Dialog2Flow: Pre-training Action-Driven Soft Contrastive Learning Embeddings for Automatic Dialog Flow Extraction

Anonymous ACL submission

Abstract

Efficiently deriving structured workflows from unannotated dialogs remains an underexplored and formidable challenge in computational linguistics. Automating this process could significantly accelerate the manual design of workflows in new domains and enable the grounding of large language models in domain-specific flowcharts, enhancing transparency and controllability. In this paper, we introduce Dialog2Flow (D2F) embeddings, which differ from conventional sentence embeddings by mapping utterances to a latent space where they are grouped according to their communicative and informative functions (i.e., the actions they represent). D2F allows for modeling dialogs as continuous trajectories in a latent space with distinct action-related regions. By clustering D2F embeddings, the latent space is quantized, and dialogs can be converted into sequences of region/action IDs, facilitating the extraction of the underlying workflow. To pre-train D2F, we build a comprehensive dataset by unifying twenty task-oriented dialog datasets with normalized per-turn action annotations. We also introduce a novel soft contrastive loss that leverages the semantic information of these actions to guide the representation learning process, showing superior performance compared to standard supervised contrastive loss. Evaluation against various sentence embeddings, including dialog-specific ones, demonstrates that D2F yields superior qualitative and quantitative results across diverse domains.[1]

1 Introduction

Conversational AI has seen significant advancements, especially with the rise of Large Language Models (LLMs) (Bubeck et al., 2023; Lu et al., 2022; Hendrycks et al., 2021a,b; Cobbe et al., 2021). Dialog modeling can be divided into open-domain dialogs and task-oriented dialogs (TOD), with the latter focusing on helping users achieve specific tasks (Jurafsky, 2006). In TOD, structured workflows guide agents in assisting users effectively. This paper explores the underexplored terrain of automatically extracting such workflow from a collection of conversations.

User: i'm looking for the transplant unit department please
Action: **INFORM** DEPARTMENT

System: okay the transfer unit department give me a second let me look okay yes i found the transplant unit department can i help
Action: **REQMORE**

User: may you please provide me with the phone number please
Action: **REQUEST** PHONE

System: get no problem okay so the number is 1223217711
Action: **INFORM** PHONE

User: okay um just repeat it it's 1 2 2 3 2 1 7 1 1
Action: **CONFIRM** PHONE

System: okay thank you very much
Action: **THANK_YOU**

Figure 1: Example segment of the dialog `SNG1533` from the `hospital` domain of the SpokenWOZ dataset. Actions are defined by concatenating the dialog act label (in bold) with the slot label(s) associated to each utterance.

Extracting workflows automatically is crucial for enhancing dialog system design, discourse analysis, data augmentation (Qiu et al., 2022), and training human agents (Sohn et al., 2023). Additionally, it can ground LLMs in domain-specific workflows, improving transparency and control (Raghu et al., 2021; Chen et al., 2024). Recent works have attempted to induce structural representations from dialogs using either ground truth annotation or *ad hoc* methods (Hattami et al., 2023; Qiu et al., 2022, 2020), we believe that models specifically pre-trained for this purpose could significantly advance the field. Instead of pre-training dialog state encoders, we focus on pre-training utterance encoders in a workflow-related manner. By focusing on ut-

[1] *(Github and HuggingFace links removed for review).*

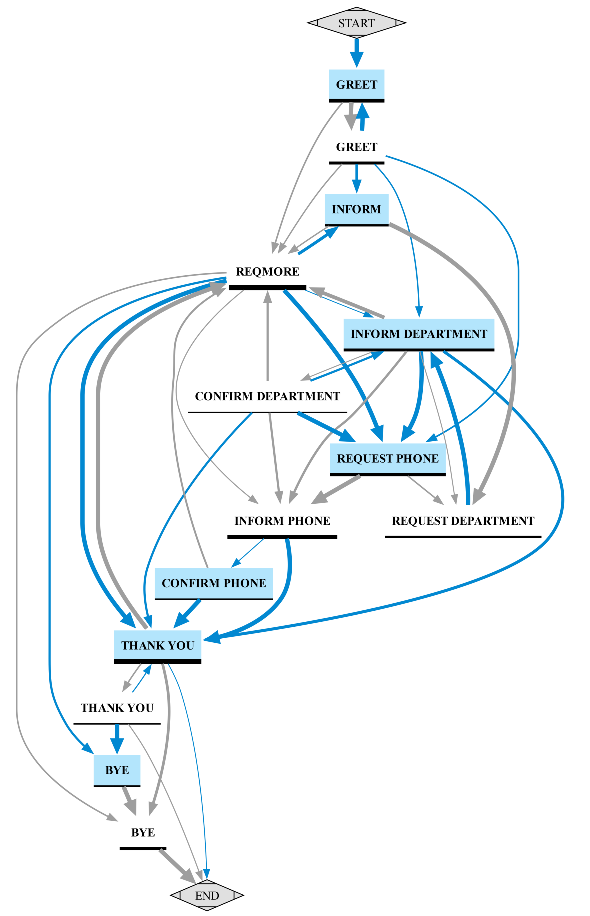


Figure 2: Directed graph representing the `hospital` domain workflow obtained from all the `hospital` dialogs in the SpokenWOZ dataset. Nodes correspond to individual actions. The width of edges and the underline thickness of nodes indicate their frequency. User actions are colored to distinguish them from system actions.

terances, we focus on how to convert sequences of utterances into "meaningful" trajectories in a latent space, disentangling them from how they are effectively condensed to task-dependent dialog states.

In TOD, *dialog acts* and *slots* are key concepts (Jurafsky, 2006). Dialog acts denote the communicative intent, while slots are pieces of task-specific information. A *dialog action* includes both the dialog act and slots. Actions allow us to transform dialogs into sequences of canonical steps carrying both their communicative and informative functions (Figure 1). Thus, aggregating sequences from multiple dialogs can reveal a common workflow (Figure 2). The main contributions of this work can be summarized as follows: (a) consolidating twenty task-oriented dialog datasets to create the largest dataset with standardized action annotations; (b) introducing a soft contrastive loss leveraging the semantic information of actions to guide the representation learning process, showing superior performance compared to standard supervised contrastive loss; and (c) introducing and releasing Dialog2Flow (D2F), to the best of our knowledge, the first utterance embedding encoder pre-trained specifically for dialog flow extraction.

