Multi-Stage Coarse-to-Fine Contrastive Learning for Conversation Intent Induction

Caiyuan Chu[1]*†, Ya Li[2]†‡, Yifan Liu[2], Jia-Chen Gu[4], Quan Liu[2,3], Yongxin Ge[1], Guoping Hu[2,3]

[1]Chongqing University, Chongqing, China
[2]iFLYTEK Research, Hefei, China [3]State Key Laboratory of Cognitive Intelligence
[4]University of Science and Technology of China, Hefei, China
{Chucy,yongxinge}@cqu.edu.cn, gujc@ustc.edu.cn,
{yali8,yfliu7,quanliu,gphu}@iflytek.com

Abstract

Intent recognition is critical for task-oriented dialogue systems. However, for emerging domains and new services, it is difficult to accurately identify the key intent of a conversation due to time-consuming data annotation and comparatively poor model transferability. Therefore, the automatic induction of dialogue intention is very important for intelligent dialogue systems. This paper presents our solution to Track 2 of Intent Induction from Conversations for Task-Oriented Dialogue at the Eleventh Dialogue System Technology Challenge (DSTC11). The essence of intention clustering lies in distinguishing the representation of different dialogue utterances. The key to automatic intention induction is that, for any given set of new data, the sentence representation obtained by the model can be well distinguished from different labels. Therefore, we propose a multi-stage coarse-to-fine contrastive learning model training scheme including unsupervised contrastive learning pre-training, supervised contrastive learning pre-training, and fine-tuning with joint contrastive learning and clustering to obtain a better dialogue utterance representation model for the clustering task. In the released DSTC11 Track 2 evaluation results, our proposed system ranked first on both of the two subtasks of this Track.

