Dial2vec: Self-Guided Contrastive Learning of Unsupervised Dialogue Embeddings

Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li*, Fei Huang
DAMO Academy, Alibaba Group
{liuche.lc,wr224079,jiangjunfeng.jjf,shuide.lyb,f.huang}@alibaba-inc.com

Abstract

In this paper, we introduce the task of learning unsupervised dialogue embeddings. Trivial approaches such as combining pre-trained word or sentence embeddings and encoding through pre-trained language models (PLMs) have been shown to be feasible for this task. However, these approaches typically ignore the conversational interactions between interlocutors, resulting in poor performance. To address this issue, we proposed a self-guided contrastive learning approach named dial2vec. Dial2vec considers a dialogue as an information exchange process. It captures the conversational interaction patterns between interlocutors and leverages them to guide the learning of the embeddings corresponding to each interlocutor. The dialogue embedding is obtained by an aggregation of the embeddings from all interlocutors. To verify our approach, we establish a comprehensive benchmark consisting of six widely-used dialogue datasets. We consider three evaluation tasks: domain categorization, semantic relatedness, and dialogue retrieval. Dial2vec achieves on average 8.7, 9.0, and 13.8 points absolute improvements in terms of purity, Spearman's correlation, and mean average precision (MAP) over the strongest baseline on the three tasks respectively. Further analysis shows that dial2vec obtains informative and discriminative embeddings for both interlocutors under the guidance of the conversational interactions and achieves the best performance when aggregating them through the interlocutor-level pooling strategy. All codes and data are publicly available at https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/dial2vec.

1 Introduction

Dialogue embedding, as a critical prerequisite of semantically understanding a dialogue, has been a central issue in dialogue-related research such as dialogue clustering (Shi et al., 2018; Lv et al., 2021), conversational sentiment analysis (Wang et al., 2020; Lv et al., 2021), context-dependent text-to-SQL (Hui et al., 2021; Wang et al., 2022), and dialogue summarization (Liu et al., 2019b; Liu and Chen, 2021). Trivial unsupervised approaches generally encode dialogues by combining their pre-trained word or sentence embeddings (Pennington et al., 2014; Reimers and Gurevych, 2019) or using PLMs (Wu et al., 2020a; Bao et al., 2020; He et al., 2022a,b,c). However, such methods are not specifically designed for dialogues and thus fail to adequately capture the key conversational information. In this paper, we formally introduce the task of learning unsupervised dialogue embeddings, which aims to learn dialogue embeddings that can well reflect conversational semantics without any additional manual annotations.

Previous studies have extensively demonstrated the importance of encoding token-level interactions for learning semantic textual embeddings. However, for dialogue embedding, encoding interlocutor-level interactions is also essential but is overlooked in trivial approaches. Figure 1 shows

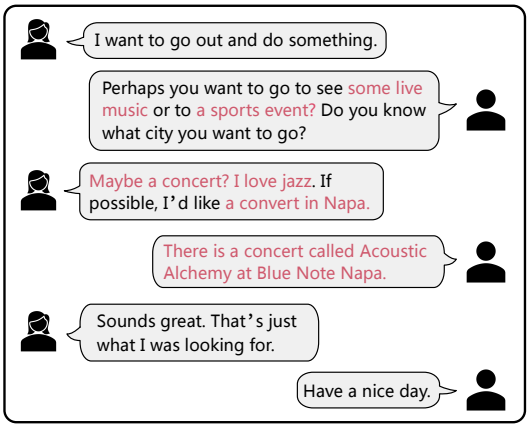

Figure 1: A dialogue from the SGD dataset.

an example. We highlight the significant interaction patterns between the interlocutors with red

* Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7272 - 7282
December 7-11, 2022

color. As we can see, although these patterns only appear in three utterances, they highly represent the key conversational semantics (e.g., topics) and are more important than the other parts (e.g., greetings and chit-chats). We hold that capturing and leveraging them is one of the keys to learning high-quality unsupervised dialogue embeddings.

In this work, we propose dial2vec, a self-guided contrastive learning approach to solve the proposed task. Dial2vec considers a dialogue as an information exchange process between the two interlocutors and learns embeddings for both interlocutors with the help of each other. Specifically, dial2vec firstly encodes a dialogue through a PLM and assigns each interlocutor a self-representation by masking the non-corresponding positions in the encoding outputs. Then it calculates a matching matrix via the token-level dot-product operation between the two self-representations, obtaining a cross-representation for each interlocutor. Finally, the two cross-representations are leveraged as guidance to help the two self-representations gradually learn the interlocutor-level interaction-aware information and eliminate the interaction-free information during the training procedure.

To verify our model, we build a comprehensive benchmark comprising a total of 98,879 dialogues by introducing six widely-used dialogue datasets, including BiTOD (Lin et al., 2021), Doc2dial (Feng et al., 2020), MetalWOZ (Lee et al., 2019), MultiWOZ (Eric et al., 2019), Self-dialogue (Fainberg et al., 2018), and SGD (Rastogi et al., 2020). Each dataset consists of thousands of dialogues, where each dialogue is provided with a domain label (e.g., hotel booking and movie). We leverage these labels and design three evaluation tasks: domain categorization, semantic relatedness, and dialogue retrieval. We categorize them into intrinsic and extrinsic tasks according to their different focus.

