DORIC : Domain Robust Fine-Tuning for Open Intent Clustering through Dependency Parsing

Jihyun Lee[1], **Seungyeon Seo**[1], **Yunsu Kim**[1,2], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
{jihyunlee, ssy319, yunsu.kim, gblee}@postech.ac.kr

Abstract

We present our work on Track 2 in the Dialog System Technology Challenges 11 (DSTC11). DSTC11-Track2 aims to provide a benchmark for zero-shot, cross-domain, intent-set induction. In the absence of in-domain training dataset, robust utterance representation that can be used across domains is necessary to induce users' intentions. To achieve this, we leveraged a multi-domain dialogue dataset to fine-tune the language model and proposed extracting Verb-Object pairs to remove the artifacts of unnecessary information. Furthermore, we devised the method that generates each cluster's name for the explainability of clustered results. Our approach achieved 3rd place in the precision score and showed superior accuracy and normalized mutual information (NMI) score than the baseline model on various domain datasets.

1 Introduction

Understanding the user's intent plays an important role in task-oriented dialogue (TOD) systems. Traditionally, understanding the user's intent requires supervised training using intent-annotated dialogue datasets (Xu et al., 2013; Wang et al., 2015; Kim et al., 2017; Goo et al., 2018). However, for new emerging domains and services, defining the intent set is challenging and also requires an expert's knowledge. Therefore, finding an automatic method that identifies intents from raw conversational data is desirable to reduce costs.

The intent clustering task of the 11th Dialog System Technology Challenge (DSTC11) aims to provide a realistic benchmark for the intent induction problem. This track evaluates automatic customer intent induction methods from dialogues between human agents and customers, and the DSTC11 challenge participants are required to create a set of intent labels based on the conversations. To provide a realistic setting, the number of intents and domain of the test set are not provided until the development phase ends.

In this paper, we introduce an automatic intent induction framework that effectively utilizes a public TOD dataset. First, we fine-tuned the language model with multi-domain TOD datasets so that it has a domain-robust semantic representation. Here, we extract verbs and object from utterance to remove the artifacts of unnecessary information. For the training, we applied supervised contrastive learning (SCL), which is known to be stable in language model fine-tuning (Gunel et al., 2020). Second, to infer a new intent set from the unseen domain dataset, we applied a clustering technique that groups the utterances based on the fine-tuned embedding model's representation. Furthermore, we generate a label name for each cluster to obtain an interpretable result. The cluster label generation method could reduce the effort of examining each set manually to understand the clustering results.

In the experiment with the test dataset (finance and banking domain), we achieved 3rd place in terms of precision and demonstrated superior accuracy (ACC) and a higher normalized mutual information (NMI) score than the baseline. Furthermore, the generated clustering labels reasonably explain each cluster. Finally, we analyzed our model with comparable options and demonstrated the result on development, test, and TOD datasets. We named our framework DORIC, which means **DO**main **R**obust fine-tuning for open **I**ntent **C**lustering through dependency parsing.

2 Related Work

2.1 Intent Classification

Traditionally, benchmark datasets (Price, 1990; Coucke et al., 2018; Eric et al., 2019) for intent identification have sufficient labeled datasets for training, and the task has been solved through the classification method. For example, Goo et al. (2018) classified the intent and slot information using the attention mechanism, and Kim et al. (2017)

Proceedings of The Eleventh Dialog System Technology Challenge, pages 40–47
September 11, 2023 ©2023 Association for Computational Linguistics

enriched word embeddings by using semantic lexicons and adapted this strategy to the intent classification. In addition, Wang et al. (2015) grouped the intent of tweets into six categories, used a graph embedding consisting of tweet nodes, and classified their intents. However, labeling the intent for the raw dialogue dataset requires extensive human labor, so building a new labeled intent dataset in the real world is challenging. Therefore, a robust intent induction model that can be applied to a new domain as an unsupervised method is required.

