Interspeech 2022
18-22 September 2022, Incheon, Korea

Induce Spoken Dialog Intents via Deep Unsupervised Context Contrastive Clustering

Ting-Wei Wu[1], *Biing-Hwang Juang*[1]

[1]Georgia Institute of Technology
School of Electrical and Computer Engineering
waynewu@gatech.edu, juang@ece.gatech.edu

Abstract

Intent detection is one of most critical tasks in spoken language understanding. However, most systems could only identify a predefined set of intents, without covering a ubiquitous space of real-world semantics. Discovering new dialog intents with clustering to explore additional requests is crucial particularly in complex domains like customer support services. Leveraging the strong coherence between the user query utterance and their following contexts in the dialog, we present an effective intent induction approach with fine-tuning and clustering with contrastive learning. In particular, we first transform pretrained LMs into conversational encoders with in-domain dialogs. Then we conduct context-aware contrastive learning to reveal latent intent semantics via the coherence from dialog contexts. After obtaining the initial representations on both views of the query and their contexts, we propose a novel clustering method to iteratively refine the representation by minimizing semantic distances between pairs of utterances or contexts, under the same cluster assignment on the opposite view. The experimental results validate the robustness and versatility of our framework, which also achieves superior performances over competitive baselines without the label supervision.

Index Terms: Spoken Language, Intent, Clustering, Context

1. Introduction

Recent advances of task-oriented agents such as Apple Siri, Amazon Alexa or Google DialogFlow have prompted the success of improving more natural customer service automation [1]. They have been widely adapted in several domains like booking flights, restaurants, and customer support [2]. Usually, the very first step of establishing such systems is to determine appropriate ontologies that constrain the dialog intent and state space for a specific task [3]. However, most works with such assumptions usually categorize utterances into few simple intents; yet exclude further pragmatic intent discovery in the real world [4]. In complex domains like customer support and healthcare, users may not be fully cognizant of the system capabilities, where emerging intents would reasonably appear during dialogs that are excluded from few simple system-supported intents [5]. Systems which can barely enumerate a complete view of user intents will easily fail to recognize an emerging intent during shifts between domains or chit-chats [6].

To alleviate the constraints of exploring predefined intents, [7] introduced the dialog induction task to automatically discover user intent in dialogs by employing data-driven clustering approaches. Further analysis could determine the most precious intents for refining dialog systems by contrasting the complete view of a user intent space. However, unlike (written) texts, spoken utterances do not come with clear sentential segmentation by topographic means [8]. The weak noisy signals and sparsity pose a significant challenge for clustering short spoken texts. Learning a discriminative classifier that maps utterances directly to intent labels may suffer from poor purity when representations still have tremendous overlap across categories before the decision boundaries are formed [9].

Recently, contrastive learning based on pretrained language models (LM) such as BERT [10] or RoBERTa [11], offers striking state-of-the-art (SOTA) performances across multiple natural language processing tasks (NLP) [12]. Multiple works leverage contrastive learning to learn unsupervised representation on languages with self-supervised signals such as text augmentation [13, 14, 15] or text masking [16]. Unlike monological discourses, dialog-based contrastive learning is also popular to learn inherent dialog characteristics like sequential orders [17] or coherence [18, 19] among multi-turn interactions for response generation [19] or summarization tasks [20]. However, unsupervised contrastive learning on task-oriented spoken dialogs for intent induction has less been explored where intents of agent context responses could be even more strongly related to user utterances. For instance, in customer services, agent shall always reply to user requests with similar targeted entities. Relying on such context similarity for contrastive learning to enhance model's discrimination abilities would align with the clustering objectives more intuitively than direct modifications on original texts [15]. Inspired by recent research threads of contrastive learning, two immediate questions then surge: Instead of seeking a classifier to directly decide discriminative boundaries between short utterances, is it possible to learn a good intent space where projected samples are densely distributed in clusters and easily segregated, by leveraging information from their corresponding task-oriented contexts? Second, how could we quickly adapt pretrained LMs into a task-specific operator that generates such representations?