2 Related Work

Sentence Embeddings Transformer-based encoders like Universal Sentence Encoder (Cer et al., 2018) and Sentence-BERT (Reimers and Gurevych, 2019) outperformed RNN-based ones such as Skip-Thought (Kiros et al., 2015) and InferSent (Conneau et al., 2017). These models use a *pooling strategy* (e.g., mean pooling, `[CLS]` token) to obtain a single sentence embedding optimized for semantic similarity. However, specific domains require different similarity notions. For task-oriented dialogs, TOD-BERT (Wu et al., 2020) and Dialog Sentence Embedding (DSE) (Zhou et al., 2022) show that conversation-based similarity outperforms semantic similarity across TOD tasks. Likewise, we hypothesize that action-based similarity can yield meaningful workflow-related sentence embeddings.

Contrastive Learning Contrastive learning has achieved success in representation learning for both images (Chen et al., 2020; He et al., 2020; Henaff, 2020; Tian et al., 2020; Chen et al., 2020; Hjelm et al., 2019) and text (Zhou et al., 2022; Zhang et al., 2022, 2021; Gao et al., 2021; Wu et al., 2020). It learns a representation space where similar instances cluster together and dissimilar instances are separated. More precisely, given an *anchor* with *positive* and *negative* counterparts, the goal is to minimize the distance between anchor-positive pairs while maximizing the distance between anchor-negative pairs. Negatives are typically obtained through in-batch negative sampling, where positives from different anchors in the mini-batch are used as negatives.

3 Method

3.1 Representation Learning Framework

Following common practices (Zhou et al., 2022; Chen et al., 2020; Tian et al., 2020; Khosla et al., 2020), the main components of our framework are:

- **Encoder**, $f(\cdot) \in \mathbb{R}^n$, which maps x to a representation vector, $\mathbf{x} = f(x)$. Following Sentence-BERT (Reimers and Gurevych, 2019) and DSE (Reimers and Gurevych, 2019), $f(\cdot)$ consists of a BERT-based encoder with mean pooling strategy trained as a bi-encoder with shared weights (siamese network).

• **Contrastive head**, $g(\cdot) \in \mathbb{R}^d$, used during training to map representations $\mathbf{x}$ to the space where contrastive loss is applied. Following Chen et al. (2020) and DSE, we instantiate $g(\cdot)$ as the multi-layer perceptron with a single hidden layer $\mathbf{z} = g(\mathbf{x}) = \mathrm{ReLU}(\mathbf{x} \cdot W_1)W_2$ where $W_1 \in \mathbb{R}^{n\times n}$ and $W_2 \in \mathbb{R}^{n\times d}$.

• **Similarity measure**, $sim(\mathbf{u}, \mathbf{v})$, used to learn the representation is cosine similarity. Thus, similarity is then measured only by the angle between $\mathbf{u}$ and $\mathbf{v}$, making our latent space geometrically a unit hypersphere. Hence, in this study, we treat similarity and alignment interchangeably. Additionally, we assume $f(\cdot)$ and $g(\cdot)$ vectors are L2-normalized, leading to $sim(\mathbf{u}, \mathbf{v}) = cos(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$.

3.1.1 Supervised Contrastive Loss

For a batch of N randomly sampled anchor, positive, and label triples, $B = \{(x_i, x_i^+, y_i)\}_{i=1}^N$, the supervised contrastive loss (Khosla et al., 2020), for each i-th triplet (x_i, x_i^+, y_i) is defined as:

$$\ell_i^{sup} = -\sum_{j\in\mathcal{P}_i} \frac{1}{|\mathcal{P}_i|} \log \frac{e^{\mathbf{z}_i \cdot \mathbf{z}_j^+/\tau}}{\sum_{k=1}^N e^{\mathbf{z}_i \cdot \mathbf{z}_k^+/\tau}} \qquad (1)$$

where $\mathcal{P}_i = \{j \mid y_i = y_j\}$ is the set of indexes of all the samples with the same label as the i-th sample in the batch, and τ is the softmax temperature parameter that controls how soft/strongly positive pairs are pulled together and negative pairs pushed apart in the embedding space.[2] The final loss is computed across all the N pairs in the mini-batch as $\mathcal{L}^{sup} = \frac{1}{N}\sum_{i=1}^N \ell_i^{sup}$.

3.1.2 Supervised *Soft* Contrastive Loss

Let $\delta(y_i, y_j)$ be a semantic similarity measure between both y_i, y_j labels, we define our soft contrastive loss as follows:

$$\ell_i^{soft} = -\sum_{j=1}^N \frac{e^{\delta(y_i,y_j)/\tau'}}{\sum_{k=1}^N e^{\frac{\delta(y_i,y_k)}{\tau'}}} \log \frac{e^{\mathbf{z}_i \cdot \mathbf{z}_j^+/\tau}}{\sum_{k=1}^N e^{\frac{\mathbf{z}_i \cdot \mathbf{z}_k^+}{\tau}}}$$

where τ' is the temperature parameter to control the "softness" of the negative labels (impact analysis in Appendix E). Unlike Equation 3, this loss encourages the encoder to *separate anchors and negatives in proportion to the semantic similarity of their labels* (details in Appendix D). Finally, the mini-batch loss $\mathcal{L}^{soft}$ is computed as in $\mathcal{L}^{sup}$.

[2] The lower τ, the sharper the softmax output distribution and the stronger the push/pull factor.