2.2 Intent induction with unsupervised method

The representative method of unsupervised intent induction utilizes the clustering method. Liu et al. (2021) is one example of research that enhanced the clustering algorithm. They proposed a balanced score metric to obtain similar-sized clusters in K-means clustering and found proper K-values that were more stable than naive K-means. Chatterjee and Sengupta (2020) also enhanced the clustering algorithm by utilizing the outlier information of the density-based clustering model, which is called ITER-DBSCAN. Their work shows greater accuracy on imbalanced intent data. On the other hand, there has been research that improved the dialogue representations for better clustering results. For example, Perkins and Yang (2019) iteratively enhanced the dialogue embedding by reflecting the clustering score, and Lin and Xu (2019) proposed a BiLSTM (Mesnil et al., 2014) embedding model with margin loss that is effective in detecting unknown intents. However, robust intent induction in diverse domains was not examined in previous research. Therefore, as a strategy for enhancing the embedding dialogue model, we propose the DORIC method, which robustly embeds diverse dialogue domains.

3 DSTC11 Intent Clustering Task

In this task, participants are required to assign an intent label to each dialogue turn. A set of dialogues are provided as input, and each turn is pre-labeled with both its speaker role (i.e., Agent or Customer) and dialogue acts (i.e., InformIntent or not). One development dataset and two test datasets are provided, and each dataset consists of approximately 1K customer support spoken conversations with manual transcriptions and annotations. The development dataset derives from an insurance-related customer support service, and each conversation has an average of 70 turns. In addition, the development dataset contains ground truth intent annotations that allow participants to test and evaluate the model. The number of intent types and the domains of the test dataset are not revealed until the development phase ends. Note that no training dataset is given, as this challenge aims to zero-shot intent induction.

4 Method

4.1 Semantic representation

With the advance of the pre-trained langue model, leveraging these models for embedding dialogue has exhibited promising results (Ham et al., 2020; Lin et al., 2020; Yang et al., 2021; Lee and Lee, 2022). Following their success, we utilize the pre-trained SBERT (Reimers and Gurevych, 2019) as a backbone model; SBERT has a siamese network structure and performs well in classification (Reimers and Gurevych, 2019), summarization (Zhong et al., 2020), and intent clustering (Liu et al., 2021).

The SBERT model is pre-trained on a written form text dataset that has a different linguistic pattern with the dialogue utterances. This difference could hurt the accuracy on dialogue related tasks (Wu et al., 2020). Therefore, we fine-tuned the model with the multi-domain public task-oriented dialogue dataset MultiWOZ 2.2 (Eric et al., 2019). This dataset has nine intents types, and we fine-tuned the model to learn the embedding of utterances according to intent type. In this data, intents are spanned multiple turns, and some utterances contain multiple intents in one utterance. To clarify the match between the utterance and the intent label, we used only the first utterance of the spanned intent dialogue and excluded utterances that contained multiple intents. We analyze the processed fine-tuning dataset in Table 1.

As we aim for unsupervised intent induction, domain robust fine-tuning is crucial to identify the intents across the domain. To do so, we extract Verb-Object structure from the utterance using the dependency parser[1]. This additional pre-process removes the effect of non-relevant words or utterance styles when fine-tuning the SBERT model. However, at inference time, we used whole utterances for clustering as preliminary experiments demon-

[1]https://spacy.io/

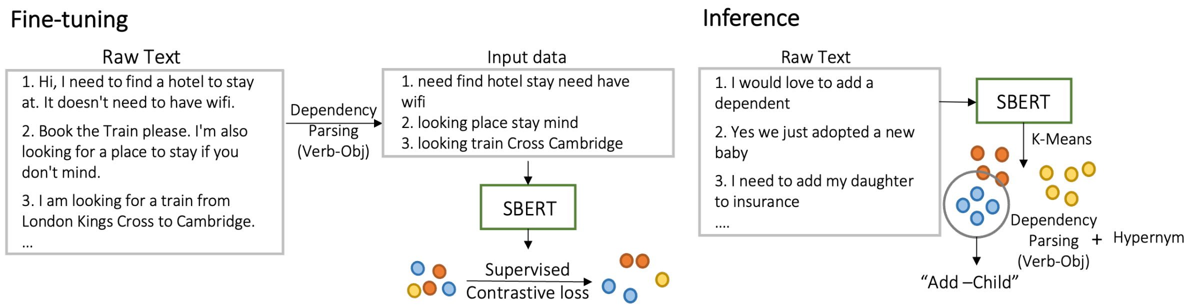

Figure 1: Fine-tuning and inference method of our proposed methods.