To address these problems, we propose a new unsupervised framework to induce dialog intents. The key is to learn query utterance representations in a semantic intent space where the same intent cluster of two different user utterances could be induced by their similar agent contexts. We first divide each dialog into the user query utterance and their following contexts and perform two stages of fine-tuning with the Masked Context Modeling (MCM) pretext task and the proposed context-aware contrastive learning. After obtaining an in-domain encoder, we iteratively perform k-means-style updates to compute the cluster assignment on these representations and perform additional contrastive learning with pseudo labels to further manifest each intent cluster. Our work diverges from previous work on multi-view clustering [7, 21], where it learns robust representations by explicitly comparing query utterances' distance with their corresponding contexts. Experimental results on two intent induction datasets show our method outperforms several strong baselines substantially.

Copyright © 2022 ISCA

10.21437/Interspeech.2022-240

Our contributions are summarized as follows:
1) We make the first attempt of introducing contrastive learning in the dialog induction task without any label supervision.
2) We introduce a novel contrastive clustering method that jointly learns representations and cluster assignments mutually.
3) Experimental results verify the SOTA performance of our approach on two intent induction datasets and t-SNE visualization verifies the formation of intent clusters.

2. Methodology

Our framework could be divided into three phases: conversational fine-tuning, context-aware contrastive learning, and contrastive clustering. Given any set of dialogs $X \in \mathcal{X}$, we view each sample as the following format: $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{c}_i)$, $\mathbf{u}_i$ is the first query utterance and $\mathbf{c}_i$ with $L-1$ length is the following contexts. These two views of data will be used in the following three phases.

2.1. Conversational Fine-Tuning

We first cast pretrained LMs, which are BERT [10], RoBERTa [11], DistilRoBERTa [22] in this paper, into in-domain conversational encoders. We implicitly learn the token-level characteristics by adopting Mask Context Modeling (MCM) task. We first concatenate each utterance $\mathbf{u}_i$ and their contexts $\mathbf{c}_i$ with `[SEP]` as $\mathbf{x}_i$. Different from Mask Language Modeling, we only mask tokens in the contexts $\mathbf{c}_i$ with 80% of the time a special token `[MASK]`, 10 % of the time random word tokens from vocabulary and others remaining the same. This strengthens the model to take account of a complete utterance to predict the mask tokens in $\mathbf{c}_i$ which empirically works better. We then retrieve the feature representation $\mathbf{h}_i$ for the masked dialog $\tilde{\mathbf{x}}_i$ with each token as $\tilde{\mathbf{x}}_i^t$ with the encoder $\phi_\theta(\cdot)$ and project it with a projection matrix $\mathbf{W}_v \in \mathbb{R}^{H \times |V|}$ to transform $\mathbf{h}_i$ into one-hot vectors $\hat{\mathbf{y}}_i$, where H is BERT hidden size and $|V|$ is vocabulary size. We then update the parameters θ and $\mathbf{W}_v$ to minimize the cross entropy losses between the predicted probabilities and their real words at the masked timesteps only.

2.2. Context-aware Contrastive Learning (CACL)

Inspired by the recent success of contrastive learning in the few-shot setting [9, 23], we found similarity-based pretraining benefits the metric-based clustering to generate distant-sensitive representations. Given a large dialog unannotated corpus, we assume that utterances and their corresponding contexts, especially in customer services, should induce the same or related domain intent, since goals could be gradually inferred through both parties. If we try to learn a good representation $\phi(\mathbf{u}_i)$ close to the optimal ϕ^* by semantically minimizing distances between $\mathbf{u}_i$ and $\mathbf{c}_i$, according to [24], we are implicitly forcing LMs to first predict their latent intent Y given $\mathbf{u}_i$; and generate $\mathbf{c}_i$ from Y, which is exactly how agents respond to the intent of the request. It elucidates that with a good representation from ϕ^*, we can cluster $\mathbf{u}_i$ based on similarity by inducing their underlying intents Y. Hence, we leverage the dual-encoder architecture [25] to model the relevance between a pair $(\mathbf{u}, \mathbf{c})$ of an utterance and their contexts.

