of $\tilde{x}_i'$ and $\tilde{x}_j'$ depending on the initial sample pair x_i and x_j. d_{ij} is the euclidean distance between h_i and h_j, s is the distance threshold, $N_i^{k_3}$ is the top-k_3 nearest neghborhood of x_i.

After the whole contrastive learning process, we use a non-parametric clustering algorithm (for simplicity, we use k-means) to obtain the final clustering results.

KCR	Methods	BANKING			StackOverflow			M-CID		
		NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
0%	GloVe-AG	52.76	14.41	31.18	23.45	4.85	24.48	51.23	32.57	42.35
	SAE-KM	60.12	24.00	37.38	48.72	23.36	37.16	51.03	43.51	52.95
	SAE-DEC	62.92	25.68	39.35	61.32	21.17	57.09	50.69	44.52	53.07
	BERT-KM	36.38	5.38	16.27	11.60	1.60	13.85	37.37	14.02	33.81
	MTP-CLNN	81.80	55.75	65.90	**78.71**	**67.63**	**81.43**	79.95	66.71	79.14
	NCL	**82.87**	**58.57**	**68.67**	77.24	66.85	80.50	**80.80**	**68.37**	**80.97**
25%	BERT-DTC	56.05	20.19	32.91	22.28	16.45	30.32	36.00	13.64	27.51
	CDAC+	67.65	34.88	48.79	74.33	39.44	74.30	43.89	19.65	39.37
	DAC	69.85	37.16	49.67	53.97	36.46	53.96	49.83	27.21	43.72
	MTP-DAC	81.48	55.64	66.12	77.22	61.42	78.60	77.79	62.88	77.02
	MTP-CLNN	84.11	61.29	71.43	79.68	70.17	83.77	80.24	66.77	79.20
	NCL	**85.40**	**65.28**	**75.47**	**82.09**	**75.03**	**87.30**	**80.49**	**67.70**	**80.80**
50%	BERT-DTC	69.68	35.98	48.87	53.94	36.79	51.78	51.90	28.94	44.70
	CDAC+	70.62	38.61	51.97	76.18	41.92	76.30	50.47	26.01	46.65
	DAC	76.41	47.28	59.32	70.78	56.44	73.76	63.27	43.52	57.19
	MTP-DAC	83.43	59.78	70.42	78.91	67.37	81.27	78.17	63.41	77.68
	MTP-CLNN	85.62	64.93	75.23	81.03	73.02	85.64	79.48	65.71	77.85
	NCL	**85.78**	**65.11**	**75.42**	**82.03**	**76.02**	**87.7**	**80.81**	**68.15**	**80.63**
75%	BERT-DTC	74.51	44.57	57.34	67.02	55.14	71.14	60.82	38.62	55.42
	CDAC+	71.76	40.68	53.46	76.68	43.97	75.34	55.06	32.52	53.70
	DAC	79.99	54.57	65.87	75.31	60.02	78.84	71.41	54.22	69.11
	MTP-DAC	85.78	65.28	75.43	80.89	71.17	84.20	80.94	68.27	80.89
	MTP-CLNN	87.52	70.00	79.74	82.56	75.66	87.63	83.75	73.22	84.36
	NCL	**87.84**	**71.24**	**81.07**	**84.87**	**79.11**	**89.4**	**84.85**	**75.53**	**86.82**

Table 1: *Overall performance on the three datasets. We use NMI, ARI and ACC to evaluate each model. The model is unsupervised when KCR=0, otherwise semi-supervised. The LAR is set to 10%.*

3. Experiments

3.1. Dataset

We evaluate NCL on three popular public datasets of NID. **BANKING** [19] contains more than 13000 customer messages with 77 intents. **STACKOVERFLOW** [20] is a large-scale questions dataset published online. **M-CID** [21] is a small corpus dataset collected for covid-19 study. In addition, we choose **CLINC150** [17] as the external dataset in our pre-training stage for its high quality annotations and the coverage of diverse domains. Table 2 illustrates the statistics of these datasets.

3.2. Training details

We use *bert-base-uncased* model [22] as the backbone of our model, while taking *[CLS]* token as the representation. For the pre-training stage, we follow the settings of previous works [12]. For the head of NCL, we use a two-layer fully connected network to convert the representation dimensionality from 768 to 128. In experiments, We set the batch size as 128, and τ as 0.1. The known class ratio (KCR) is $\{0\%, 25\%, 50\%, 75\%\}$, labeled ratio (LAR) is 10%[12]. For datasets BANKING, STACKOVERFLOW and M-CID, the anchor neighborhood size k_1 is set to be $\{3,3,3\}$, the nearby neighborhood size k_2 is set to be $\{40,300,20\}$, the judging neighborhood size k_3 is set to be $\{100,800,80\}$, and the distance threshold s is set to be $\{330,300,270\}$. The neighborhoods are updated every 5 epochs. We use Random Token Replacement (RTR) [12] as our augmentation method, and the replacement probability is set to be 0.25. For model optimazation, we use AdamW [22]. All the experiments are conducted on a single piece of Tesla P40 24GB.