Dataset	#U	#D	#DA	#S
ABCD (Chen et al., 2021)	20.4K	10	0	10
BiTOD (Lin et al., 2021)	72.5K	6	13	33
Disambiguation (Qian et al., 2022)	114.3K	8	9	28
DSTC2-Clean (Mrkšić et al., 2017)	25K	1	2	8
FRAMES (El Asri et al., 2017)	20K	1	21	46
GECOR (Quan et al., 2019)	2.5K	1	2	10
HDSA-Dialog (Chen et al., 2019)	91.9K	8	6	24
KETOD (Chen et al., 2022)	107.7K	20	15	182
MS-DC (Li et al., 2018)	71.9K	3	11	56
MulDoGO (Peskov et al., 2019)	74.8K	6	0	63
MultiWOZ2.1 (Eric et al., 2020)	108.3K	8	9	27
MultiWOZ2.2 (Zang et al., 2020)	55.9K	8	2	26
SGD (Rastogi et al., 2020)	479.5K	20	15	184
Taskmaster1 (Byrne et al., 2019)	30.7K	6	1	59
Taskmaster2 (Byrne et al., 2019)	147K	11	1	117
Taskmaster3 (Byrne et al., 2019)	589.7K	1	1	21
WOZ2.0 (Mrkšić et al., 2017)	4.4K	1	2	10
SimJointMovie (Shah et al., 2018)	7.2K	1	14	5
SimJointRestaurant (Shah et al., 2018)	20K	1	15	9
SimJointGEN (Zhang et al., 2024)	1.3M	1	16	5
Total	**3.4M**	**52**	**44**	**524**

Table 1: Details of used TOD datasets, including the number of utterances (#U), unique domains (#D), dialog act labels (#DA), and slot labels (#S).

3.2 Training Targets

We experiment with four types of training targets, which differ in whether the dialog action label is used as-is or decomposed into dialog act and slot labels, and the type of contrastive loss used. Specifically, we have the following two targets using the proposed soft contrastive loss:

• **D2F**$_{single}$: $\mathcal{L} = \mathcal{L}^{soft}_{\textbf{act+slots}}$

• **D2F**$_{joint}$: $\mathcal{L} = \mathcal{L}^{soft}_{\textbf{act}} + \mathcal{L}^{soft}_{\textbf{slots}}$

and the two corresponding targets using the default supervised contrastive loss:

• **D2F-Hard**$_{single}$: $\mathcal{L} = \mathcal{L}^{sup}_{\textbf{act+slots}}$

• **D2F-Hard**$_{joint}$: $\mathcal{L} = \mathcal{L}^{sup}_{\textbf{act}} + \mathcal{L}^{sup}_{\textbf{slots}}$

The subscript in bold indicates the type of label used to compute the loss, either the dialog action as a single label (*act+slots*), or the dialog act and slots separately. In the case of the joint loss, separate contrastive heads $g(\cdot)$ are employed.

4 Training Corpus

We identified and collected 20 TOD datasets from which we could extract dialog act and/or slot annotations, as summarized in Table 1. We then manually inspected each dataset to locate and extract the necessary annotations, manually standardizing domain names and dialog act labels across datasets. Finally, we unified all datasets under a consistent format, incorporating per-turn dialog act and slot annotations. The resulting unified TOD dataset comprises 3.4 million utterances annotated with 18 standardized dialog acts, 524 unique slot labels, and 3,982 unique action labels (dialog act + slots) spanning across 52 different domains (details in Appendix A).

5 Experimental Setup

For training D2F we mostly follow the experimental setup of DSE (Zhou et al., 2022) and TOD-BERT (Wu et al., 2020), using BERT$_{base}$ as the backbone model for the encoder to report results in the main text. Additional configurations are reported in the ablation study (Appendix C) while implementation details are given in Appendix B.

5.1 Baselines

General sentence embeddings. • **`GloVe`**: the average of GloVe embeddings (Pennington et al., 2014). • **`BERT`**: the vanilla BERT$_{base}$ model with mean pooling strategy, corresponding to our untrained encoder. • **`Sentence-BERT`**: the model with the best average performance reported among all Sentence-BERT pre-trained models, namely the `all-mpnet-base-v2` model pre-trained using MPNet (Song et al., 2020) and further fine-tuned on a 1 billion sentence pairs dataset. • **`GTR-T5`**: the Generalizable T5-based dense Retriever (Ni et al., 2022) pre-trained on a 2 billion web question-answer pairs dataset, outperforming previous sparse and dense retrievers on the BEIR benchmark (Thakur et al., 2021).

Dialog sentence embeddings. • **`TOD-BERT`**: the `TOD-BERT-jnt` model reported in Wu et al. (2020) pre-trained to optimize a contrastive response selection objective by treating utterances and their dialog context as positive pairs. The pre-training data is the combination of 9 publicly available task-oriented datasets around 1.4 million total utterances across 60 domains. • **`DSE`**: pre-trained on the same dataset as TOD-BERT, DSE learns utterance embeddings by simply taking consecutive utterances of the same dialog as positive pairs for contrastive learning. DSE has shown to achieve better representation capability than the other dialog and general sentence embeddings on TOD downstream tasks (Gung et al., 2023; Zhou et al., 2022). • **`SBD-BERT`**: the `TOD-BERT-SBD`$_{MWOZ}$ model reported in Qiu et al. (2022) in which utterances are represented as the mean pooling of the tokens that are part of the slots of the utterance, as identified by a Slot Boundary Detection (SBD) model trained on the original MultiWOZ dataset (Budzianowski et al., 2018). • **`DialogGPT`**: following TOD-BERT and DSE, we also report results with DialogGPT (Zhang et al., 2020) using the mean pooling of its hidden states as the sentence representation.

5.2 Evaluation Data

Most of the TOD datasets are constructed solely based on written texts, which may not accurately reflect the nuances of real-world spoken conversations, potentially leading to a gap between academic research and real-world spoken TOD scenarios. Therefore, we evaluate our performance not only on a subset of our unified TOD dataset but also on SpokenWOZ (Si et al., 2023), the first large-scale human-to-human speech-text dataset for TOD designed to address this limitation. More precisely, we use the following two evaluation sets:

• **Unified TOD evaluation set:** 26,910 utterances with 1,794 unique *action* labels (dialog act + slots) extracted from the training data. These utterances were extracted by sampling and removing 15 utterances for each *action* label with more than 100 utterances in the training data.

• **SpokenWOZ:** 31,303 utterances with 427 unique *action* labels corresponding to all the 1,710 single domain conversations in SpokenWOZ. We are only using complete single-domain conversations so that we can also use them later to induce the domain-specific workflow for each of the 7 domains in SpokenWOZ.[3]

6 Similarity-based Evaluation

Before the dialog flow-based evaluation, we assess the quality of the representation space geometry through the similarity of the embeddings representing different *actions*. We use the following methods as quality proxies:

• **Anisotropy.** Following Jiang et al. (2022); Ethayarajh (2019), we measure the anisotropy of a set of embeddings as the average cosine (absolute) similarity among all embeddings in the set.[4] Ideally,

[3] There are no single-domain calls for the `profile` domain so it is not included.