Experimental results on this benchmark show that dial2vec outperforms the baselines by a substantial margin. Compared with the strongest baseline, dial2vec achieves on average 8.7, 9.0, and 13.8 points absolute improvements in terms of purity, Spearman's correlation, and mean average precision (MAP) on the three tasks respectively. We also conduct experiments with the single interlocutor's embeddings, their aggregation strategies, and the overall dialogue embedding distributions to study how dial2vec achieves such advanced performance. The results demonstrate that dial2vec learns both informative and discriminative embeddings for the two interlocutors and achieves the best performance when combining them through the proposed interlocutor-level pooling aggregation strategy.

2 Related Work

2.1 Text Embedding

Text embedding aims to encode a piece of text into a distributed vector that could represent its semantics. Early works (Bengio et al., 2003; Mikolov et al., 2013; Pennington et al., 2014) learn unsupervised word embeddings by making use of word-level co-occurrence information in the skip-gram or CBOW tasks. Recently, Devlin et al. (2018); Liu et al. (2019a); Yang et al. (2019); Raffel et al. (2020) pre-train deep transformer (Vaswani et al., 2017) with a series of pretext tasks, setting a new state-of-the-art across the GLUE benchmark (Wang et al., 2018) as well as exhibiting a strong potential in producing general text embeddings. Along this line, Gao et al. (2021); Yan et al. (2021); Liu et al. (2021); Chuang et al. (2022); Nishikawa et al. (2022); Zhou et al. (2022); Klein and Nabi (2022) fine-tune the PLMs with contrastive learning objectives, achieving remarkable improvements in learning unsupervised sentence embeddings. Luo et al. (2021) introduce a data augmentation-based contrastive learning approach in learning document embeddings, achieving superior performance over word2vec-based approaches (Le and Mikolov, 2014; Chen, 2017).

For dialogue embedding, the above approaches are generally unsatisfactory, as they typically obtain dialogue embeddings by averaging the pre-trained word or sentence embeddings, ignoring the interlocutor-level conversational interactions. Although conversational-PLMs pre-trained with dialogue data can solve this problem to some extent (Wu et al., 2020a; Bao et al., 2020; Roller et al., 2021), they mainly focus on learning end-to-end models which are not sufficient for our task. As a comparison, we study how to produce high-quality dialogue embeddings by fully exploiting the conversational information.

2.2 Contrastive Learning

Contrastive learning is an emerging self-supervised learning method which can improve the representation capability of PLMs in both pre-training and fine-tuning stages. Wu et al. (2020b); Meng et al.

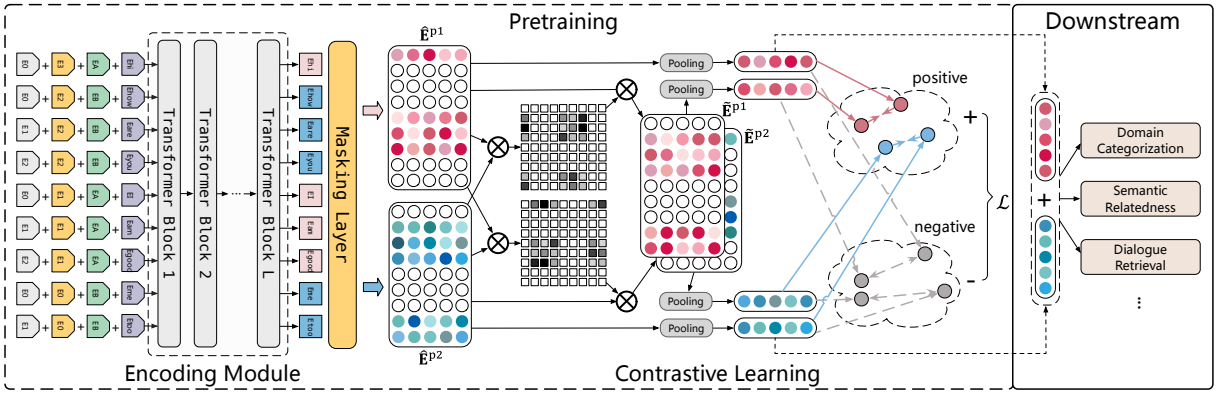

Figure 2: Architecture of dial2vec. Firstly, it encodes a dialogue through a PLM and assigns each interlocutor a self-representation through a masking layer (highlighted with yellow). Hollow circles in each self-representation represent zero embeddings. Then two matching matrices are calculated through the dot-product multiplication, based on which two cross-representations are generated. Each cross-representation and its corresponding self-representation are complementary in the token sequence dimension. Finally, the cosine distance between them will be minimized or maximized according to whether the training sample is positive or negative.