3.3. Baselines

We compare NCL with several baselines: 1) **Glove-AG** [23] is based on Glove embeddings and evaluated with agglomerative clustering. 2) **SAE-KM** and **SAE-DEC** [24] are k-means and deep embedding clustering based on stacked autoencoder. 3) **BERT-KM** adopts k-means on BERT embeddings. 4) **Bert-DTC** [25] extends DEC into semi-supervised scenario. 5) **CDAC+** [4] uses a pseudo-labeling process. 6) **DAC** [5] puts forward a method of aligning clusters. 7) **MTP-DAC** and **MTP-CLNN** [12] adopt a multi-task pre-training strategy, CLNN uses an improved contrastive learning method for clustering.

3.4. Main results

Following [12], we use normalized mutual information (NMI), adjusted rand index (ARI), and accuracy (ACC) as the evaluation metrics. The overall results of all the models on three datasets are shown in Table 1, in which we can get some interesting conclusions. Firstly, all methods perform better under the semi-supervised (KCL = 25%, 50%, 75%) scenario compared to the unsupervised (KCL = 0%) scenario , which indicates labeled data can help the model to learn the granularity of clustering. Then, contrastive learning based methods outperform other approaches in our experiments, which demonstrates contrastive learning can benefit the representation learning of utterances. In addition, our proposed NCL performs a little bit better than MTP-CLNN in the unsupervised scenario, while it reaches the new state-of-the-art performance in the semi-supervised setting. Specifically, NCL achieves 1.11% improvement on average in NMI, 1.90% improvement on average in ARI and 2.20% im-

Datasets	Domain	Intents	Volume
CLINN150	general	120	18000
BANKING	banking	77	13083
STACKOVERFLOW	questions	20	20000
M-CID	covid-19	16	1745

Table 2: *The statistics of experimental datasets.*

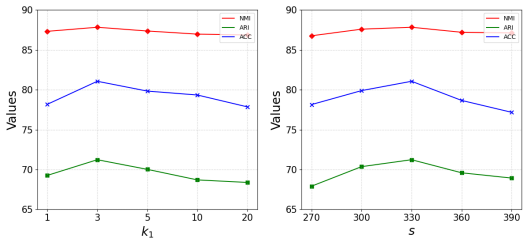

Figure 3: *Performances on different k_1 and s.*

provement on average in ACC over all semi-supervised settings.

4. Discussion

4.1. Hyper-parameter Tuning of k_1 & s

Here we study how k_1 & s influence the model performance in Figure 3. We set various k_1 on BANKING while LAR = 10% and KCR = 75%. With the growth of k_1, we can find MNI, ARI and ACC first increase and then decrease. Our model achieves the best performance when k_1 = 3, better than k_1 = 1 (*it is equal to the model without anchor neighborhood*), which indicates the proposed anchor neighborhood can help improve the model. We argue that the method of picking diverse top-k enhances the robustness indeed, while the k_1 value should not be too large, in order to maintain high similarity between x_i and x_i'. On the other hand, we experiment on various s, using the same settings as above. The results indicates experiments with too large s may not decrease noise adequately, while too small s leads to the generalization reduction. Moreover, our model achieves the best performances when s = 330, so this threshold is more suitable and closer to the real boundary distance for most intents in the dataset.

4.2. Ablation Study of Distance-based Constraint

To judge positive or negative pairs in the contrastive learning, we use a boundry distance s combined with a judging neighborhood $N_i^{k_3}$. To analyse the importance of this combination, we conduct a set of ablation experiments in Figure 4. The three sets of histograms respectively show the results of experiments with original settings, without s, and without k_3. Obviously, while maintaining both s & k_3, we obtain the best performance. Once we remove either of them, the results will get worse. The method without s performs better than the method without k_3, which indicates the influence of judging neighborhood $N_i^{k_3}$ is dominant compared with boundry distance s.

4.3. Visualization of Clustering

In Figure 5, we use Principal Component Analysis (PCA) to show the visualization of embeddings on STACKOVERFLOW by comparing strong baseline MTP-CLNN [12] and our NCL model. It shows our method makes the confused clusters more

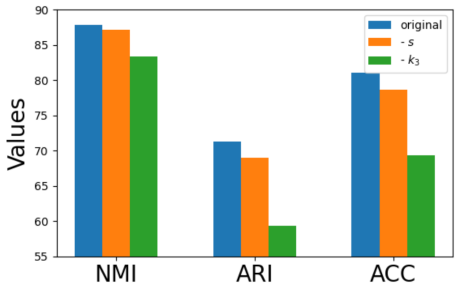

Figure 4: *Ablation study of distance-based constraint.*

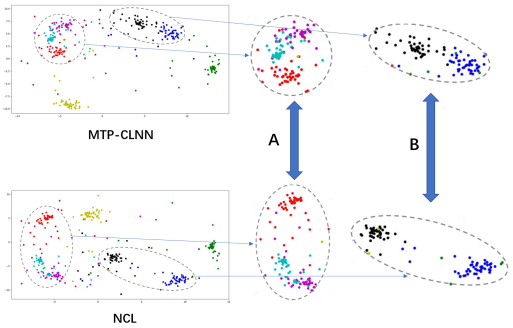

Figure 5: *Visualization results on STACKOVERFLOW.*

distinct with each other. Especially, we concentrate on comparison of two groups' clustering effect. In group A, we find NCL makes a better distinction between the red cluster and the other two clusters, while the three clusters are relatively close in MTP-CLNN. In group B, NCL reduces the confusion of black and blue clusters compared to MTP-CLNN. Results on other datasets also show similar effects.