[4] $\frac{1}{n^2-n}\left|\sum_i \sum_{j\neq i} \cos(\mathbf{x}_i, \mathbf{x}_j)\right|$ for given $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$

embeddings of the same *action* should be similar (high intra-action anisotropy) while being dissimilar to those of other actions (low inter-action anisotropy). We report the average intra- and inter-action anisotropy across all actions.

• **Similarity-based few-shot classification.** We use Prototypical Networks (Snell et al., 2017) to perform a similarity-based classification. A prototype embedding for each *action* is calculated by averaging k of its embeddings (k-shot). All other embeddings are then classified based on the closest prototype embedding. We report the *macro averaged* F_1 score and *Accuracy* for $k = 1$ and $k = 5$ (i.e., 1-shot and 5-shot classification).

• **Ranking.** For each action, we randomly select one utterance as the query and retrieve the top-k closest embeddings, creating a ranking with their actions. Ideally, the top-k retrieved embeddings should predominantly correspond to the same *action* as the query, thus ranked first. We report *Normalized Discounted Cumulative Gain* (nDCG@10), averaged over all actions.

6.1 Similarity-based Results

Tables 2 and 3 present the similarity-based classification and anisotropy results on the unified TOD evaluation set and SpokenWOZ, respectively. Results are averaged over 1,794 and 427 different action labels for both datasets, respectively. For classification results, we report the mean and standard deviation from 10 repetitions, each sampling different embeddings for the 1-shot and 5-shot prototypes. All D2F variants outperform baselines in all metrics, indicating a representation space where embeddings are clustered by their actions. However, baseline results provide a proxy for the quality of their representation spaces for our end goal. For instance, general embeddings, which cluster by semantic similarity, are outperformed by DSE, which clusters by utterance context in TOD dialogs. Notably, D2F embeddings trained with the proposed soft contrastive loss outperform D2F-Hard embeddings trained with the vanilla supervised contrastive loss, especially in the 1-shot setting. In Table 3, the difference among the various embeddings narrows, and standard deviations increase significantly compared to Table 2. This indicates that results vary considerably depending on the sampled prototypes, suggesting that the SpokenWOZ data is noisier than the unified TOD evaluation set. This is expected as SpokenWOZ utterances were obtained by an ASR model from real-world human-to-human spoken TOD conversations, thus affected by ASR noise and various linguistic phenomena such as back-channels, disfluencies, and incomplete utterances.[5]

Classification results provide a local view of the representation space quality around the different sampled prototypes. Actions spread into multiple sub-clusters could still yield good classification results. Thus, we also consider anisotropy results for a more global view of the representation space quality. Among the baselines, TOD-BERT has the highest intra-action anisotropy but also the highest inter-action value, meaning different actions are more similar than embeddings of the same action on average (negative Δ values!). Sentence-BERT has the lowest inter-action anisotropy, indicating different actions are the most dissimilar, although embeddings of the same action are less similar ($\Delta = 0.094$) compared to DSE ($\Delta = 0.108$) in Table 2. D2F embeddings exhibit the best anisotropy values, with a similarity difference between intra- and inter-action embeddings of 0.597 and 0.451, or 0.193 and 0.103 on SpokenWOZ, for single and joint targets, respectively, roughly doubling their D2F-Hard counterparts. This improvement could be attributed to a better overall arrangement of the embeddings, guided by the semantics of the actions during the representation learning process. For instance, Figure 3 shows the projection of the embeddings onto the unit sphere for a subset of six related actions.[6] Sentence-BERT clusters embeddings into roughly two main semantic groups, with price-related actions on top and others at the bottom. D2F-Hard correctly clusters embeddings of the same action together while maintaining separation among centroids of different actions. However, the arrangement among different clusters is better in D2F, guided by action semantics –namely, all clusters are adjacent, with •`[request` **`price`**`]` next to •`[inform` **`price`**`]`; •`[inform` **`name price`**`]` between •`[inform` **`name`**`]` and •`[inform` **`price`**`]`; and •`[inform` **`name price area`**`]` between •`[inform` **`name price`**`]` and •`[inform` **`name area`**`]`.

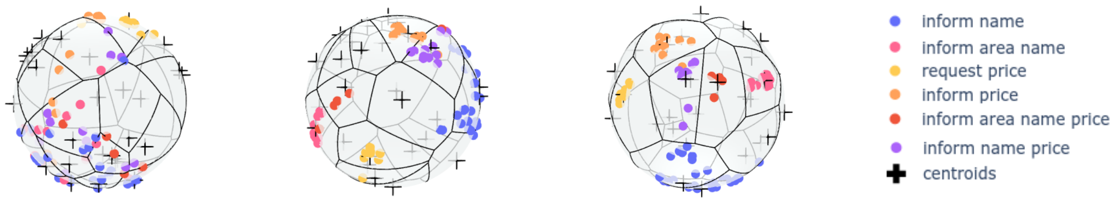
Finally, Table 4 presents the ranking-based results on both evaluation sets. We report the mean and standard deviation from 10 repetitions, each

[5]SpokenWOZ authors conducted experiments using newly proposed LLMs and dual-modal models, showing that current models still have substantial room for improvement on this realistic spoken dataset (Si et al., 2023).

[6]The original n-1 manifold in which utterances are embedded correspond to the unit hyper-sphere, thus, the unit sphere provides a more truthful visualization than a 2D plane.