(2021); Giorgi et al. (2020) introduce the token-level and sentence-level contrastive learning tasks by correcting corrupted texts to encourage PLMs to learn noise-invariant representations. Zhang et al. (2022) propose phrase-guided and tree-guided contrastive learning objectives to inject syntactic knowledge into PLMs. Kim et al. (2021) propose a self-guided learning objective through which a PLM fine-tunes itself under the guidance of its different layers. Inspired by these works, we propose to leverage the interlocutor-level conversational interactions to guide the learning of dialogue embeddings in an unsupervised learning manner.

3 Proposed Approach

In this section, we take a two-party dialogue as an example to describe how dial2vec works. It is worth mentioning that dial2vec can be extended to the multi-party version through the OVR (one vs. the rest) scheme with no modification of the architecture.

3.1 Training Samples Generation

We first describe how we construct the positive and the negative training samples, which plays a key role in the self-guided contrastive learning approach. Suppose that we have a dialogue dataset $\mathcal{D} = \{S_k\}_{k=1}^{K}$, where $S_k = \{u_1^{p_1}, u_2^{p_2}, u_3^{p_1}, u_4^{p_2}, \ldots, u_{t-1}^{p_1}, u_t^{p_2}\}$ is the k-th dialogue session with t utterances. p_1 and p_2 represent two interlocutors. We treat each utterance in a dialogue as a turn, regardless of which interlocutor it corresponds to. For the convenience of narration, k in S_k is omitted in the following sections.

We treat S (i.e., the original dialogue) as a positive sample. To construct a negative sample S', we first randomly select an interlocutor in S, say p_1, and keep all the turns of it. Then we fill the other turns of S with the utterances of p_2 randomly sampled from all dialogue sessions. For each positive sample, we repeat this operation multiple times to generate the desired number of negative samples.

3.2 Model Architecture

Figure 2 shows the architecture of dial2vec, which consists of two parts: encoding and contrastive learning. After training, dial2vec aggregates the embeddings from both interlocutors to obtain the final dialogue embedding, which is further used for downstream tasks.

3.2.1 Encoding

Following Bao et al. (2020), we use four types of embeddings as input to dial2vec: token embedding, relative positional embedding, turn embedding, and role embedding. To encode the dialogue, we first concatenate all the utterances and then tokenize them through WordPiece (Wu et al., 2016) to obtain a long token sequence. The tokens along with their corresponding position, turn, and role indices are respectively mapped into four embedding spaces and summed to form the final input embedding.

3.2.2 Contrastive Learning

Suppose that the output embeddings from the encoder are $\{\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \ldots, \mathbf{h}_n\}$, where $\mathbf{h}_i \in \mathbb{R}^d$ is the output embedding corresponding to the i-th input token and n is the length of the input sequence,

we stack the output embeddings as a matrix denoted as $\mathbf{E} \in \mathbb{R}^{n \times d}$.

To obtain the self-representations, we first generate two binary mask vectors $\mathbf{m}^{p_1}$ and $\mathbf{m}^{p_2}$ for two interlocutors respectively. Let $m_i^{p_1}$ be the i-th element in $\mathbf{m}^{p_1}$, then $m_i^{p_1}$ is set to 1 only when $\mathbf{h}_i$ is derived from the input token of p_1, otherwise it is 0. Similar operation is applied to generate $\mathbf{m}^{p_2}$. Then, the self-representations are obtained by:

$$\begin{aligned} \hat{\mathbf{E}}^{p_1} &= \mathbf{E} \odot (\mathbf{m}^{p_1})^T, \\ \hat{\mathbf{E}}^{p_2} &= \mathbf{E} \odot (\mathbf{m}^{p_2})^T, \end{aligned} \tag{1}$$

where $\odot$ denote the broadcast element-wise multiplication.

To extract the interaction information, we perform the token-level dot-product multiplication between $\hat{\mathbf{E}}^{p_1}$ and $\hat{\mathbf{E}}^{p_2}$ and compute a correlation score matrix for each interlocutor, which is formulated by:

$$\begin{aligned} \mathbf{C}^{p_1} &= \hat{\mathbf{E}}^{p_2} \left(\hat{\mathbf{E}}^{p_1}\right)^T, \\ \mathbf{C}^{p_2} &= \hat{\mathbf{E}}^{p_1} \left(\hat{\mathbf{E}}^{p_2}\right)^T, \end{aligned} \tag{2}$$

where $\mathbf{C}^{p_1}$ and $\mathbf{C}^{p_2}$ are both $n \times n$ square matrices and they are transposed to each other. Then we generate the cross-representations by:

$$\begin{aligned} \tilde{\mathbf{E}}^{p_1} &= \mathbf{C}^{p_1} \hat{\mathbf{E}}^{p_1}, \\ \tilde{\mathbf{E}}^{p_2} &= \mathbf{C}^{p_2} \hat{\mathbf{E}}^{p_2}. \end{aligned} \tag{3}$$

Note that $\tilde{\mathbf{E}}$ can be regarded as a refined representation of $\hat{\mathbf{E}}$, which highlights the conversational interaction information in the trivial encoding results. The fact that $\tilde{\mathbf{E}}$ and $\hat{\mathbf{E}}$ share the same semantic space allows us to directly optimize their cosine distance without any additional transformations. In this circumstance, $\tilde{\mathbf{E}}$ acts as guidance for leading $\hat{\mathbf{E}}$ to be an interaction-aware self-representation, and this is why we call dial2vec works in a self-guided manner.