Context Filtering: We will use the fine-tuned encoder $\phi(\cdot)$ from section 2.1 to obtain the hidden representation $\mathbf{h}$ by taking the final encoding from the `[CLS]` token via a *pooling* operation. But before sending the dialog into the pretrained encoder $\phi(\cdot)$, some redundant sentences $\mathbf{c}_j$ in the dialog contexts $\mathbf{c} = \{\mathbf{c}_j\}$ may cause frictions to obfuscate user intents. Instead of encoding all dialog contexts, we first obtain the hidden representations for each sentence in a dialog $(\mathbf{h}_u, \mathbf{h}_c^1, ..., \mathbf{h}_c^{L-1})$ where L is the dialog length. Then we calculate the cosine distance between each dialog context representation and the utterance. Eventually, we take the closest context as the final sentence pair $(\mathbf{u}, \mathbf{c}_{j^*})$, where $j^* = \arg\min_{j \in [1, L-1]}(1 - cos(\mathbf{h}_u, \mathbf{h}_c^j))$ and $cos(\cdot, \cdot)$ is the cosine similarity.

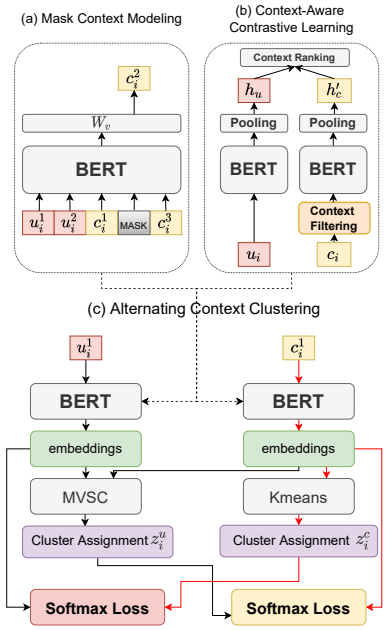


Figure 1: *Illustration of the proposed framework for dialog intent induction. There are three phases including two phases of fine-tuning and detection. Part (c) will be performed iteratively by refining the encoder obtained from (a) and (b).*

Context Ranking: We introduce two new tokens `[QUERY]` and `[CONTEXT]` for the sequences $\mathbf{u}$ and $\mathbf{c}_{j^*}$ respectively and send them into the encoder to obtain hidden representations $(\mathbf{h}_u, \mathbf{h}_c^{j^*})$ at the head from `[QUERY]` and `[CONTEXT]`. For each pair of $(\mathbf{h}_u, \mathbf{h}_c^{j^*})$, we use the context ranking to fine-tune our encoder ϕ. We leverage the standard multiple negatives ranking loss (MNEG) [30] to rank the correct context $\mathbf{c}_{j^*}$ for the utterance u_i over other negative contexts $\mathbf{c}_{k \setminus j^*}$ in a single batch. MNEG loss for a single batch $(u_1, c_1), ..., (u_{N_B}, c_{N_B})$ is calculated as the following:

$$\mathcal{L}_{MNEG} = -\sum_{i=1}^{N_B}(S \cdot cos(u_i, c_i) - log \sum_{j=1, j \neq i}^{N_B} e^{S \cdot cos(u_j, c_j)}) \tag{1}$$

where S is a scaling factor. It will try to maximize the correct match score and abate negative scores. The output will be the sentence encoder ϕ^* which could be used for iterative clustering assignment and refinement.