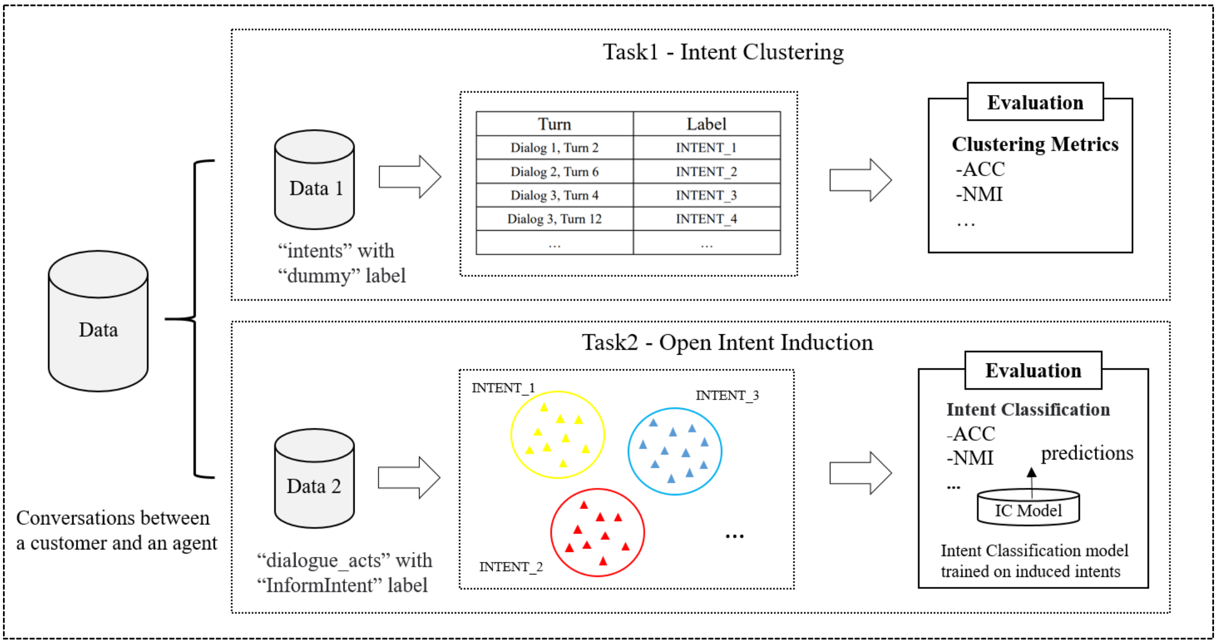
1 Introduction

The design of dialogue mode is very important for the development of task-oriented dialogue system. It typically consists of a set of intents with corresponding slots for capturing and handling domain-specific dialogue box states. Previous work on schema-guided dialogue (Rastogi et al., 2020a,b; Ruan et al., 2020; Lee et al., 2022) focused on data-efficient joint dialogue state modeling across domains and zero-shot generalization to new APIs. However, for new emerging domains and novel services, the identification of key intents of such schema typically requires domain expertise and/or laborious analysis of a large volume of conversation transcripts. As the demand for and adoption of virtual assistants continues to increase, recent work has investigated ways to accelerate pattern design through the automatic induction of intentions (Hakkani-Tür et al., 2015; Haponchyk et al., 2018; Perkins and Yang, 2019; Chatterjee and Sengupta, 2020) or the induction of slots and dialogue states (Min et al., 2020; Hudeček et al., 2021). However, the lack of realistic shared benchmarks with public datasets, metrics, and task definitions has made it difficult to track progress in this area. For this reason, the Eleventh Dialogue System Technology Challenge (DSTC11) proposed a track of Intent Induction from Conversations for Task-Oriented Dialogue. This Track is composed of two subtasks, which are organized as shown in Fig. 1. 1) Intent Clustering, which needs to cluster the given dialogue statements and evaluate them using standard clustering metrics. 2) Open Intent Induction, in which participants are required to generate a set of intents, each represented by a list of sample utterances. The induced intents and utterances will be evaluated using their performance on a downstream classification task over reference intents. Both subtasks aim to discover intents from conversations. This Track is very challenging due to the lack of labeled data and the unknown number of conversation intents.

This paper presents a system that is evaluated in this Track. For these tasks, in order to obtain a better dialogue utterance representation model under the condition that the data is unlabeled, we propose a multi-stage coarse-to-fine contrast learning model training scheme. The backbone of our multi-stage training scheme is the RoBERTa-large (Liu et al., 2019). Firstly, pre-training is performed

* Work done during the internship at iFLYTEK Research.
† Equal contribution.
‡ Corresponding author.

Proceedings of The Eleventh Dialog System Technology Challenge, pages 31–39
September 11, 2023 ©2023 Association for Computational Linguistics



Turn	Label
Dialog 1, Turn 2	INTENT_1
Dialog 2, Turn 6	INTENT_2
Dialog 3, Turn 4	INTENT_3
Dialog 3, Turn 12	INTENT_4
...	...

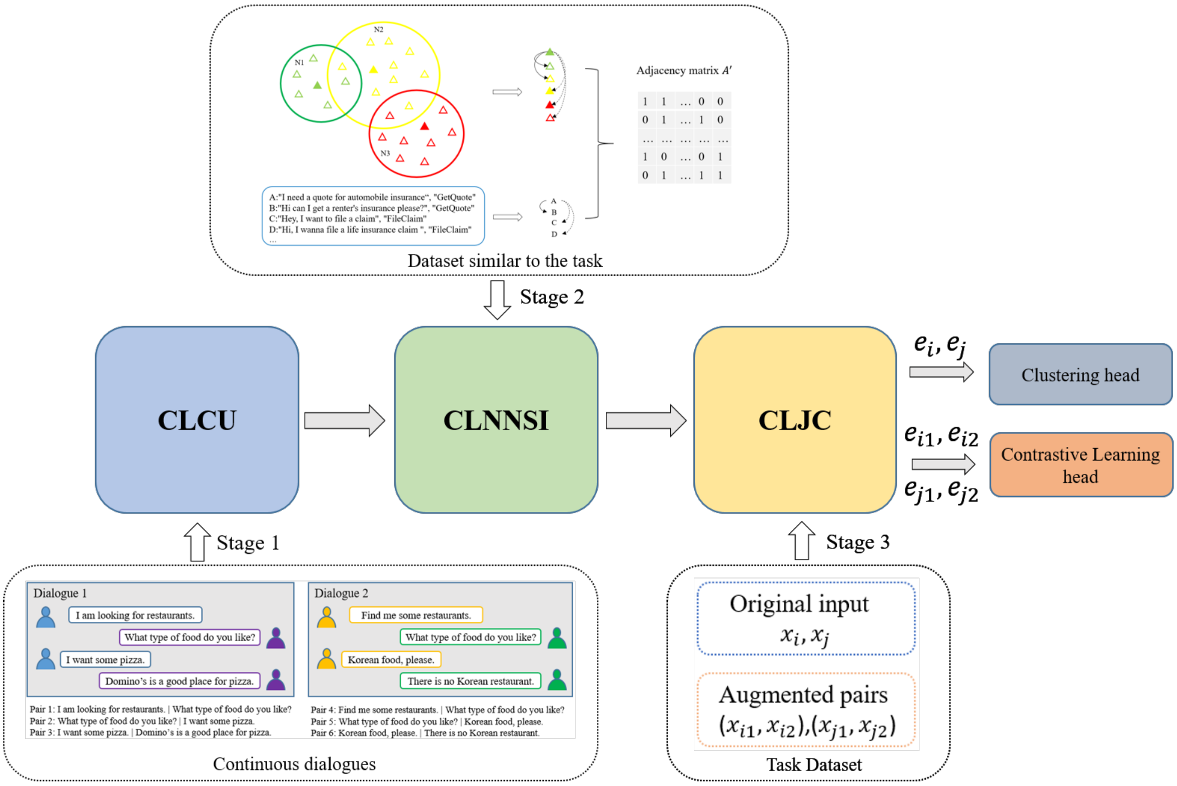
Figure 1: An overview of the tasks in this Track, including the tasks of Intent Clustering and Open Intent Induction.