We further introduce w as a restriction hyper-parameter to mask the long-range semantic correlations among the utterances of p_1 and p_2. Specifically, let $\mathbf{C}[i, j]$ denotes the element in the i-th row and the j-th column of $\mathbf{C}$ in Eq. (2). $T(i)$ represents a function which returns the turn index for the i-th output embedding in $\mathbf{E}$. Then $\forall i, j \in 1, 2, \ldots, n$, $\mathbf{C}[i, j]$ is masked with zero where $abs(T(i) - T(j)) > w$, otherwise remains unchanged. Here we omit the superscript p_1 and p_2 in $\mathbf{C}$ since they are processed in the same way.

3.2.3 Aggregation

To obtain the dialogue embedding $\mathbf{e}$, we compare two aggregation strategies. In the first strategy, we directly perform average pooling across all entire output embeddings $\mathbf{E}$ (here we do not distinguish between p_1 and p_2). We further propose the interlocutor-level pooling strategy, formulated as:

$$\mathbf{e} = \sum_{r=1}^{R} \frac{\sum_{i=1}^{n} m_i^{p_r} \mathbf{h}_i}{\sum_{i=1}^{n} m_i^{p_r}}, \tag{4}$$

where $m_i^{p_r}$ is the i-th value in $\mathbf{m}^{p_r}$ and R is the number of interlocutors. We compare the results of the two strategies in Section 5.3.2.

3.2.4 Learning Objective

We adopt the NT-Xent loss proposed in (Oord et al., 2018) to train our model. Let N be the number of all training samples associated with S, which actually equals one positive sample plus the number of its corresponding negative samples. The loss l is defined as:

$$l = -\sum_{r=1}^{R} \log \frac{e^{\text{sim}(\hat{\mathbf{E}}^{p_r}, \tilde{\mathbf{E}}^{p_r})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\hat{\mathbf{E}}_j^{p_r}, \tilde{\mathbf{E}}_j^{p_r})/\tau}}, \tag{5}$$

where τ is the hyper-parameter of temperature. $\text{sim}(\cdot, \cdot)$ is defined as an average pooling operation followed by the cosine distance calculation. For all K dialogues in the dataset D, the loss $\mathcal{L}$ is given by:

$$\mathcal{L} = \frac{1}{K} \sum_{i=1}^{K} l_i. \tag{6}$$

4 Experiments Setup

4.1 Evaluation Tasks

We introduce three evaluation tasks: domain categorization, semantic relatedness, and dialogue retrieval. We categorize them into intrinsic and extrinsic tasks. The intrinsic tasks, including domain categorization and semantic relatedness, focus on assessing the overall distribution of the learned dialogue embeddings. The extrinsic task (i.e., dialogue retrieval) is more concerned with the performance of embeddings in dense retrieval scenarios.

Domain Categorization. Given a dataset of dialogues, the task is to assign each dialogue the corresponding domain label. Following Schnabel et al. (2015), we conduct this experiment as an unsupervised clustering task. All the dialogue embeddings

Datasets	Train			Dev			Test			#Domain
	#Sample	#Turn	#Word	#Sample	#Turn	#Word	#Sample	#Turn	#Word	
BiTOD	2952	19	217	70	11	109	106	10	106	6
Doc2dial	3474	11	187	661	12	182	661	12	182	4
MetalWOZ	30307	11	83	3788	11	82	3789	11	82	47
MultiWOZ	8437	13	177	1077	9	110	1084	9	110	7
Self-dialogue	19331	15	151	2416	15	151	2417	15	152	28
SGD	16142	20	199	836	14	140	1331	12	124	45

Table 1: Statistics of the six dialogue datasets used in our experiments. #Turn and #Word represent the average number of turns and words per dialogue. #Domain represents the total number of domains in the dataset.

are clustered into n categories with KMeans++ Algorithm (Arthur and Vassilvitskii, 2006), where n is the number of domains in the dataset. We adopt the purity metric in this task.

Semantic Relatedness. We pair each dialogue with a dialogue randomly selected from the same dataset and evaluate their semantic relatedness score based on their cosine similarity. The ground-truth label assigned to each dialogue pair is a binary value and is decided by whether the two dialogues share the identical domain. Following Baroni et al. (2014), we calculate Spearman's correlation between the sorted semantic relatedness scores and their corresponding ground-truth labels. This task is more stable than the domain categorization task since it gets rid of the high variance of clustering algorithms when the embedding distribution changes.

Dialogue Retrieval. Given a dialogue as a query, this task requires a model to rank all the candidates based on the cosine similarities. We use mean average precision (MAP) as the evaluation measure.