5. Conclusions

In this paper, we propose a novel neighbor-based contrastive learning framework for NID task. We first introduce diverse top-k parameters into our novel contrastive learning, which enhances the robustness by using $N_i^{k_1}$ and $N_i^{k_2}$ as the selection range of initial positive pairs before data augmentation. Then we use a boundry distance threshold combined with $N_i^{k_3}$ range to determine the positive or negtive relationship between augmented data in every minibatch, which enhances the robustness by decreasing noise. Experimental results on three public datasets indicate our model outperforms all the baseline models on the IND task. In the future work, we will further improve our method from how to form initial positive pairs and judging positive or negative relationship between augmented data. For example, we may set dynamic parameters k_1 and k_2 for different samples, or we can propose more universal multi-metrics for judging the relationship of augmented data.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, U1636211, 61672081), the 2022 Tencent Big Travel Rhino-Bird Special Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2021ZX-18).

7. References

[1] D. Hakkani-Tür, Y. Ju, G. Zweig, and G. Tür, "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*, 2015, pp. 1854–1858.

[2] Padmasundari and S. Bangalore, "Intent discovery through unsupervised semantic text clustering," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, 2018, pp. 606–610.

[3] G. Forman, H. Nachlieli, and R. Keshet, "Clustering by intent: A semi-supervised method to discover relevant clusters incrementally," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III*, 2015, pp. 20–36.

[4] T. Lin, H. Xu, and H. Zhang, "Discovering new intents via constrained deep adaptive clustering with cluster refinement," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 2020, pp. 8360–8367.

[5] H. Zhang, H. Xu, T. Lin, and R. Lyu, "Discovering new intents with deep aligned clustering," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 2021, pp. 14 365–14 373.

[6] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2021, pp. 6894–6910.

[7] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 2021, pp. 5065–5075.

[8] J. M. Giorgi, O. Nitski, B. Wang, and G. D. Bader, "Declutr: Deep contrastive learning for unsupervised textual representations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 2021, pp. 879–895.

[9] Y. Mou, K. He, Y. Wu, Z. Zeng, H. Xu, H. Jiang, W. Wu, and W. Xu, "Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2022, pp. 46–53.

[10] Y. Mou, K. He, P. Wang, Y. Wu, J. Wang, W. Wu, and W. Xu, "Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for OOD intent discovery," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2022, pp. 1517–1529.

[11] D. Zhang, F. Nan, X. Wei, S. Li, H. Zhu, K. R. McKeown, R. Nallapati, A. O. Arnold, and B. Xiang, "Supporting clustering with contrastive learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 2021, pp. 5419–5430.

[12] Y. Zhang, H. Zhang, L. Zhan, X. Wu, and A. Y. S. Lam, "New intent discovery with pre-training and contrastive learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2022, pp. 256–269.

[13] J. Zhang, K. Hashimoto, W. Liu, C. Wu, Y. Wan, P. S. Yu, R. Socher, and C. Xiong, "Discriminative nearest neighbor few-shot intent detection by transferring natural language inference," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 2020, pp. 5064–5082.

[14] J. Zhang, T. Bui, S. Yoon, X. Chen, Z. Liu, C. Xia, Q. H. Tran, W. Chang, and P. S. Yu, "Few-shot intent detection via contrastive pre-training and fine-tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2021, pp. 1906–1912.

[15] H. Zhang, Y. Zhang, L. Zhan, J. Chen, G. Shi, X. Wu, and A. Y. S. Lam, "Effectiveness of pre-training for few-shot intent classification," in *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, 2021, pp. 1114–1120.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[17] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*.

[18] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *CoRR*, vol. abs/2004.11362, 2020.

[19] I. Casanueva, T. Temcinas, D. Gerz, M. Henderson, and I. Vulic, "Efficient intent detection with dual sentence encoders," *CoRR*, vol. abs/2003.04807, 2020.

[20] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, and H. Hao, "Short text clustering via convolutional neural networks," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, 2015, pp. 62–69.

[21] A. Arora, A. Shrivastava, M. Mohit, L. S.-M. Lecanda, and A. Aly, "Cross-lingual transfer learning for intent detection of covid-19 utterances," 2020.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *CoRR*, vol. abs/1910.03771, 2019.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*.

[24] J. Xie, R. B. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016, pp. 478–487.

[25] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019, pp. 8400–8408.