by unsupervised contrastive learning in a large number of consecutive conversations. Secondly, the model is fine-tuned using a dataset with labels from the same domain as that of the target data for supervised comparative learning. Finally, the model obtained after the above two training steps is further fine-tuned on the target data by joint clustering and contrastive learning to obtain the final model. In addition, for the selection of the number of clustering categories k, we adopted the automatic parameter optimization method based on silhouette coefficients (Rousseeuw, 1987) provided in the baseline code. As shown in the released evaluation results, our proposed model ranked first on both subtasks. Furthermore, experimental results are analyzed by ablation tests. Finally, we draw conclusions and give an overview of our future work.

2 Related Work

Labeled data for task-oriented dialogue systems is often scarce because of the high cost of data annotation. Consequently, learning generic dialogue representations that effectively capture dialogue semantics at various granularities (Hou et al., 2020; Krone et al., 2020; Gu et al., 2019; Yu et al., 2021) lays a good foundation for handling a variety of downstream tasks (Vinyals et al., 2016; Snell et al., 2017). Contrastive learning has recently demonstrated promising results in the processing of natural language. Among which SimCSE (Gao et al., 2021) and TOD-BERT (Wu et al., 2020) get a very good performance on general texts and dialogues, respectively. DSE (Zhou et al., 2022) set new state-of-the-art results on general dialogues.

SimCSE (Gao et al., 2021) uses Dropout (Srivastava et al., 2014) to construct positive pairs by passing a sentence through the encoder twice to generate two different embeddings. Despite the fact that SimCSE performs better than ordinary data augmentation that directly manipulates discrete text, it has proven to be a poor performer in the field of dialogue, which is confirmed in the DSE (Zhou et al., 2022). Moreover, TOD-BERT takes an utterance and the concatenation of all the previous utterances in the dialogue as a positive pair. Although showing promise on the same tasks, TOD-BERT's semantic granularity and data statistics in many other dialogue tasks differ from those evaluated in their paper. DSE learns from dialogues by taking consecutive utterances of the same dialogue as positive pairs for contrastive learning, and state-of-the-art results are obtained in several tasks, including intention classification. Recently, MTP-CLNN (Zhang et al., 2022) set new state-of-the-art results in the New Intent Discovery field. However, the supervised pre-training part of it requires the data to have a small number of labels to get better results. According to our experimental verification, the generalization ability of the pre-trained model in the first stage is poor for data with no labels at all. In addition, recent research results



1	1	…	0	0
0	1	…	1	0
…	…	…	…	…
1	0	…	0	1
0	1	…	1	1

A:"I need a quote for automobile insurance“, "GetQuote"
B:"Hi can I get a renter's insurance please?", "GetQuote"
C:"Hey, I want to file a claim", "FileClaim"
D:"Hi, I wanna file a life insurance claim ", "FileClaim"
…

I am looking for restaurants.
What type of food do you like?
I want some pizza.
Domino’s is a good place for pizza.