Intent	Count
FindRestaurants	3561
SearchHotel	3375
FindTrains	3262
FindAttractions	2795
ReserveHotel	1951
GetTrainTickets	1926
ReserveRestaurant	1600
GetRide	1262
FindPolice	229
Total	19961

Table 1: Type and number of intents of the fine-tuning dataset.

Inference data	NMI	ACC
Verb-Object	41.89	30.79
Sentence	**65.16**	**56.68**

Table 2: The comparison of using whole sentence and Verb-Object format in inference. NMI and accuracy result on DSTC11 development are reported.

strated better results (Table 2). We demonstrate our overall method in Figure 1.

4.2 Supervised contrastive learning

Recently, there have been several successful studies using contrastive learning (CL) in the computer vision and language domain (Chen et al., 2020; Liu and Abbeel, 2020; Wu et al., 2018; Gunel et al., 2020) and CL shows more generalize and robustness to language model training than cross-entropy loss (Gunel et al., 2020). Following their success, we utilize supervised contrastive learning (SCL) (Khosla et al., 2020) in fine-tuning.

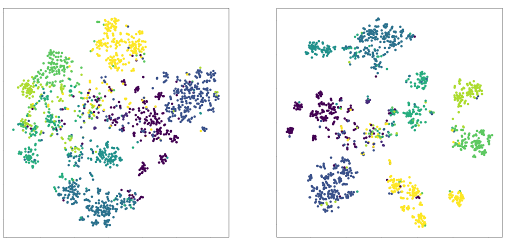

Figure 2: Visualization of Multiwoz 2.2 test dataset. The left is utterance representation before training, and the right is after training.

SCL is a modified version of the CL approach, which utilize the label information. In CL, only the anchor and its augmented object are regarded as positive objects, and others are used as negative object in training. However, SCL set the same label objects as positive and others as negative, so more accurate embedding representation learning is possible.

For the N randomly sampled object $\{x_k, y_k\}_{k=1...N}$, mini batch used for training is consists of $2N$ pair,$\{\tilde{x}_k, \tilde{y}_k\}_{k=1...2N}$. In $\{\tilde{x}_k, \tilde{y}_k\}_{k=1...2N}$, $\{\tilde{x}_k, \tilde{y}_k\}_{k=1...N}$ is same as original sampled $\{x_k, y_k\}_{k=1...N}$ and $\{\tilde{x}_k, \tilde{y}_k\}_{k=N...2N}$ is augmented pair of original sampled data. Wordnet (Miller, 1995) based synonym augmentation[2] is used for augmentation. Φ denotes the SBERT encoder and τ is scalar temperature parameter to adjust the separation strength. The overall loss is given by the following equations, and visualization of the training results are shown in Figure 2:

$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \mathcal{L}_i^{sup} \tag{1}$$

$$L_i^{sup} = -\frac{1}{2N_{\tilde{y}} - 1} \sum_{j=1}^{2N} \mathbf{1}_{i \neq j} \cdot \mathbf{1}_{\tilde{y} = \tilde{y}} \cdot \log \frac{\exp(\Phi_i \cdot \Phi_j / \tau)}{\sum_{k=1}^{2N} \mathbf{1}_{i \neq k} \exp(\Phi_i \cdot \Phi_j / \tau)} \tag{2}$$

4.3 Intent label generation

To improve the explainability of the clustering results, we automatically generated the semantic labels from the clustering results. Following the intent datasets, which usually represent the intent

[2]https://github.com/makcedward/nlpaug

name as a verb and object pair (Zang et al., 2020; Coucke et al., 2018; Rastogi et al., 2020), we also named our induced clusters as Verb-Object forms using a dependency parser. In the previous method, Liu et al. (2021) counted the common Verb-Object pairs in the cluster and used the most common pair as the cluster name. However, this method did not create a proper label when detailed words appeared in the object. For example, a pair of `call-son` and `call-daughter` cannot be grouped as `call-child` using the previous method.