4.2 Datasets

We collect six widely-used dialogue datasets as below. We choose these datasets because they hold clear domain labels. Other datasets either provide non-semantic labels (e.g., logical labels that are less relevant to conversational semantics) (Li et al., 2017) or provide the domain labels automatically annotated by algorithms (Chen et al., 2021), thus are not suitable in our experiments. We split each dataset into training, validation, and testing sets and filter out dialogues with multiple domains in validation and test sets to fit our evaluation tasks. All domain labels are invisible to the model during the training procedure. Table 1 shows their statistics.

BiTOD (Lin et al., 2021) is a bilingual multi-domain dataset proposed for end-to-end task-oriented dialogue modeling. It provides thousands of dialogues and a large and realistic bilingual knowledge base. We use the dialogues and conduct experiments under the monolingual setting.

Doc2dial (Feng et al., 2020) includes goal-oriented dialogues that are grounded in the associated documents. We take the document topics as the domain labels of the dialogues.

MetalWOZ (Lee et al., 2019) is proposed for DSTC8, aiming at helping models more accurately predict user responses in new domains.

MultiWOZ (Eric et al., 2019) is a multi-domain dialogue dataset that poses significant challenges to task-oriented dialogue modeling due to its complexity. We use the 2.1 version in our experiments.

Self-dialogue (Fainberg et al., 2018) consists of large-scale self-dialogues with a broad set of topics. Modeling these dialogues is relatively difficult since they have more turns and topics.

SGD (Rastogi et al., 2020) is another larger-scale multi-domain dialogue dataset. We take the service field as the domain label of the dialogues.

4.3 Comparison Methods

The baseline approaches compared to our model are categorized into four groups as follows.

Non-DL Approaches. We treat a dialogue as a document and apply Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to assign each dialogue a topic. LDA is only used in the domain categorization task, since it cannot give a similarity score between two dialogues.

Embedding-based Approaches. We consider a dialogue as a sequence of words or sentences, and we obtain dialogue embeddings by combining their pre-trained embeddings. We adopt GloVe (Pennington et al., 2014) to obtain pre-trained word embeddings, SimCSE (Gao et al., 2021) to obtain pre-trained universal sentence embeddings, and DialogueCSE (Liu et al., 2021) to obtain pre-trained dialogue-based sentence embeddings. Also, we

Model	Domain Categorization							Semantic Relatedness						
	bit	**doc**	**met**	**mul**	**sel**	**sgd**	**Average**	**bit**	**doc**	**met**	**mul**	**sel**	**sgd**	**Average**
LDA	44.7	35.2	19.3	45.9	24.7	20.2	31.7	-	-	-	-	-	-	-
GloVe	64.3	54.7	40.5	79.0	35.0	51.6	54.2	34.3	25.9	15.8	38.9	15.7	27.6	26.4
Doc2Vec	82.7	70.9	43.9	86.1	40.9	63.6	64.7	43.4	24.8	14.6	30.6	10.7	26.9	25.2
SimCSE	79.3	64.7	45.1	85.5	46.8	66.7	64.7	39.5	33.2	14.9	38.1	18.3	26.9	28.5
DialogueCSE	85.8	68.4	77.5	94.9	53.2	72.1	75.3	42.4	44.5	23.9	65.2	27.6	31.7	39.2
BERT	49.1	54.0	31.6	61.3	44.4	31.3	45.3	24.3	22.4	11.6	30.5	16.7	17.9	20.6
RoBERTa	63.2	40.4	46.4	62.8	44.9	40.8	49.8	30.2	14.8	15.4	28.5	17.5	16.7	20.5
T5	78.7	55.2	67.6	89.5	43.8	69.5	67.4	38.6	28.6	20.8	42.5	20.0	29.7	30.0
TOD-BERT	75.6	63.1	**82.9**	94.3	50.0	50.3	69.4	47.0	32.6	24.3	48.9	24.6	24.8	33.7
Blender	80.9	56.4	62.3	82.4	45.4	73.1	66.8	37.0	28.1	19.9	44.4	18.3	31.1	29.8
PLATO	74.1	79.0	73.9	82.5	62.5	71.0	73.8	46.6	38.7	22.7	45.3	35.1	32.4	36.8
Dial2Vec	**90.6**	**90.2**	77.2	**96.7**	**63.1**	**86.2**	**84.0**	**68.8**	**50.7**	**24.5**	**71.0**	**37.2**	**36.9**	**48.2**

Table 2: Evaluation results of the intrinsic tasks on the six dialogue datasets, including BiTOD (**bit**), Doc2dial (**doc**), MetalWOZ (**met**), MultiWOZ (**mul**), Self-dialogue (**sel**) and SGD (**sgd**). The metrics are purity and Spearman's correlation for the two tasks respectively. All results reported are averaged across 10 independent runs to reduce the variance. Boldface and underline highlight the best and the second-best scores.

consider a dialogue as a document and embed it with Doc2Vec (Le and Mikolov, 2014).

PLMs. We consider three representative pre-trained language models, including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019a), and T5 (Raffel et al., 2020).