2.3. Context Contrastive Clustering (CC)

After two-phase fine-tuning, our main goal is to induce the intents from a collection of our interested user queries. We follow [7] to formulate it as an unsupervised clustering task on these utterances. We present a new method, context contrastive clustering to continuously refine the neural encoders by encouraging utterances and their corresponding contexts to yield similar

Dataset		TwACS				AskUbuntu			
Encoder	**Clustering**	**P**	**R**	**F1**	**ACC**	**P**	**R**	**F1**	**ACC**
Autoencoder [7]	k-means	27.1	18.7	22.2	15.9	27.9	19.3	22.8	14.4
Autoencoder [7]	MVSC [21]	29.3	21.2	24.6	18.9	28.6	13.0	17.9	11.2
DialoGPT [26]	k-means	28.2	16.2	20.6	15.1	20.1	15.4	17.5	13.3
BERT [10]	k-means	30.6	24.1	27.0	19.5	38.0	27.9	32.2	26.4
TOD-BERT [27]	k-means	31.7	24.5	27.7	19.9	36.8	30.5	33.3	28.6
SBERT [28]	k-means	36.5	27.8	31.6	25.2	50.9	44.7	47.6	41.3
SBERT [28]	SCCL [15]	30.8	34.4	32.5	26.5	48.4	37.6	42.3	36.2
Autoencoder [7]	Avkmeans	46.0	33.6	38.9	31.6	50.2	43.8	46.8	36.2
QuickThoughts [29]	Avkmeans	51.1	38.1	43.7	35.9	56.6	55.1	55.8	43.2
BERT [10] + NCM + CACL (ours)	k-means	51.9	39.4	44.8	36.1	71.1	64.3	67.6	58.4
BERT [10] + NCM + CACL (ours)	MVSC [21]	51.9	42.2	46.6	41.1	69.1	64.4	66.7	62.4
ROB [11] + NCM + CACL (ours)	Contrastive clustering	53.0	43.8	47.9	42.7	73.3	66.7	69.9	62.8
DROB [22] + NCM + CACL (ours)	Contrastive clustering	**54.3**	45.3	49.4	43.1	73.6	67.9	70.6	64.0
BERT [10] + NCM + CACL (ours)	Contrastive clustering	54.1	**46.4**	**49.9**	**43.3**	**76.4**	**71.4**	**73.8**	**68.1**

Table 1: *Experimental results (%) for various encoders and cluster strategies on two evaluation datasets. The best results are **in bold**.*

cluster assignments. We first initialize the pretrained encoder ϕ obtained from the section 2.2 and obtain two views of representations: $\mathbf{h}_u^i, \mathbf{h}_c^i$. To note, we empirically found better to use the entire contexts for $\mathbf{h}_c$ instead of the single sentence after context filtering. Then, we obtain the initial cluster assignment for utterances $\{z_i^u\}$ by using *multi-view spectral clustering (MVSC)* [21] that also considers context-view for matching. To further induce the prior where agent contexts will be more similar than utterances under same request intents, we retrieve the cluster assignment for contexts $\{z_i^c\}$ alone with k-means clustering. Two assignments may not be exactly matched since utterances and contexts may not be aligned in terms of low-level semantics.

During each iteration, we refine our encoder with the context ranking described in the section 2.2 with a loss $\mathcal{L}_{MNEG}$ and another supervised learning technique by leveraging $\{z_i^u\}$ and $\{z_i^c\}$. In particular, we wish utterances and contexts should have the same cluster assignments. Hence we project the context cluster assignment $\{z_i^c\}$ onto the utterances and perform supervised contrastive learning depicted in [23]. We treat two utterances from the same class as a positive pair and the two utterances across different classes as a negative pair. Then we update our encoder with the following loss:

$$\mathcal{L}_{CAC}^u = -\frac{1}{T}\sum_{i=1}^{N_B}\sum_{j=1}^{N_B}\mathbb{1}_{y_i=y_j} log\frac{exp(cos(\mathbf{h}_u^i, \mathbf{h}_u^j)/\tau)}{\sum_{n=1}^{N_B} exp(cos(\mathbf{h}_u^i, \mathbf{h}_u^n)/\tau)} \quad (2)$$

$\mathbb{1}$ is the indicator function. We repeat the procedure for contexts too by applying the utterance cluster assignment $\{z_i^u\}$. Finally, we perform the multi-task learning with the final loss for contrastive learning as $\mathcal{L}_{all} = \mathcal{L}_{MNEG} + \mathcal{L}_{CAC}^u + \mathcal{L}_{CAC}^c$.