To overcome this limitation, we propose a method that uses the object word's hypernyms. By adopting hypernyms, we could obtain a proper word containing detailed information. More precisely, we generate *verb-hyper(object)* and *verb-hyper(hyper(object))* pairs from existing Verb-Object pairs and calculate the most common pair from this generated result. We employ this rule when the number of the most common pair and second place does not differ by more than α times, and we set α to 2.0 in the experiment. Word-net (Miller, 1995) is used to get the hypernyms.

5 Experiment

5.1 Experimental setup

Dataset To demonstrate the performance of our model, we used the development (dev) (insurance), test1 (banking), and test2 (finance) datasets. These datasets have domains that are different from the fine-tuning dataset (tourism), so we were able to examine our method's effectiveness in diverse domains. Additionally, we used the Schema-Guided Dialogue Dataset (SGD) dataset; we extracted tourism-related domains from the SGD dataset to make the same domain environment with fine-tuning dataset. The number of intents for each dataset is shown in Table 3.

Dataset	Domain	# of intents
Dev	Insurance	22
Test1	Banking	29
Test2	Fianace	39
SGD	Tourism	6

Table 3: Domain and number of intents type for each dataset.

Evaluation NMI and accuracy were the primary metrics used for the evaluation, and to provide additional metrics, precision was also used. The higher NMI value denotes that clustering has reduced more entropy. 1:1 alignments between the induced intents and the gold intents were computed by the Hungarian algorithm (Kuhn, 1955).

Setup We employ the pre-trained SBERT (Reimers and Gurevych, 2019) for the baseline embedding model. The pre-trained parameters were from the huggingface (Wolf et al., 2019) *all-mpnet-base-v2* version. In the SCL function, we set the τ as 0.07 and trained a maximum of five epochs with early stopping. In the K-means clustering, we set the minimum cluster number as five and the max cluster number as 50 and use silhouette score for comparing the clustered result, which is based on tightness and separation (Rousseeuw, 1987).

5.2 Intent clustering result

Model	Data	NMI	ACC	Precision
Different Domain with Fine-tuning Dataset				
Baseline	Dev(Insurance)	59.31	46.14	65.98
DORIC	Dev(Insurance)	**65.16**	**56.68**	**67.63**
Baseline	Test1(Banking)	65.71	51.85	60.68
DORIC	Test1(Banking)	**71.02**	**52.35**	**73.92**
Baseline	Test2(Finance)	60.26	59.75	69.25
DORIC	Test2(Finance)	**69.64**	**65.14**	**75.14**
Same Domain with Fine-tuning Dataset				
Baseline	SGD(Tourism)	60.54	63.67	49.90
DORIC	SGD(Tourism)	**65.32**	**68.36**	**51.00**

Table 4: Comparison of baseline and DORIC in different dataset. NMI, ACC and Precision are reported.

The results of DORIC after evaluation on the dev (insurance), test1 (finance), test2 (banking), and SGD (tourism) datasets are shown in Table 4. These results show that our model outperforms the baseline model in terms of the NMI, ACC, and precision on all datasets. Except for the SGD dataset, the dataset's domains are all different from the fine-tuning dataset MultiWOZ2.2 (tourism), which demonstrates that our intent induction framework is robust to diverse domain datasets. The visualization of experimental results in Figure 3 also exhibits the aligned result with Table 4; compared to the baseline, DORIC embeds utterances with the same label at a closer distance.

5.3 Intent label generation with hypernym

Table 5 shows examples of the generated intent labels, and cluster with †denotes the clusters with the hypernyms following section 4.3. As shown in the table, our proposed method successfully explains the cluster results compared to the ground truth label. Furthermore, using hypernyms enables the grouping of detailed information in the cluster. For