Find me some restaurants.
What type of food do you like?
Korean food, please.
There is no Korean restaurant.

Pair 1: I am looking for restaurants. | What type of food do you like?
Pair 2: What type of food do you like? | I want some pizza.
Pair 3: I want some pizza. | Domino’s is a good place for pizza.
Pair 4: Find me some restaurants. | What type of food do you like?
Pair 5: What type of food do you like? | Korean food, please.
Pair 6: Korean food, please. | There is no Korean restaurant.

Figure 2: The framework of our proposed multi-stage coarse-to-fine contrastive learning model training scheme.

on short text clustering show that the combined training method based on clustering and contrastive learning, SCCL (Zhang et al., 2021) achieves very good results. However, a better pre-trained model for the clustering task as encoded by SCCL can further improve the results of clustering. To this end, the goal of our work in the first two stages is to continue pre-training language models to derive better representations for downstream tasks.

3 Methodology

Our proposed multi-stage coarse-to-fine model training scheme consists of three stages: Contrastive Learning with Consecutive Utterances (CLCU), Contrastive Learning with the Nearest Neighbors and the Same Intent (CLNNSI) and Contrastive Learning with Joint Clustering (CLJC). This is shown in Fig. 2. In the first stage, a pre-trained model is obtained by performing unsupervised contrastive learning on a large amount of dialogue data using consecutive discourses of the same dialogue as positive pairs. In the second stage, for labeled data in the same domain, we treat that sample with its neighboring samples or samples with the same intention as a positive pairs, and then fine-tune the model by contrastive learning. The third stage further fine-tunes the model by joint contrastive learning and clustering jointly on the target data. The negative pairs for contrastive learning are collected by small batches of negative sampling in the above three stages of model training. After training, we employ a simple non-parametric clustering algorithm named k-means to obtain clustering results. In this section, we describe our multi-stage coarse-to-fine contrastive learning model training scheme in detail below.

3.1 Stage 1: CLCU

CLCU encourages the model to treat an utterance as similar to its neighboring utterances and dissimilar to utterances that are not subsequent to it or that belong to other dialogues when doing contrastive learning on consecutive utterances. Consecutive utterances contain implicit categorical information, which benefits dialogue classification tasks (e.g., intent classification). Consider pairs 1 and 4 in Fig. 2 stage 1: We implicitly learn similar representations of *I am looking for restaurants* and *Find me some restaurants*, since they are both consecutive with *What type of food do you like?*. In contrast, SimCSE does not enjoy these benefits by simply using Dropout as data augmentation. Although

TOD-BERT also leverages the intrinsic semantics of dialogue by combining an utterance with its dialogue context as a positive pair, the context is often a concatenation of 5 to 15 utterances. Due to the large discrepancy in both semantics and data statistics between each utterance and its context, simply optimizing the similarity between them leads to less satisfying representations on many dialogue tasks. Just like the experimental results in DSE (Zhou et al., 2022). TOD-BERT can even lead to degenerated representations on some downstream tasks when compared to the original BERT (Devlin et al., 2019) model. Therefore, in the first stage, we adopt the method of DSE, which learns from dialogues by taking consecutive utterances of the same dialogue as positive pairs for contrastive learning, and directly use the model they trained on a large number of datasets as our first stage model.

3.2 Stage 2: CLNNSI

In the second stage, a small number of labeled datasets in the same domain as the target data are used, which makes our model have stronger generalization ability. And adopt an objective that pulls neighboring instances together and pushes distant ones away in the embedding space to learn compact representations for clustering. To be specific, we first encode the utterances with the pre-trained model from stage 1. The inner product is then used as a distance metric to find the top-K nearest neighbors of each utterance x_i in the embedding space, forming a neighborhood N_i. During training, for each minibatch of utterances B $= \{x_i\}_{i=1}^{M}$ and each utterance $x_i \in$ B, we uniformly sample one neighbor x_i' from its neighborhood N_i. Then use data augmentation to generate $\tilde{x}_i$ and $\tilde{x}_i'$ for x_i and x_i' respectively. Here, $\tilde{x}_i$ and $\tilde{x}_i'$ are treated as a positive pair. We then obtain an augmented batch $B' = \{\tilde{x}_i, \tilde{x}_i'\}_{i=1}^{M}$ with all the generated samples. To compute contrastive loss, we construct an adjacency matrix A' for B', which is a 2M $\times$ 2M binary matrix where 1 indicates positive relations (either being neighbors or having the same intent label) and 0 indicates negative relations. Hence, the contrastive loss can writed as:

$$l_i = -\frac{1}{|C_i|} \sum_{j \in C_i} \log \frac{\exp\left(\operatorname{sim}\left(\tilde{e}_i, \tilde{e}_j\right)/\tau\right)}{\sum_{k \neq i}^{2M} \exp\left(\operatorname{sim}\left(\tilde{e}_i, \tilde{e}_k\right)/\tau\right)}, \tag{1}$$

$$L_{stg2} = \frac{1}{2M} \sum_{i=1}^{2M} l_i, \tag{2}$$

where $C_i = \left\{A'_{ij} = 1 \mid j \in \{1, \ldots, 2M\}\right\}$ denotes the set of instances having positive relation with $\tilde{x}_i$ and $|C_i|$ is the cardinality. $\tilde{e}_i$ is the embedding for utterance $\tilde{x}_i$, τ is the temperature parameter. $sim(.,.)$ is a similarity function on a pair of normalized feature vectors. Has the following advantages by introducing the notion of neighborhood relationships in contrastive learning: 1) Similar instances are pulled together and dissimilar instances are pushed away to achieve more compact clusters; and 2) known intents are naturally incorporated with the adjacency matrix.

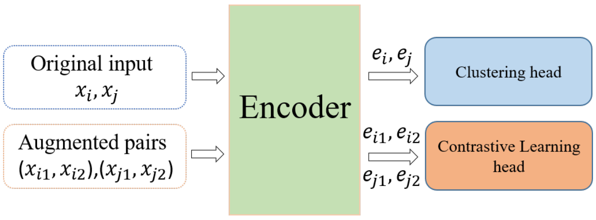


Figure 3: The training framework for stage 3.

3.3 Stage 3: CLJC

Previous research efforts focused on integrating clustering with deep representation learning by optimizing a clustering objective defined in the representation space (Zhang et al., 2017; Shaham et al., 2018). Despite promising improvements, the clustering performance is still inadequate, especially in the presence of complex data with a large number of clusters. One possible reason is that, even with a deep neural network, data still has significant overlap across categories before clustering starts. Consequently, the clusters learned by optimizing various distance or similarity-based clustering objectives suffer from poor purity. Moreover, contrastive learning has recently achieved remarkable success in self-monitoring (Wu et al., 2018; Chen et al., 2020), as the name suggests, a contrastive loss is adopted to pull together samples augmented from the same instance in the original dataset while pushing apart those from different ones. This beneficial property can be leveraged to support clustering by scattering apart the overlapped categories. Hence, in the third stage, for target data, we adopt the method of joint clustering and contrastive learning to further finetune the model. The training framework stage3 is shown in Fig. 3. Among them, x_{i_1},x_{i_2} and x_{j_1},x_{j_2} are obtained by means of data augmentation. In this paper, we adopt the contextual augmenter data

dataset	dial.	test-utt.	task1-utt.	task2-utt.
dev	948	913	1205	4332
banking	1000	407	1503	3696
finance	2000	1130	1597	6676

Table 1: Statistics of the DSTC11-Track 2 datasets. utt.: utterance, dial.: dialogue.

augment in the form of a pre-trained transformers to find the top-n suitable words of the input text for insertion or substitution. We used word substitution to augment the data and chose Bertbase and Robertabase to generate augmented pairs. And the overall objective is:

$$L_{stg3} = L_{CL} + \eta L_{Clu}, \tag{3}$$

where L_{CL} and L_{Clu} are the loss functions of comparative learning and clustering, respectively. η balances between the contrastive loss and the clustering loss of stage3, for simplicity, it is set to 10 in our experiment.

4 Experiments

4.1 Dataset

For this Track, one development dataset and two test datasets are provided. Each dataset consists of 1) some human-to-human conversations between a customer and an agent, and 2) a set of testing samples with corresponding intent annotations. The task1-utterances and task2-utterances are utterances in which "intents" are non-empty and "dialogue_acts" are "InformIntent" in dialogues in each dataset, respectively. Detailed statistics of the dataset are summarized in Table 1.