3. Experimental Setup

3.1. Datasets and baselines

We use two datasets (TwACS, AskUbuntu) collected from [7] that have smaller amount of annotated user intents and a large amount of raw dialogs. TwACS is a customer support corpus released by Kaggle[1] which has total 28,857 dialogs with 491 of them being annotated into 14 intents. AskUbuntu is maintained for duplicate question detection task with technical support questions which has total 69,927 dialogs with 1029 of them being annotated into 20 intents. For fine-tuning purposes, we only use unlabeled corpus for training. We use the annotated dialogs in TwACS and AskUbuntu to verify our model's performance on the dialog intent induction task.

[1] https://www.kaggle.com/thoughtvector/customer-support-on-twitter

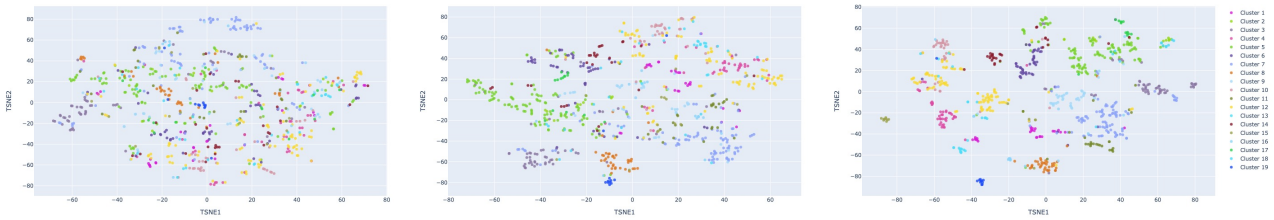
Dataset				TwACS		AskUbuntu	
MCM	**CACL**	**CF**	**CC**	**F1**	**ACC**	**F1**	**ACC**
				27.0	19.5	32.2	26.4
✓				28.4	20.4	48.4	41.7
			✓	38.7	32.4	61.9	55.7
✓	✓			43.3	35.0	62.7	54.2
✓	✓	✓		44.8	36.1	67.6	58.4
	✓	✓		46.7	40.3	64.9	58.8
✓			✓	48.6	41.4	67.1	61.1
	✓	✓	✓	46.6	41.6	70.8	65.5
✓	✓	✓	✓	**49.9**	**43.3**	**73.8**	**68.1**

Table 2: *Ablation study (%) of our proposed approach by permuting components from our final model (✓ indicates the specified component is added). **MCM** is Mask Context Modeling. **CACL** is Context-aware Contrastive Learning. **CF** is whether to use the whole contexts or only the filtered one. **CC** denotes Context Contrastive Clustering. We use kmeans to replace **CC** if no ✓ is presented.*

We experiment with competitive baselines from some popular transformer-based LMs and clustering approaches. The encoders we compare include:

- **Autoencoders** [7]: fine-tuned on evaluation datasets.
- **Quick thoughts (QT)** [29]: a strong representation baseline that is adopted in BERT.
- **BERT** [10]: a contextualized model based on transformers.
- **DialoGPT** [26]: a generative model trained on dialog corpus.
- **TOD-BERT** [27]: BERT fine-tuned on conversational data based on BERT.
- **Sentence-BERT (SBERT)** [28]: a dual encoder structure for similarity training.
- **SCCL** [15]: an instance-wise contrastive learning approach with contextual augmentation on SBERT which simultaneously minimizes KL divergence between the cluster assignment probability and a target distribution.