Conversational-PLMs. We adopt TOD-BERT (Wu et al., 2020a), Blender (Roller et al., 2021), and PLATO (Bao et al., 2020) as baselines in this group. TOD-BERT (Wu et al., 2020a) is pre-trained with nine dialogue datasets, including MultiWOZ and MetalWOZ. We adopt it as a strong baseline to compare against our model, especially on the MultiWOZ and MetalWOZ datasets. Blender (Roller et al., 2021) and PLATO (Bao et al., 2020) are pre-trained with large-scale open domain dialogue data including Twitter and Reddit (Cho et al., 2014; Zhou et al., 2018; Galley et al., 2019). We also include them as strong baselines for comparison.

For all PLMs and Conversational-PLMs, we use the average of the output embeddings from the top layer as the dialogue embedding. We do not experiment with the `[CLS]` token embedding since it may be relatively weak in representing long conversational texts.

4.4 Implement Details

Our model is implemented in PyTorch (Paszke et al., 2019). We initialize our encoder with PLATO's pre-trained parameters. During fine-tuning, we freeze the bottom 6 layers of the encoder to avoid the catastrophic forgetting problem. The maximum sequence length is limited to 512 but is sufficient for most dialogues in our experiments. The temperature τ and the window size w are set to 0.2 and 10 respectively, since such configuration performs best across all datasets. We optimize the model parameters with Adam optimizer (Kingma and Ba, 2015), using a learning rate of 1e-5 and a batch size of 5 per GPU. All models are trained with 4 NVIDIA Tesla V100 GPUs.

5 Experimental Results

5.1 Intrinsic Task

Table 2 shows the experimental results of the intrinsic tasks. For each task, dial2vec achieves on average 8.7 and 9.0 absolute improvements in terms of purity and Spearman's correlation against the strongest baseline DialogueCSE. We attribute the strong performance to the introduction of the interlocutor-level conversational interaction information in learning dialogue embeddings.

Conversational-PLMs show overwhelming superiority over PLMs, indicating that pre-training with conversational data plays a key role in producing better dialogue embeddings. The phenomenon that TOD-BERT achieves very competitive results against dial2vec on MultiWOZ and MetalWOZ also confirms this fact. Even so, dial2vec easily bridges or reverses the gaps between PLATO and TOD-BERT on both datasets, demonstrating its superiority in exploiting conversational information.

PLATO generally performs better than TOD-BERT and Blender. We hypothesize that the turn and role embeddings also play a crucial role in our task. To verify this, we employ BERT as the en-

coder and train dial2vec(BERT)[1] under the same setting as dial2vec(PLATO). However, the results on the three tasks degrade rapidly after reaching the best performances. We believe that under such a setting, the inputs provide insufficient information for the model to maintain turn-aware and role-aware semantic information in the encoding outputs, making the training not robust.

The embedding-based methods suffer from poor performance since they ignore the weights when combining word and sentence embeddings. Among them, SimCSE releases the inherent representation capability of PLMs by introducing the twice-dropout operation in fine-tuning, achieving better results than GloVe and Doc2Vec. But since such an operation is generic and orthogonal to our work, we do not incorporate it into our model. Particularly, DialogueCSE yields superior results compared with other embedding-based methods and even shows competitive performances against our model. This is reasonable since it is the only baseline in this group that leverages conversational interactions to learn sentence embeddings. However, its performance is still unsatisfactory since it performs sentence-level instead of interlocutor-level interactions and fails to model the entire dialogue.

5.2 Extrinsic Task

Table 3 shows dial2vec's performances on the dialogue retrieval task. Compared to the experiment results on the intrinsic tasks, dial2vec achieves more significant improvements on all datasets. We attribute it to dial2vec's capability of understanding fine-grained conversational semantics. Since dial2vec is forced to distinguish the positive samples composed of the exact matching question-answers from the negative ones, the semantic information it learned is more fine-grained than tasks trained with only domain labels. Such characteristic makes it adept at ranking semantically similar candidates, resulting in better performance on the MAP metric.

5.3 Analysis

To further investigate the property of our model, we adopt PLATO as the baseline to conduct experiments with the single interlocutor's embeddings, the aggregation strategies, and the embedding distributions. We report the average results across all datasets.

[1]In this case, `[EOU]` tokens are inserted into the sequence to denote the separation of different turns.

Model	Dialogue Retrieval						
	bit	doc	met	mul	sel	sgd	AVG
GloVe	63.8	49.1	29.4	65.9	24.1	52.6	47.5
Doc2Vec	67.7	43.7	15.1	50.9	16.3	43.1	39.5
SimCSE	62.5	52.5	23.8	62.1	27.0	44.8	45.5
DialogueCSE	72.9	58.2	66.7	82.9	34.5	62.5	62.9
BERT	52.4	44.8	17.0	56.4	25.8	26.0	37.1
RoBERTa	62.2	40.6	30.4	57.4	25.5	35.0	41.9
T5	67.3	49.9	43.9	69.7	27.8	53.8	52.1
TOD-BERT	73.2	53.0	65.7	84.2	33.0	45.3	59.1
Blender	69.1	50.1	44.6	70.1	25.4	63.0	53.7
PLATO	71.6	59.7	54.5	68.7	45.9	63.2	60.6
Dial2Vec	**94.4**	**69.4**	**68.0**	**96.4**	**49.4**	**82.8**	**76.7**

Table 3: Evaluation results of the dialogue retrieval task. We use the mean average precision (MAP) as the evaluation metric. Boldface and underline highlight the best and the second-best scores.