4.2 Metrics

Task 1 and task 2 are both evaluated by the following six metrics,: accuracy (ACC) (Huang et al., 2014), normalized mutual information (NMI), F1-score, Recall, Precision, and adjusted rand index (ARI). But the ACC is the primary metric used for ranking system submissions. Metrics dependent on reference intents will be computed using an automatic alignment of cluster labels to reference intent labels. For task 1, alignments will be computed based on turn-level reference intent labels. For task 2, to avoid the need to assign labels to turns in the input transcripts, alignments will be computed using classifier predictions on the set of utterances held out for evaluation. In both cases, 1:1 alignments between induced intents and reference intents will be computed using the Hungarian algorithm (Kuhn, 1955).

Team	ACC	P	R	F1	NMI	ARI
T23	**69.79**	76.09	76.12	76.00	75.05	59.23
T07	69.59	72.01	81.64	**76.50**	73.48	60.13
T35	69.31	**78.41**	73.35	75.78	74.30	58.67
T05	69.06	73.26	78.32	75.70	**75.54**	**61.38**
T02	68.83	73.04	78.06	75.46	75.13	60.88
T17	67.15	70.49	78.48	74.10	73.20	60.86
T36	66.32	71.86	74.46	73.13	73.64	58.05
T24	66.19	71.17	77.36	74.13	74.72	58.15
T00	64.92	71.42	74.80	73.05	71.08	50.37
T34	63.73	71.01	74.84	72.59	72.98	52.77
...	...	...	...	...	...	...
baseline	55.80	64.97	62.98	63.26	62.98	39.85

Table 2: The summary of the task 1 evaluation results on the two datasets, with the best score in bold.

Team	ACC	P	R	F1	NMI	ARI
T23	**76.30**	78.68	**89.86**	83.55	87.42	**71.82**
T02	75.34	78.18	88.18	82.86	87.32	68.87
T36	74.85	78.81	87.59	82.85	87.00	71.36
T24	74.70	79.76	87.62	83.42	87.45	70.71
T05	74.52	79.50	87.49	83.17	87.88	70.26
T17	73.79	**83.25**	84.87	**83.99**	**88.11**	71.42
T14	69.55	70.68	87.97	78.27	83.67	65.11
T13	69.43	82.66	74.47	78.26	82.29	62.13
T27	68.70	80.29	77.51	78.76	83.69	63.83
T19	67.50	69.09	89.4	77.92	83.73	63.67
...	...	...	...	...	...	...
baseline	63.61	68.93	79.59	73.86	80.09	57.24

Table 3: The summary of the task 2 evaluation results on the two datasets, with the best score in bold.

4.3 Experiment Results

The final summary results of the test-banking dataset and the test-finance dataset for the two subtasks are shown in Tables 2 and 3. Our model ranks first in the test sets for both subtasks. Compared with the baseline, where the baseline model in the above tables is the official provided baseline model mpnet by Track 2 of DSTC11. It is trained on a large and diverse dataset of over 1 billion training pairs. And the test ACC of our model is improved by 13.99% and 12.69% on tasks 1 and 2, respectively. It shows the effectiveness of

	task1	task2
GloVe (Pennington et al., 2014)	20.58	29.12
MPNet (Song et al., 2020)	46.14	60.23
SimCSE (Gao et al., 2021)	47.39	57.74
DSE (Zhou et al., 2022)	58.67	64.18
SCCL (Zhang et al., 2021)	65.32	78.34
Our model	**75.68**	**85.76**

Table 4: Comparison of the accuracy of different models on the development dataset for the two tasks.

our method.

In addition, we give the results of our method on the development dataset compared with some representative baseline models. We compared with the following baseline models: **(1) GloVe** (Pennington et al., 2014). One of the official baseline models provided by the Track 2 of DSTC11, which is a sentence-transformers model. It maps sentences or paragraphs to a 300 dimensional dense vector space and can be used for tasks like clustering or semantic search. **(2) MPNet** (Song et al., 2020). Another official baseline model provided by Track 2 of DSTC11, which is trained on a large and diverse dataset of over 1 billion training pairs. The base model is MPNet-base. **(3) SimCSE** (Gao et al., 2021). The model we chose is trained on 10^6 randomly sampled sentences from English Wikipedia by unsupervised contrastive learning, and its base model is RoBERTa-large. **(4) DSE** (Zhou et al., 2022). It is trained on a large number of training pairs constructed in the style of consecutive conversational sentences as positive pair using multiple conversational datasets. And the base model is RoBERTa-large. **(5) SCCL** (Zhang et al., 2021). Fine-tuning of the baseline model using the development dataset by joint training with clustering and contrastive learning, where the baseline model is the DSE-trained model described above.