We equip and test our framework with three LMs, BERT [10], RoBERTa (ROB) [11], an improved variant of BERT and DistilRoBERTa (DROB) [22], a distilled version of RoBERTa, trained with around four times fewer data than the teacher RoBERTa model. We compare their performances against an array of cutting-edge conversational sentence encoders. All



(a) BERT (no fine-tuning) (b) BERT (Phase 1: MCM) (c) BERT (Phase 1: MCM + Phase 2: CACL)

Figure 2: *t-SNE plots of encoded utterances from AskUbuntu. The representations are created by BERT-based encoder ϕ (a) without any fine-tuning (b) after MCM fine-tuning and (c) our final model. The colors are true labels.*

of models contain 768-dimensional transformer layers with 12 (BERT, ROB) or 6 (DROB) attention layers. There are five different clustering strategies for comparison: k-means, multi-view spectral clustering (MVSC) [21], SCCL [15], Av-Kmeans [7], and our contrastive clustering. We evaluate different strategies of dialog induction tasks by comparing their predicted cluster assignments with the real labels in two datasets.

3.2. Experimental Setting

We use Sentence-Transformers package [28] to implement our framework and choose *paraphrase-distilroberta-base-v1* as the backbone. We first fine-tune the LMs with conversational fine-tuning. We train for 4 epochs in batches of 16. We also set the max length as 200 and mask probability of 0.15. For context-aware contrastive learning, we train each model with 2 epochs in batches of 8 to prevent overfitting. We run the iterative clustering for $T = 10$ iterations. We use the AdamW optimizer and $2e-5$ learning rate and weight decay rate is 0.01. We run each experiment 5 times and report the average. We follow the metrics in [7] for clustering precision/recall, f1 score and accuracy.

4. Results

4.1. Main Clustering Results

The main results are summarized in Table 1. We observe that directly deploying universal DialoGPT or BERT representations may suffer from significant overlap between intent clusters which causes low performances. Training with conversational data like TOD-BERT offers better intent disentanglement. The results suggest that the pretraining objectives may substantially influence the downstream clustering task. We observe even competitive performance when we apply SBERT dual-encoder structure which intrinsically compares pairs of sentences for similarity. However, we only see slight or no improvement with SCCL which indicates that augmenting only on utterances may not be enough regardless of applying with contrastive learning. Imbalanced labels may also contradict with the objective of KL divergence minimization to form uniform clusters.

By leveraging the two-phase fine-tuning of adapting BERT into a context-aware encoder, the representations could be unanimously separated better with clustering algorithms k-means and MVSC, which shows stronger performance over the previous SOTA Avkmeans algorithm baseline. Training the encoder with context ranking successfully establish an intent subspace where utterances are close to their contexts, which congruently approximates the process of disclosing their underlying latent intents. Finally, we see an improved performance by adopting our proposed clustering approach. Such cross-instance similarities are especially useful when we iteratively refine the cluster assignment to reach the agreement between utterances and their corresponding contexts. We finally see BERT in our case outperforms ROB and DROB which may suggest the MCM task may align with BERT pretraining more.

4.2. Ablation Study

A more careful ablation comparison is conducted in Table 2 to investigate the effects of different components, all of which are necessary to achieve the peak performance. Using the raw BERT model without learning any in-domain data will result in the poor discriminative performance. Fine-tuning with in-domain data along with contexts is important to refine the representations for better dialog clustering, particularly with our proposed contrastive learning method. Around 3-6 % performance increase is observed when we perform context filtering where redundant noises may disturb. Finally, we see by adopting further iterative context clustering will benefit the model discrimination ability of semantically aligning similar utterances with their contexts.

4.3. t-SNE visualization

To understand how well the latent space is formulated where encoded utterances are mapped onto, we perform t-SNE visualization in Figure 2 on three types of representations generated with disparate pretrained encoders. First, we could see without any fine-tuning, utterances with distinct intents are entangled into a single mixed cluster. After conversational fine-tuning of transforming LMs into conversational encoders, we start to see some clusters with intents apart. Eventually, our final model further specializes the sentence encoder to learn meaningful task-related semantic clusters even without any label supervision.