5.3.1 Single Interlocutor's Embeddings

Intuitively, each interlocutor holds their own unilateral information of the dialogue. However, such interaction-free information usually contains noises or overlaps with that from other interlocutors. Table 4 shows the experiment results of PLATO and dial2vec. We find that the PLATO's embeddings for individual interlocutors usually perform close to or even better than the aggregated results. As a comparison, dial2vec yields significantly better embeddings for both interlocutors, and achieves further improvements when aggregating them. We conclude that under the guidance of the conversational interactions, dial2vec eliminates the interlocutor-level interaction-free information and highlights the interaction-aware information, thus achieving better performances.

Model	Interlocutor	Purity	Spearman	MAP
PLATO	$p1$	71.78	36.23	59.38
	$p2$	75.29	36.67	60.73
	$p1 + p1$	73.83	36.80	60.60
	diff	-1.46	0.13	-0.13
Dial2Vec	$p1$	83.27	47.82	75.72
	$p2$	82.21	46.74	74.56
	$p1 + p1$	**83.97**	**48.19**	**76.73**
	diff	0.70	0.37	1.01

Table 4: Performances of dialogue embeddings for each interlocutor. $p1$ represents that we use the embeddings from the interlocutor who starts the conversation, and $p2$ represents the opposite. $p1 + p2$ represents the aggregated results. $diff$ shows improvements of the aggregated results over the best single interlocutor's results. Boldface represents the best scores among the models.

5.3.2 Aggregation Strategies

As described in Section 3.2.3, we experiment with two aggregation strategies: average pooling and interlocutor-level pooling. For average pooling, we average all the output embeddings as the final dialogue embedding, while for interlocutor-level pooling, we sum the average pooling results of the output embeddings corresponding to each interlocutor.

Model	Purity	Spearman	MAP
Dial2Vec$_{avg}$	83.69	47.93	76.39
Dial2Vec$_{int}$	**83.97**	**48.19**	**76.73**
diff	+0.28	+0.26	+0.34

Table 5: Comparison between the average-pooling (denoted as avg) and interlocutor-level pooling (denoted as int) strategies. Boldface highlights the best scores.

Table 5 shows the results for the two strategies on all datasets. The interlocutor-level pooling strategy performs consistently better than the average pooling strategy. We hold that the interlocutor-level pooling strategy acts as a normalization operation that balances the weight of semantic information from different interlocutors.

5.3.3 Alignment and Uniformity Analysis

Inspired by Wang and Isola (2020), we employ the alignment and uniformity metrics to study the variation of dialogue embedding distribution during training. Given a set of data pairs and their corresponding labels, the alignment metric is calculated as the expected value of Euclidean distances of each positive data pair, formulated as:

$$\ell_{alignment} \triangleq \mathop{\mathbb{E}}_{x,x^+\sim p_{pos}} \|f(x) - f(x^+)\|^2. \quad (7)$$

The alignment metric is suitable for tasks such as Gao et al. (2021) since the positive pairs are encoded from a unique text. However, in our scenario, positive pairs generated from two different dialogues are only expected to have closer distances than negative pairs. Thus, we revise the Eq. (7) to be:

$$\ell_{adj\_alignment} \triangleq \mathop{\mathbb{E}}_{x,x^+\sim p_{pos}} \|f(x) - f(x^+)\|^2, - \mathop{\mathbb{E}}_{x,x^-\sim p_{neg}} \|f(x) - f(x^-)\|^2. \quad (8)$$

We name $\ell_{adj\_alignment}$ as the adjusted alignment metric.

The uniformity metric is defined to measure how close the embeddings are to the uniform distribution:

$$\ell_{uniformity} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d}{\sim} p_{data}} e^{-2\|f(x)-f(y)\|^2}, \quad (9)$$

where p_{data} denotes the data distribution.

Figure 3 shows how the adjusted alignment and uniformity vary with the training iterations. 'Start' marks the results at the very beginning of training, which also stands for the vanilla PLATO's performances. As we can see, the two metrics decrease rapidly in the first few iterations. We believe dial2vec learns discriminative embeddings by pushing the embeddings for the two interlocutors in the negative samples away from each other in this stage. Since the dialogue embeddings are spread out over the unit hypersphere, both metrics decrease. As the training proceeds, the model learns the fine-grained informative dialogue embeddings from the positive samples. This makes the dialogues with similar semantics close to each other, causing the uniformity to increase again. The two metrics finally converge to the values much better than the start points, showing that dial2vec learns both informative and discriminative embeddings.