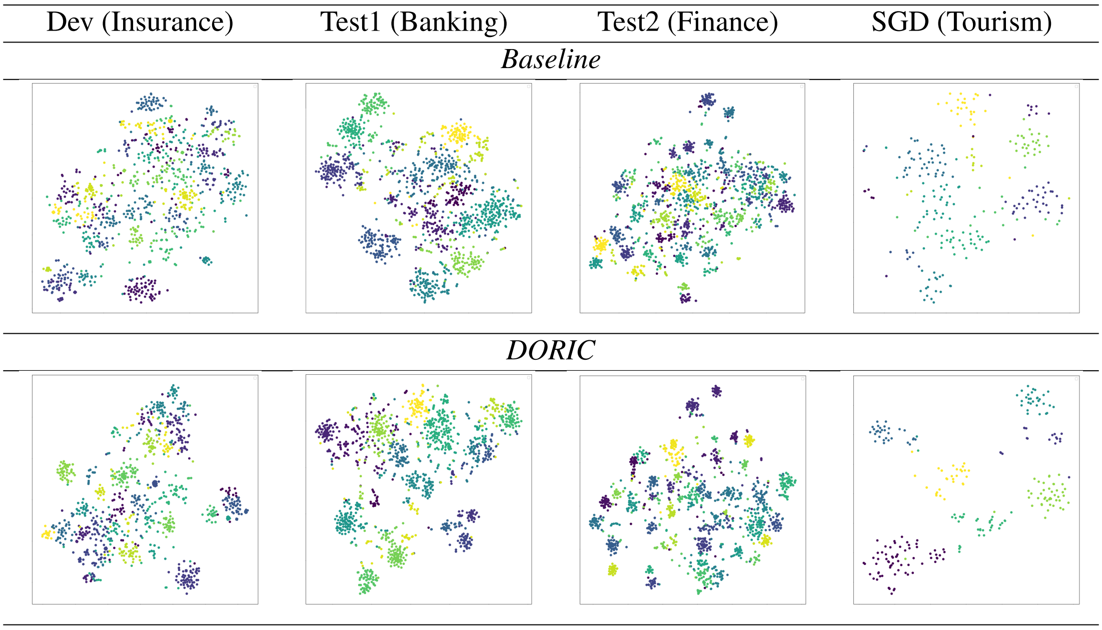

Figure 3: Visulization on dev (insurance), test1 (banking), test2 (finance) and SGD (tourism) dataset.

Idx	Generated name	Ground-truth
0	create-account	CreateAccount
1	cancel-billing	CancelAutomaticBilling
2†	**add-child**	AddDependent
3	report-accident	ReportAutomobileAccident
4	change-address	ChangeAddress
5†	get-quote	GetQuote
6	change-plan	ChangePlan
7	file-claim	FileClaim
8	pay-bill	PayBill
9	check-balance	CheckAccountBalance
10	change-question	ChangeSecurityQuestion
11	cancel-plan	CancelPlan
Without hypernym		
2	**add-son**	AddDependent
5	get-quote	GetQuote

Table 5: Example of generated intent labels and ground truth. Cluster name with †means using hypernym.

Method	Dataset (domain)	NMI	ACC
Different Domain with Fine-tuning Dataset			
Sentence	Dev (Insurance)	62.13	55.35
Verb-Obj	Dev (Insurance)	**65.16**	**56.68**
Sentence	Test1 (Banking)	68.91	**53.22**
Verb-Obj	Test1 (Banking)	**71.02**	52.35
Sentence	Test2 (Finance)	64.94	**67.07**
Verb-Obj	Test2 (Finance)	**69.64**	65.14
Same Domain with Fine-tuning Dataset			
Sentence	SGD (Tourism)	65.24	68.35
Verb-Obj	SGD (Tourism)	**65.32**	**68.36**

Table 6: The NMI and accuracy result on DSTC11 development, test, and SGD dataset according to fine-tuning utterance format.

instance, Cluster 2 obtains a more comprehensive label, `add-child` than `add-son` by using the hypernyms. We also present the more detailed results for the dev and test data in Appendix A.1.

6 Analysis

6.1 Verb-Object structure in fine-tuning

To examine the effect of extracting Verb-Object structures from the sentence, we compare our proposed method with methods that use the whole sentence during the fine-tuning stage (Table 6). Using the Verb-Object structure demonstrates superior NMI results in both different-domain and same-domain environments; this result indicates that fine-tuning with Verb-Object information has helped reduce the clustering uncertainty. However, the accuracy doesn't significantly differ between the Verb-Object form and the whole sentence form in the tourism domain, which is identical to the fine-tuning dataset domain.