Embeddings	F_1 score 1-shot	F_1 score 5-shot	Accuracy 1-shot	Accuracy 5-shot	Anisotropy intra(↑)	Anisotropy inter(↓)	Anisotropy Δ (↑)
GloVe	23.24 ± 0.87	24.45 ± 0.94	26.04 ± 0.81	30.01 ± 0.86	0.674	0.633	0.041
BERT	23.85 ± 0.47	28.22 ± 0.60	26.32 ± 0.62	32.92 ± 0.38	0.737	0.781	-0.044
Sentence-BERT	27.86 ± 0.93	33.30 ± 0.68	30.55 ± 0.82	38.22 ± 0.46	0.527	**0.433**	0.094
GTR-T5	30.86 ± 0.39	38.38 ± 0.64	33.34 ± 0.29	42.96 ± 0.60	0.694	0.706	-0.012
DSE	**35.43** ± 0.96	**42.21** ± 0.90	**38.12** ± 0.77	**46.85** ± 0.79	0.649	0.541	**0.108**
TOD-BERT	27.58 ± 0.92	33.35 ± 0.58	29.63 ± 1.06	36.88 ± 0.87	**0.840**	0.864	-0.024
DialoGPT	25.86 ± 0.34	31.34 ± 0.73	28.24 ± 0.53	36.15 ± 0.83	0.734	0.758	-0.024
SBD-BERT	24.31 ± 0.95	27.71 ± 0.38	26.40 ± 0.96	31.53 ± 0.44	0.687	0.604	0.083
D2F-Hard$_{single}$	**58.84** ± 0.62	**67.82** ± 0.52	**61.52** ± 0.54	**70.69** ± 0.43	**0.646**	**0.313**	**0.332**
D2F-Hard$_{joint}$	56.25 ± 1.16	66.22 ± 0.62	58.98 ± 1.08	69.23 ± 0.48	0.629	0.399	0.230
D2F$_{single}$	**<u>65.36</u>** ± 0.91	70.89 ± 0.30	**<u>68.06</u>** ± 0.87	**<u>74.15</u>** ± 0.40	**0.782**	**<u>0.186</u>**	**<u>0.597</u>**
D2F$_{joint}$	63.70 ± 1.35	**<u>70.94</u>** ± 0.41	66.53 ± 1.15	74.03 ± 0.31	0.741	0.289	0.451

Table 2: Similarity-based few-shot classification results on our unified TOD evaluation set. The intra- and inter-action anisotropy are also provided along their difference (Δ). **Bold** indicates the best values in each group while **<u>underlined</u>** the global best.



(a) **Sentence-BERT** (b) **D2F-Hard**$_{joint}$ (c) **D2F**$_{joint}$

Figure 3: Spherical Voronoi diagram of embeddings projected onto the unit sphere using UMAP with cosine distance as the metric. The embeddings represent system utterances from the `police` domain of the MultiWOZ2.1 dataset. Legends indicate the ground-truth action associated to each embedding and the centroids used to generate the partitions for all the actions in this domain.

Embeddings	F_1 score 1-shot	F_1 score 5-shot	Accuracy 1-shot	Accuracy 5-shot	Anisotropy intra(↑)	Anisotropy inter(↓)	Anisotropy Δ (↑)
GloVe	19.47 ± 2.47	24.54 ± 2.45	26.07 ± 4.52	33.30 ± 4.19	0.653	0.642	0.010
BERT	21.93 ± 2.40	31.11 ± 2.56	28.33 ± 3.76	39.98 ± 3.56	0.711	0.761	-0.049
Sentence-BERT	23.48 ± 2.62	35.71 ± 2.94	33.03 ± 4.70	47.47 ± 3.60	0.440	**0.404**	0.036
GTR-T5	26.53 ± 2.29	**41.10** ± 2.37	35.76 ± 4.00	**52.73** ± 3.16	0.681	0.714	-0.033
DSE	**27.53** ± 2.70	39.90 ± 3.08	**35.93** ± 4.54	51.73 ± 3.41	0.633	0.608	0.026
TOD-BERT	21.23 ± 2.03	32.28 ± 2.33	29.26 ± 3.99	41.71 ± 3.68	**0.848**	0.885	-0.038
DialoGPT	21.74 ± 2.10	32.01 ± 2.38	27.65 ± 3.47	41.05 ± 3.64	0.700	0.726	-0.026
SBD-BERT	19.09 ± 2.10	23.83 ± 2.22	25.80 ± 3.56	32.14 ± 3.62	0.651	0.596	0.055
D2F-Hard$_{single}$	**34.64** ± 2.90	**49.63** ± 2.87	**42.77** ± 4.61	**58.63** ± 3.27	**0.526**	**0.424**	**0.103**
D2F-Hard$_{joint}$	31.46 ± 2.61	46.89 ± 2.50	39.45 ± 4.22	56.43 ± 2.98	0.514	0.481	0.033
D2F$_{single}$	**<u>35.55</u>** ± 3.51	**<u>49.75</u>** ± 2.48	**<u>43.15</u>** ± 5.24	**<u>59.93</u>** ± 3.06	0.516	**<u>0.321</u>**	**<u>0.195</u>**
D2F$_{joint}$	33.19 ± 2.95	46.90 ± 2.66	41.22 ± 4.40	57.07 ± 2.92	**0.545**	0.429	0.116

Table 3: Similarity-based few-shot classification results on SpokenWOZ. The intra- and inter-action anisotropy are also provided along their difference (Δ).

using different query utterances for all actions. We observe a similar pattern across both datasets: an increase in variability and a drop in performance for all embedding types in SpokenWOZ. However, D2F embeddings still outperform all baselines and their D2F-Hard counterparts. For a more qualitative analysis, Table 5 provides an example of the rankings obtained for the query "your phone please" with the target action `[request phone_number]` on SpokenWOZ. As seen, DSE errors arise due to embeddings being closer if they correspond to consecutive utterances (`inform` and `request` utterances). Sentence-BERT errors occur due to the retrieval of utterances semantically re-

Embeddings	NDCG@10♣	NDCG@10★
GloVe	26.55 ± 0.57	25.09 ± 2.28
BERT	26.98 ± 0.80	27.74 ± 2.00
Sentence-BERT	30.88 ± 0.70	30.07 ± 2.23
GTR-T5	33.21 ± 0.60	32.74 ± 2.44
DSE	**38.09** ± 0.71	**33.94** ± 2.47
TOD-BERT	30.55 ± 0.74	25.63 ± 1.88
DialoGPT	28.86 ± 0.71	27.92 ± 2.01
SBD-BERT	27.20 ± 0.83	22.24 ± 1.93
D2F-Hard$_{single}$	**60.87** ± 0.47	**42.48** ± 2.77
D2F-Hard$_{joint}$	58.38 ± 0.72	40.03 ± 2.52
D2F$_{single}$	**67.31** ± 0.42	**43.12** ± 2.92
D2F$_{joint}$	66.50 ± 0.49	40.97 ± 2.61

Table 4: Ranking-based results on the unified TOD evaluation set (♣) and SpokenWOZ (★).

lated to "number" and "phone." In contrast, all D2F-retrieved utterances correctly represent different ways to request a phone number, even though half were considered incorrect due to the lack of slot label standardization across different domains (e.g., `phone_number` and `phone`).[7] Nonetheless, for clustering utterances by similarity to extract a dialog flow without annotation, D2F would successfully cluster these 10 utterances together as they correspond to semantically equivalent actions (`[request phone_number]` and `[request phone]`).