As can be seen from the experimental results in Table 4, our model obtained the best results in both subtasks compared to several other representative baseline models, with accuracy rates of 75.68% and 85.76%, respectively. On task 1, the accuracy of our model outperformed GloVe by 55.1%, outperformed MPNet by 29.54%, outperformed SimCSE by 28.29%, outperformed DSE by 17.01%, and outperformed SCCL by 10.36%. On task 2, the accuracy of our model outperformed GloVe by 56.64%, outperformed MPNet by 25.53%, outperformed SimCSE by 28.02%, outperformed DSE by 21.56%, and outperformed SCCL by 7.42%. This illustrates the soundness of our approach.

	task1	task2
model	75.68	85.76
w/o. stage1,stage2 and stage3	46.14	60.24
w/o. stage2 and stage3	58.59	60.90
w/o. stage3	70.71	80.28

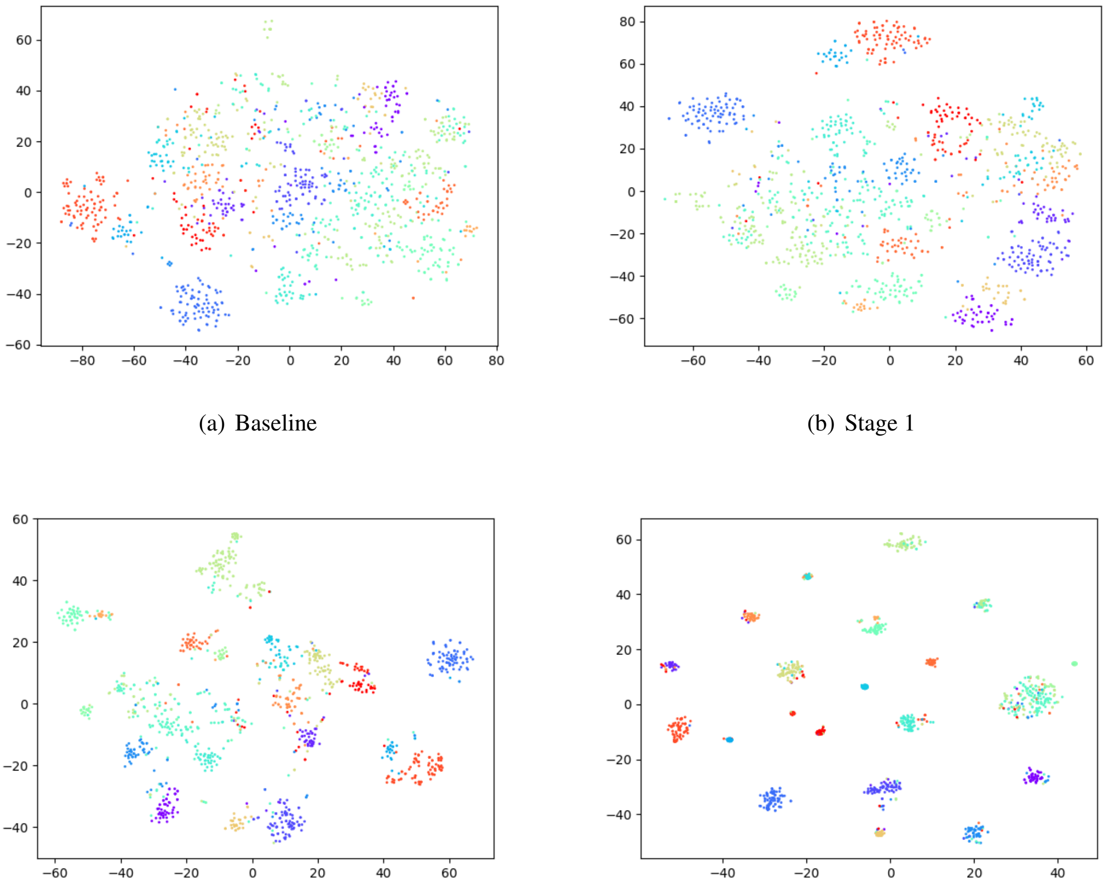
Table 5: Accuracy of ablation experiment on the development set.