5. Conclusions

In this work, we present an effective approach to induce dialog intents. We propose to transform LMs into conversational encoders and retrain them with context-aware contrastive learning. Finally, we alternatively meet the consistency of cluster assignments on utterances and their contexts to reach the best performance. We demonstrate that by extracting such fine-grained intent-aware representations, it is possible to separate complex dialogs in customer services based on their latent intents with the mutual agreement between utterances and their contexts. Comparing existing system-supported intents with intent clusters could better exude the new undiscovered intents.

6. References

[1] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog system," 2020.

[2] L. L. Jianfeng Gao, Michel Galley, "Neural approaches to conversational ai," *arXiv:1809.08267*, 2019.

[3] H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han, "A survey of joint intent detection and slot-filling models in natural language understanding," 2021.

[4] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2018.

[5] 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1311–1316. [Online]. Available: https://aclanthology.org/D19-1131

[6] D. Chen and Z. Yu, "GOLD: Improving out-of-scope detection in dialogues using data augmentation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 429–442. [Online]. Available: https://aclanthology.org/2021.emnlp-main.35

[7] H. Perkins and Y. Yang, "Dialog intent induction with deep multi-view clustering," 2020.

[8] R. F. Jonathan Ginzburg, *Computational Models of Dialogue*, 2010.

[9] I. Vulić, P.-H. Su, S. Coope, D. Gerz, P. Budzianowski, I. Casanueva, N. Mrkšić, and T.-H. Wen, "Convfit: Conversational fine-tuning of pretrained language models," 2021.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[12] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić, "Efficient intent detection with dual sentence encoders," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Online: Association for Computational Linguistics, Jul. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.nlp4convai-1.5

[13] K. Wang, N. Reimers, and I. Gurevych, "Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning," in *EMNLP*, 2021.

[14] 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*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552

[15] D. Zhang, F. Nan, X. Wei, S.-W. Li, H. Zhu, K. 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*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5419–5430. [Online]. Available: https://aclanthology.org/2021.naacl-main.427

[16] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," 2021.

[17] J. Wu, X. Wang, and W. Y. Wang, "Self-supervised dialogue learning," 2019.

[18] C. Liu, R. Wang, J. Liu, J. Sun, F. Huang, and L. Si, "DialogueCSE: Dialogue-based contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2396–2406. [Online]. Available: https://aclanthology.org/2021.emnlp-main.185

[19] S. Dai, G. Wang, S. Park, and S. Lee, "Dialogue response generation via contrastive latent representation learning," in *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*. Online: Association for Computational Linguistics, Nov. 2021, pp. 189–197. [Online]. Available: https://aclanthology.org/2021.nlp4convai-1.18

[20] J. Liu, Y. Zou, H. Zhang, H. Chen, Z. Ding, C. Yuan, and X. Wang, "Topic-aware contrastive learning for abstractive dialogue summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1229–1243. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.106

[21] S. Kanaan-Izquierdo, A. Ziyatdinov, and A. Perera-Lluna, "Multiview and multifeature spectral clustering using common eigenvectors," *Pattern Recognition Letters*, vol. 102, pp. 30–36, Jan. 2018.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[23] J. Zhang, T. Bui, S. Yoon, X. Chen, Z. Liu, C. Xia, Q. H. Tran, W. Chang, and P. Yu, "Few-shot intent detection via contrastive pre-training and fine-tuning," 2021.

[24] J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo, "Predicting what you already know helps: Provable self-supervised learning," 2021.

[25] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[26] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," 2020.

[27] C.-S. Wu, S. Hoi, R. Socher, and C. Xiong, "Tod-bert: Pre-trained natural language understanding for task-oriented dialogue," 2020.

[28] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[29] L. Logeswaran and H. Lee, "An efficient framework for learning sentence representations," 2018.

[30] M. Henderson, R. Al-Rfou, B. Strope, Y. hsuan Sung, L. Lukacs, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil, "Efficient natural language response suggestion for smart reply," 2017.