7 Dialog Flow Extraction Evaluation

Dialog flow extraction is an underexplored hard-to-quantify and challenging task with nuances in definition. However, to evaluate embedding quality, we formally define the problem as follows: Let $\mathcal{U}$ and $\mathcal{A}$ denote sets of TOD utterances and actions, respectively. Let $\mathcal{U}$ and $\mathcal{A}$ be sets of TOD utterances and actions, respectively. Let $\alpha : \mathcal{U} \mapsto \mathcal{A}$ be a (usually unknown) function mapping an utterance to its corresponding action. Let $d_i = (u_1, \cdots, u_k)$ be a dialog with $u_j \in \mathcal{U}$, and $t_i = (\alpha(u_1), \cdots, \alpha(u_k)) = (a_1, \cdots, a_k)$ its conversion to a sequence of actions, referred to as a *trajectory*. Given a set of m dialogs, $D = \{d_1, \cdots, d_m\}$, and after conversion to a set of action trajectories, $D^t = \{t_1, \cdots, t_m\}$, the goal is to extract the common dialog flow by combining all the trajectories in D^t. This common flow is represented as a weighted actions transition graph $G_D = \langle A, E, w_A, w_E \rangle$ where A is the set of actions, E represents edges between actions, the edge weight $w_E(a_i, a_j) \in [0, 1]$ indicates how often a_i is followed by a_j, and the action weight $w_A(a_i) \in [0, 1]$ is its normalized frequency.[8]

[7] This lack of slot standardization also affects results in Tables 3 and 4.

7.1 Evaluation Details

For each domain in SpokenWOZ, we build and compare its reference graph G_D against the induced graph $\hat{G}_D$ using different embeddings. The reference graph G_D is built from the trajectories D^t generated using the ground truth action labels —*e.g.* Figure 2 is indeed $G_{hospital}$. In contrast, the induced graph $\hat{G}_D$ is built *without any annotation* by clustering all the utterance embeddings in D and using the cluster ids as action labels to generate the trajectories $\hat{D}^t$. That is, for G_D, we have $\alpha(u_i) = a_i$, while for $\hat{G}_D$, we have $\alpha(u_i) = c_i$ where c_i is the cluster id assigned to $\mathbf{u}_i$. To compare the induced and reference graphs, we report the difference in the number of nodes between them as the evaluation metric.[9] Despite its simplicity, this metric allows us to compare the complexity of the induced vs. reference graph in terms of their sizes (induced actions). Furthermore, to avoid the influence of infrequently occurring utterances/actions on graph size, we prune them by removing all nodes a with $w_A(a) < \epsilon = 0.02$ (noise threshold).

In practice, the total number of actions to cluster is unknown in advance. For instance, a hierarchical clustering algorithm can be used to approximate this number (see Appendix F). However, for evaluation purposes, we set the number of clusters in each domain to be equal to the ground truth number so that all the embeddings are evaluated under the same best-case scenario in which this number is known in advance. Therefore, all the induced graphs are built and processed equally, making the input embeddings the only factor influencing the final graph.

7.2 Dialog Flow Extraction Results

Table 6 shows the results obtained when comparing the different induced graphs. We can see that graphs obtained with baseline embeddings tend to underestimate the complexity of each domain, producing less meaningful graphs with fewer states than their references.[10] Among the baseline embed-

[8] Even though having states as individual actions makes them non-Markovian, this graph is easy to interpret and directly links the quality of individual actions to the overall flow's quality.

[9] One cluster id c_i can correspond to multiple a_is and vice versa, preventing a direct comparison between $\hat{G}_D$ and G_D.

[10] For instance, Figure A1 and A2 in Appendix show the induced $\hat{G}_{hospital}$ for Sentence-BERT and DSE containing

Rank	DSE	Sentence-BERT	D2F$_{single}$
1.	-uh my phone number is 7 4■	-okay may i have your phone number please□	-please get their phone number□
2.	-okay okay now please get your number	-may i get your phone number	-okay may i have your phone number please□
3.	-okay may i have your phone number please□	-okay may i know your telephone number please	-okay may i know your telephone number please
4.	-thank you on the phone number□	-okay can i please get your id number♣	-may i get your phone number
5.	-okay may i know your telephone number please	-okay may i have your phone name in case for cooking the table★	-um can i please have their phone number□
6.	-okay great emma please have your contact number	-okay and may i have your number please	-okay so may i have the phone number with me
7.	-my number is 2 10■	-okay and may i have your number please	-okay i'm i also need phone number□
8.	-the number is you see♠	-okay and may i have your number please	-no problem um but for the information can i have your phone number
9.	-okay and may i have your number please	-okay and your car number♡	-thank you on the phone number□
10.	-okay and may i have your number please	-this product uh may i have your phone number please	-okay can i get your phone number please to make that booking

Table 5: Top-10 retrieved utterances on SpokenWOZ for the query "your phone please" with action label `[request phone_number]`. Errors are highlighted in red with wrong action marked as: ■`[inform phone_number]`; ♠`[inform plate_number]`; ♣`[request id_number]`; ★`[request name]`; ♡`[request plate_number]`; □`[request phone]`.