6.2 Analysis of loss

To investigate the effect of SCL during fine-tuning, we compare the result with the cross-entropy loss in Table 7. In most cases, the SCL loss demonstrates better results by a large margin; however, on the SGD dataset, the NMI and ACC results were slightly or no different than the cross-entropy loss. Considering that the SGD dataset is the only dataset with the same domain with the fine-tuning dataset (tourism), this result indicates that SCL is more useful when it is used in a domain-across environment.

Loss function	Dataset (domain)	NMI	ACC
Different Domain with Fine-tuning Dataset			
Cross Entropy	Dev (Insurance)	61.98	53.69
SCL	Dev (Insurance)	**65.16**	**56.68**
Cross Entropy	Test1 (Banking)	67.67	52.16
SCL	Test1 (Banking)	**71.02**	**52.35**
Cross Entropy	Test2 (Finance)	64.09	63.07
SCL	Test2 (Finance)	**69.64**	**65.14**
Same Domain with Fine-tuning Dataset			
Cross Entropy	SGD (Tourism)	64.11	**70.31**
SCL	SGD (Tourism)	**65.32**	68.36

Table 7: The NMI and accuracy result on DSTC11 development, test, and SGD dataset according to the loss function.

7 Conclusion

In this paper, we describe our solution for the DSTC11 intent induction competition. We leveraged the SBERT model to embed sentences and fine-tuned the model using dependency parsing results. Additionally, we used supervised contrastive loss during fine-tuning to make the model robust in multiple domains. During the analysis, both dependency parsing and SCL helped to make the intent induction model more domain robust. Furthermore, our intent label generation with hypernym methods allows us to explain the clustering results. According to the results, our approach achieved 3rd place in terms of the precision score and demonstrated better NMI and accuracy compared to the baseline model.

Limitations

Our contribution has two limitations. First, although DORIC shows superior performance in the domain across the environment, the increase was insignificant in the same domain environment. Second, we thoroughly examine the embedding methods, but we adapt this method only to the K-means clustering. In the future, we plan to devise a progressed clustering method that fits our embedding method.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)) and ITRC(Information Technology Research Center) support program(IITP-2023-2020-0-01789).

References

Ajay Chatterjee and Shubhashis Sengupta. 2020. Intent mining from past conversations for conversational agent. *arXiv preprint arXiv: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.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Young-Bum Kim, Sungjin Lee, and Karl Stratos. 2017. Onenet: Joint domain, intent, slot prediction for spoken language understanding. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 547–553. IEEE.

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

Jihyun Lee and Gary Geunbae Lee. 2022. Sf-dst: Few-shot self-feeding reading comprehension dialogue state tracking with auxiliary task. *arXiv preprint arXiv:2209.07742*.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. *arXiv preprint arXiv:1906.00434*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.

Hao Liu and Pieter Abbeel. 2020. Hybrid discriminative-generative training via contrastive learning. *arXiv preprint arXiv:2007.09070*.

Pengfei Liu, Youzhang Ning, King Keung Wu, Kun Li, and Helen Meng. 2021. Open intent discovery through unsupervised semantic clustering and dependency parsing. *arXiv preprint arXiv:2104.12114*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. *arXiv preprint arXiv:1908.11487*.

Patti Price. 1990. Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Schema-guided dialogue state tracking task at dstc8. *arXiv preprint arXiv:2002.01359*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Jinpeng Wang, Gao Cong, Xin Wayne Zhao, and Xiaoming Li. 2015. Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. *arXiv preprint arXiv:2004.06871*.

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.

Juan Xu, Qi Zhang, and Xuanjing Huang. 2013. Understanding the semantic intent of domain-specific natural language query. In *Proceeding of International Joint Conference on Natural Language Processing*, pages 552–560.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

A Appendix

A.1 Generated cluster label name in detail.