4.4 Ablation Study

Through ablation experiments with a random seed set to 42 in the clustering algorithm on the development dataset, the experimental results are shown in Table 5. When there is no stage1, stage2 and stag3 (this is the case is the baseline), the ACC of the test on task1 and task2 are 46.14%, 60.24% respectively. when stage1, stage2 and stage3 are performed, the results on task1 are improved by 12.45%, 12.12% and the results on task2 improved by 0.68%, 19.38%, and 5.48%, respectively. In addition, the TSNE visualization of the utterances representation of the model obtained at each stage on the development set is shown in Fig. 4. It can be observed from the figure that each cluster becomes more and more compact after each stage compared to the baseline model. That verifies the rationality of each stage.

5 Conclusion

In this paper, we present our solution to the challenge of Intent Induction from Conversations for task-Oriented Dialogue of DSTC11. Firstly, we chose RoBERTa-large, which was pre-trained on a large number of continuous dialogues. Since continuous utterances also contain implicit classification information, they are beneficial for the task of dialogue intention classification. Secondly, the use of some labeled datasets in the same domain as the target data makes our model have stronger transfer ability. And we adopt KNN to select the positive pair for contrastive learning in order to pull neighboring instances together and push distant ones away in the embedding space to learn compact representations for clustering. Lastly, the joint training method of clustering



(a) Baseline

(b) Stage 1

(c) Stage 2

(d) Stage 3

Figure 4: TSNE visualization (Van der Maaten and Hinton, 2008) of the dialogue representations provides by baseline and three stages on the development set, each color indicates a ground truth semantic category.

and contrastive learning makes the advantages of clustering and contrastive learning complementary. Experimental results demonstrate that our methods can effectively cluster the utterances with intention in the dialogue. Our method of competitive performance achieved first place in two subtasks. In the future, we will explore better ways to obtain a better dialogue utterance representation model for the clustering task.

Acknowledgements

We thank anonymous reviewers for their valuable comments.

References

Ajay Chatterjee and Shubhashis Sengupta. 2020. Intent mining from past conversations for conversational agent. *CoRR*, abs/2005.11014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. 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*, pages 6894–6910. Association for Computational Linguistics.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*, pages 2321–2324. ACM.

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

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.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393.

Peihao Huang, Yan Huang, Wei Wang, and Liang Wang. 2014. Deep embedding network for clustering. In *2014 22nd International conference on pattern recognition*, pages 1532–1537. IEEE.

Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. 2021. Discovering dialogue slots with weak supervision. 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 2430–2442.

Jason Krone, Yi Zhang, and Mona T. Diab. 2020. Learning to classify intents and slot labels given a handful of examples. *CoRR*, abs/2004.10793.

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

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. Sgd-x: A benchmark for robust generalization in schema-guided dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10938–10946.

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. *CoRR*, abs/1907.11692.

Qingkai Min, Libo Qin, Zhiyang Teng, Xiao Liu, and Yue Zhang. 2020. Dialogue state induction using neural latent variable models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3845–3852. ijcai.org.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. 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*, pages 1532–1543. ACL.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. 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*, pages 4014–4023. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020a. Schema-guided dialogue state tracking task at DSTC8. *CoRR*, abs/2002.01359.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020b. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Yu-Ping Ruan, Zhen-Hua Ling, Jia-Chen Gu, and Quan Liu. 2020. Fine-tuning BERT for schema-guided zero-shot dialogue state tracking. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, Workshop on the Eighth Dialog System Technology Challenge, DSTC8*.

Uri Shaham, Kelly P. Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. 2018. Spectralnet: Spectral clustering using deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Chien-Sheng Wu, Steven C. H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 917–929. Association for Computational Linguistics.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. Few-shot intent classification and slot filling with retrieved examples. 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*, pages 734–749. Association for Computational Linguistics.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen R. McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. 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*, pages 5419–5430. Association for Computational Linguistics.

Dejiao Zhang, Yifan Sun, Brian Eriksson, and Laura Balzano. 2017. Deep unsupervised clustering using mixture of autoencoders. *arXiv preprint arXiv:1712.07788*.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2022. 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*, pages 256–269. Association for Computational Linguistics.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew O. Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 754–768. Association for Computational Linguistics.