Embeddings	**Taxi** (31)	**Police** (23)	**Hospital** (18)	**Train** (49)	**Restaurant** (59)	**Attraction** (45)	**AVG.**
D2F$_{single}$	9.68% (+3)	<u>4.35%</u> (-1)	**11.11%** (-2)	<u>2.04%</u> (+1)	<u>5.08%</u> (-3)	**8.89%** (+4)	<u>6.86%</u>
D2F$_{joint}$	**3.23%** (+1)	**8.70%** (-2)	<u>5.56%</u> (-1)	**10.20%** (-5)	23.73% (-14)	<u>0.00%</u> (0)	**8.57%**
D2F-Hard$_{single}$	12.90% (-4)	26.09% (-6)	16.67% (-3)	**10.20%** (-5)	**10.17%** (-6)	15.56% (+7)	15.26%
D2F-Hard$_{joint}$	<u>0.00%</u> (0)	**8.70%** (-2)	33.33% (-6)	20.41% (-10)	25.42% (-15)	13.33% (-6)	16.87%
DSE	32.26% (-10)	17.39% (-4)	33.33% (-6)	30.61% (-15)	27.12% (-16)	26.67% (-12)	27.90%
DialoGPT	32.26% (-10)	34.78% (-8)	22.22% (-4)	44.90% (-22)	64.41% (-38)	51.11% (-23)	41.61%
BERT	54.84% (-17)	30.43% (-7)	22.22% (-4)	46.94% (-23)	59.32% (-35)	42.22% (-19)	42.66%
Sentence-BERT	48.39% (-15)	43.48% (-10)	55.56% (-10)	57.14% (-28)	50.85% (-30)	55.56% (-25)	51.83%
GTR-T5	41.94% (-13)	43.48% (-10)	66.67% (-12)	51.02% (-25)	61.02% (-36)	53.33% (-24)	52.91%
SBD-BERT	77.42% (-24)	43.48% (-10)	38.89% (-7)	71.43% (-35)	86.44% (-51)	86.67% (-39)	67.39%
TOD-BERT	74.19% (-23)	78.26% (-18)	55.56% (-10)	85.71% (-42)	83.05% (-49)	82.22% (-37)	76.50%

Table 6: Comparison of induced graph size vs. reference graph size for each single-domain in SpokenWOZ, measured by the number of nodes (actions). The table shows the normalized absolute difference (%) and raw difference in parentheses. Column headers indicate the size of each reference graph (G_D). Lower differences suggest a better match in graph complexity.

dings, DSE stands out (27.90% average difference across domains), suggesting that dialogue-related embeddings are better at capturing the communicative and informative functions of dialog utterances than semantically meaningful embeddings. Notably, D2F embeddings trained with the proposed soft contrastive loss induce graphs closest in complexity to the references across domains (6.86% and 8.57% average difference for D2F$_{single}$ and D2F$_{joint}$, respectively) compared to both D2F-Hard embeddings trained with the vanilla supervised contrastive loss and baselines.[11] Finally, it is also worth noting that the D2F graphs are relatively consistent across different domains, even though some domains had only a small amount of in-domain data during training. For instance, the `hospital` and `police` domains make up only 0.11% and 0.07% of the training set (Table A1).

8 Conclusions

This paper introduced Dialog2Flow (D2F), embeddings pre-trained for dialog flow extraction grouping utterances by their communicative and informative functions in a latent space. D2F embeddings were trained on a comprehensive dataset of twenty task-oriented dialog datasets with standardized action annotations, released along with this work.

Future work will enhance D2F embeddings by exploring larger backbone models and advanced methods for sentence embeddings (Jiang et al., 2023, 2022). We will also investigate more sophisticated techniques for extracting and representing dialog flows, such as using subtask graphs (Sohn et al., 2023) or adapting dependency parsing for complex dialog structures (Qiu et al., 2020). Additionally, potential applications include using D2F embeddings to ground LLMs in domain-specific flows for improved transparency and controllability (Raghu et al., 2021), and integrating D2F embeddings into various TOD downstream tasks like dialog state tracking and policy learning.

10 and 6 less nodes than the reference graph, respectively.

[11] Figure A3 shows $\hat{G}_{hospital}$ for D2F$_{joint}$ with only 1 fewer node than the reference. Source code is provided to generate graphs for any given dialogue collection and embedding, allowing manual assessment of superior D2F graph quality.

9 Ethical Considerations

We are committed to ensuring the ethical use of our research outcomes. To promote transparency and reproducibility, we will release the source code and pre-trained model weights under the MIT license. This allows for wide usage and adaptation while maintaining open-source principles.

However, to prevent potential license incompatibilities among the various task-oriented dialogue (TOD) datasets we have utilized, we will not release our unified TOD dataset directly. Instead, we will provide a script that can generate the unified dataset introduced in this paper. This approach allows users to select the specific TOD datasets they wish to include, ensuring compliance with individual dataset licenses.

We acknowledge that gender bias present in the original data could be partially encoded in the embeddings. This may manifest as assumptions about the agent's gender, such as the agent being male or female. We advise users to be aware of this potential bias and encourage further research to mitigate such issues. Continuous efforts to audit and address biases in data and models are essential to ensure fair and equitable AI systems.

10 Limitations

Our work represents a preliminary exploration with a focus on task-oriented dialogues (TODs) using a relatively simple encoder model. While this work aims to draw attention to this underexplored area, there are a number of limitations that must be acknowledged:

1. **Scope of Dialogues:** Our study is restricted to task-oriented dialogues. Consequently, the findings and methods may not generalize well to more complex and diverse types of dialogues, particularly those of a non-task-oriented nature. Future research should explore these methods in a broader range of dialogue types to assess their generalizability.

2. **Domain Specificity:** The model has been trained on a specific collection of domains, dialogue acts, and slots. This limits its ability to generalize to unseen domains or dialogues that involve more complex and varied interactions. Expanding the range of training data to include a wider variety of domains and dialogue types is necessary to improve the model's robustness and applicability.

3. **Model Complexity:** The encoder model used in this work is relatively standard. There is potential for improvement by employing larger and more advanced models to obtained the final sentence embeddings.

4. **Data Size:** Despite being the largest dataset with standardized utterance annotations and the largest spoken TOD dataset, the datasets used in this study are limited in size. Larger datasets are necessary to fully explore and validate the proposed methods. We encourage the research community to build upon this work by utilizing more extensive datasets to enhance the reliability and validity of the results. For instance, perhaps named entity tags may be used as slots to expand annotation beyond pure task-oriented dialogues.

5. **Evaluation Metrics:** The evaluation metrics employed in this study, while standard, may not capture all aspects of performance relevant to real-world applications. Developing and utilizing a broader set of evaluation metrics would provide a more comprehensive assessment of model performance. Specifically for dialogue flow evaluation, since there is not a standard metric yet, we encourage the research community to explore better ways to represent and quantify the quality of dialogue flows.