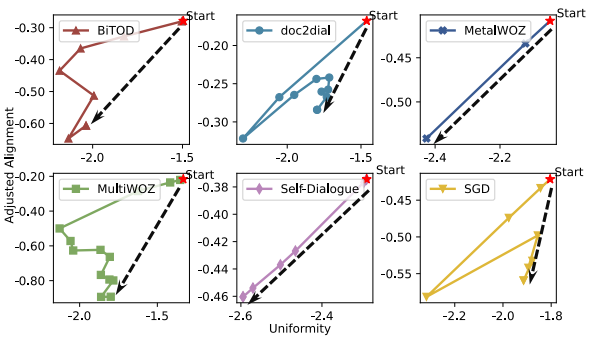

Figure 3: The scatter plot of $\ell_{adj\_alignment}$-$\ell_{uniformity}$ on the six testing sets. We plot the two metrics after every evaluation. For both metrics, lower values represent better distributions.

6 Conclusion

In this paper, we formally introduce the task of learning unsupervised dialogue embeddings and propose dial2vec to solve this task. We introduce a self-guided mechanism that leverages the conversational interactions to guide the learning of the embeddings for both interlocutors and propose the interlocutor-level strategy to aggregate them. We further release a benchmark consisting of six

widely-used dialogue datasets and three tasks designed based on the domain labels. Our model achieves superior performances on all datasets across the three tasks, and further analysis shows that the dialogue embeddings learned by our model are more informative and discriminative than the baselines. We believe there is still much room for improvement to generate satisfactory dialogue embedding.

7 Limitations

Our work has two limitations. First, although dial2vec is designed to be able to scale to multi-party conversations, we did not conduct such experiments due to the lack of a suitable multi-party evaluation dataset. As annotating for a multi-party dialogue dataset is indeed complicated, we leave it to future work. Besides, dial2vec still performs unsatisfactorily when employing BERT-like encoders. Although they also achieve very competitive results, we believe that a robust training procedure is more important since we do not know when to stop training under the unsupervised setting in practice. Dial2vec should be further improved to better adapt to the multiple formats of input embeddings.

References

David Arthur and Sergei Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Minmin Chen. 2017. Efficient vector representation for documents through corruption. *arXiv preprint arXiv:1707.02377*.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. Dialogsum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. *arXiv preprint arXiv:2204.10298*.

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

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*.

Joachim Fainberg, Ben Krause, Mihai Dobre, Marco Damonte, Emmanuel Kahembwe, Daniel Duma, Bonnie Webber, and Federico Fancellu. 2018. Talking to myself: self-dialogues as data for conversational agents. *arXiv preprint arXiv:1809.06641*.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

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

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022a. Space-2: Tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. *arXiv preprint arXiv:2209.06638*.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022b. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022c. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Binyuan Hui, Ruiying Geng, Qiyu Ren, Binhua Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, Pengfei Zhu, and Xiaodan Zhu. 2021. Dynamic hybrid relation exploration network for cross-domain context-dependent semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13116–13124.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. *arXiv preprint arXiv:2106.07345*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Tassilo Klein and Moin Nabi. 2022. Scd: Self-contrastive decorrelation of sentence embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 394–400.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Sungjin Lee, Hannes Schulz, Adam Atkinson, Jianfeng Gao, Kaheer Suleman, Layla El Asri, Mahmoud Adada, Minlie Huang, Shikhar Sharma, Wendy Tay, and Xiujun Li. 2019. Multi-domain task-completion dialog challenge. In *Dialog System Technology Challenges 8*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. Dialoguecse: Dialogue-based contrastive learning of sentence embeddings. *arXiv preprint arXiv:2109.12599*.

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

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019b. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Dongsheng Luo, Wei Cheng, Jingchao Ni, Wenchao Yu, Xuchao Zhang, Bo Zong, Yanchi Liu, Zhengzhang Chen, Dongjin Song, Haifeng Chen, et al. 2021. Unsupervised document embedding via contrastive augmentation. *arXiv preprint arXiv:2103.14542*.

Chenxu Lv, Hengtong Lu, Shuyu Lei, Huixing Jiang, Wei Wu, Caixia Yuan, and Xiaojie Wang. 2021. Task-oriented clustering for dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4338–4347.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. Ease: Entity-aware contrastive learning of sentence embedding. *arXiv preprint arXiv:2205.04260*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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)*, pages 1532–1543.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307.

Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng Wang, and Lintao Zhang. 2018. Auto-dialabel: Labeling dialogue data with unsupervised learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 684–689.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9177–9184.

Lihan Wang, Bowen Qin, Binyuan Hui, Bowen Li, Min Yang, Bailin Wang, Binhua Li, Jian Sun, Fei Huang, Luo Si, et al. 2022. Proton: Probing schema linking information from pre-trained language models for text-to-sql parsing. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1889–1898.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Chien-Sheng Wu, Steven CH Hoi, Richard Socher, and Caiming Xiong. 2020a. 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)*, pages 917–929.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Shuai Zhang, Wang Lijie, Xinyan Xiao, and Hua Wu. 2022. Syntax-guided contrastive learning for pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2430–2440, Dublin, Ireland. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130.