Cluster	TOP 3 Verb-Object Pairs	Example	Ground-truth
0	('create-account', 13), ('make-account', 5), ('open-account', 3)	∼create an account for my new renter policy.	CreateAccount
1	('cancel-billing', 6), ('stop-payment', 3), ('cancel-payment', 2)	∼to cancel the automatic billing of my account ∼	CancelAutomaticBilling
2†	('add-child', 10), ('add-son', 5), ('add-male_offspring', 5)	Oh good, help me add my son.	AddDependent
3	('report-accident', 10), ('file-claim', 1), ('start-process', 1)	Hey yeah, I have to call and report an accident.	ReportAutomobileAccident
4	('change-address', 31), ('update-address', 9), ('update-information', 2)	Hi, I'd like to change my address to a new one ∼	ChangeAddress
5†	('get-quote', 10), ('get-punctuation', 10), ('get-interruption', 10)	∼to get a quote from you guys.	GetQuote
6	('change-plan', 9), ('upgrade-plan', 1), ('get-quote', 1)	Eee yeah I think I want to change my plan	ChangePlan
7	('file-claim', 20), ('report-claim', 3), ('make-claim', 2)	I'd like to file a property claim.	FileClaim
8	('pay-bill', 17), ('get-bill', 2), ('take-care', 2)	Well, I'm calling to pay my bill.	PayBill
9	('check-balance', 4), ('pay-bill', 2), ('confirm-balance', 2)	Yes I need to check the balance ∼	CheckAccountBalance
10	('change-question', 10), ('remember-question', 2), ('keep-stuff', 1)	∼to change my security question and answer.	ChangeSecurityQuestion
11	('cancel-plan', 24), ('cancel-policy', 8), ('cancel-insurance', 6)	∼I need to cancel my plan.	CancelPlan

Table 8: Generated intent cluster label, example, and ground truth on dev (insurance) dataset. Cluster name with †means using hypernym.

Cluster	TOP 3 Verb-Object Pairs	Example	Ground-truth
0	('dispute-transaction', 17), ('have-transaction', 3), ('dispute-charge', 2)	almost. I need to dispute this transaction I found for Piggly Wiggly.	DisputeCharge
1	('open-account', 56), ('set-banking', 2), ('set-account', 2)	∼ so I would need to open a checking account then, right?	OpenBankingAccount
2	('update-address', 19), ('change-address', 7), ('do-address', 1)	I'd like to update my address I believe when ∼	UpdateStreetAddress
3†	('transfer-money', 23), ('transfer-medium_of_exchange', 23), ('transfer-standard', 18)	And transfer the money from that to my checking.	InternalFundsTransfer
4	('check-balance', 68), ('get-balance', 8), ('give-balance', 7)	yes I need to check the my account balance ∼	CheckAccountBalance
5†	('make-transfer', 23), ('make-movement', 23), ('make-change', 23)	Yeah, I needed to make a wire transfer ∼	ExternalWireTransfer
6	('find-branch', 19), ('locate-branch', 5), ('help-branch', 4)	Yes. I need help with the with finding the nearest branch to my location.	FindBranch

Table 9: Generated intent cluster label, example, and ground truth on test1 (banking) dataset. Cluster name with †means using hypernym.

Cluster	TOP 3 Verb-Object Pairs	Example	Ground-truth
0	('check-balance', 18), ('check-account', 4), ('check-checking', 2)	∼ I am calling to check my account balance.	CheckAccountBalance
1†	('update-number', 13), ('update-amount', 13), ('update-assets', 13)	∼ and next update my phone number for my business∼	UpdatePhoneNumber
2	('check-balance', 16), ('get-balance', 6), ('know-balance', 3)	I need to check the balance on my credit card please	CheckAccountBalance
3	('update-address', 17), ('change-address', 6), ('change-piece', 1)	yeah I need to update my street address on ∼	UpdateStreetAddress
4†	('open-account', 11), ('open-record', 11), ('open-evidence', 11)	∼ I was thinking about opening another account for ∼	OpenAccount
5	('add-user', 10), ('have-access', 2), ('bring-people', 1)	Yeah I need to add some additional users to my account	AddUserToAccount
6†	('make-transfer', 13), ('make-movement', 13), ('make-change', 13)	Hi Jerry. I need to make a wire transfer man.	MakeTransfer

Table 10: Generated intent cluster label, example, and ground truth on test2 (finance) dataset. Cluster name with †means using hypernym.