By highlighting these limitations, we hope to inspire further research that addresses these challenges, leading to more robust and generalizable solutions building on top of this work.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *Preprint*, arXiv:1803.11175.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017, Online. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

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.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709, Florence, Italy. Association for Computational Linguistics.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2023. Intent induction from conversations for task-oriented dialogue track at DSTC 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 242–259, Prague, Czech Republic. Association for Computational Linguistics.

Amine El Hattami, Issam H. Laradji, Stefania Raimondo, David Vázquez, Pau Rodríguez, and Christopher Pal. 2023. Workflow discovery from dialogues in the low data regime. *Transactions on Machine Learning Research*, 2023.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *Preprint*, arXiv:2307.16645.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Jurafsky. 2006. Pragmatics and computational linguistics. *The handbook of pragmatics*, pages 578–604.

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

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xiujun Li, Sarah Panda, JJ (Jingjing) Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. In *SLT 2018*.

Zhaojiang Lin, Andrea Madotto, Genta Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale N Fung. 2021. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. 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 4526–4536, Hong Kong, China. Association for Computational Linguistics.

Kun Qian, Satwik Kottur, Ahmad Beirami, Shahin Shayandeh, Paul Crook, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2022. Database search results disambiguation for task-oriented dialog systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1173, Seattle, United States. Association for Computational Linguistics.

Liang Qiu, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Structure extraction in task-oriented dialogues with slot clustering. *arXiv preprint arXiv:2203.00073*.

Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-Chun Zhu. 2020. Structured attention for unsupervised dialogue structure induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1889–1899, Online. Association for Computational Linguistics.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. 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 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Dinesh Raghu, Shantanu Agarwal, Sachindra Joshi, and Mausam. 2021. End-to-end learning of flowchart grounded task-oriented dialogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4348–4366, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. SpokenWOZ: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Sungryull Sohn, Yiwei Lyu, Anthony Liu, Lajanugen Logeswaran, Dong-Ki Kim, Dongsub Shim, and Honglak Lee. 2023. TOD-Flow: Modeling the structure of task-oriented dialogues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3355–3371, Singapore. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer.

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

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. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5786–5798, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew Arnold. 2022. Virtual augmentation supported contrastive learning of sentence representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 864–876, Dublin, Ireland. Association for Computational Linguistics.

Jianguo Zhang, Kun Qian, Zhiwei Liu, Shelby Heinecke, Rui Meng, Ye Liu, Zhou Yu, Huan Wang, Silvio Savarese, and Caiming Xiong. 2024. DialogStudio: Towards richest and most diverse unified dataset collection for conversational AI. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2299–2315, St. Julian's, Malta. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

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

A Unified TOD Dataset

Dialog acts: inform(64.66%) request(12.62%) offer(6.62%) inform_success(3.07%) good_bye(2.67%) agreement(2.45%) thank_you(2.25%) confirm(2.10%) disagreement(1.60%) request_more(1.06%) request_alternative(0.90%) recommendation(0.70%) inform_failure(0.64%) greeting(0.31%) confirm_answer(0.18%) confirm_question(0.17%) request_update(0.02%) request_compare(0.01%)

Domains: movie(32.98%) restaurant(13.48%) hotel(10.15%) train(4.52%) flight(4.30%) event(3.56%) attraction(3.50%) service(2.44%) bus(2.28%) taxi(2.21%) rentalcars(2.20%) travel(2.16%) music(1.81%) medium(1.66%) ridesharing(1.30%) booking(1.21%) home(1.01%) finance(0.79%) airline(0.69%) calendar(0.69%) fastfood(0.68%) insurance(0.61%) weather(0.58%) bank(0.47%) hkmtr(0.36%) mlb(0.35%) ml(0.31%) food(0.30%) epl(0.30%) pizza(0.25%) coffee(0.24%) uber(0.24%) software(0.23%) auto(0.21%) nba(0.20%) product_defect(0.17%) shipping_issue(0.16%) alarm(0.13%) order_issue(0.13%) messaging(0.13%) hospital(0.11%) subscription_inquiry(0.11%) account_access(0.11%) payment(0.10%) purchase_dispute(0.10%) nfl(0.09%) chat(0.08%) police(0.07%) single_item_query(0.06%) storewide_query(0.06%) troubleshoot_site(0.06%) manage_account(0.06%)

Table A1: Standardized dialog act and domain labels in our unified TOD datasets, ordered by their proportion of utterances.

Our training data is sourced from a diverse range of TOD datasets meticulously curated in DialogStudio (Zhang et al., 2024). DialogStudio comprises over 80 dialog datasets, with 30 focusing on task-oriented conversations. We conducted a comprehensive manual analysis of these 30 TOD datasets to identify those from which we could extract dialog act and/or slot annotations. From this analysis, we identified 20 datasets that met our criteria, as summarized in Table 1. The datasets in DialogStudio are unified under a consistent format while retaining their original information. However, this format only unifies the access to the conversations *per se*, omitting annotations and components of task-oriented dialogs. We then manually inspected each dataset to locate and extract the necessary annotations. This process involved identifying where and how annotations were stored originally in each dataset, extracting dialog act and/or slot annotations for each turn, either explicitly or implicitly by keeping track of the changes in the dialog state annotation from one turn to the next, and standardizing domain names and dialog act labels across datasets.

To standardize dialog act labels, we mapped the 44 unique labels found across datasets to 18 normalized dialog act labels, informed by the semantic meaning described in the original dataset papers (mapping detailed in Table A3). After this process, we unified all datasets under a consistent format, detailed in the next subsection, incorporating per-turn dialog act and slot annotations. The resulting unified TOD dataset comprises 3.4 million utterances annotated with 18 standardized dialog acts,

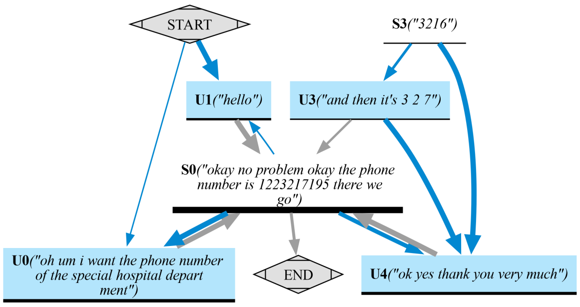


Figure A1: $\hat{G}_{hospital}$ graph obtained with Sentence-BERT (8 induced actions in total). Node labels correspond to the cluster ID along a representative utterance (the closest to the cluster centroid).

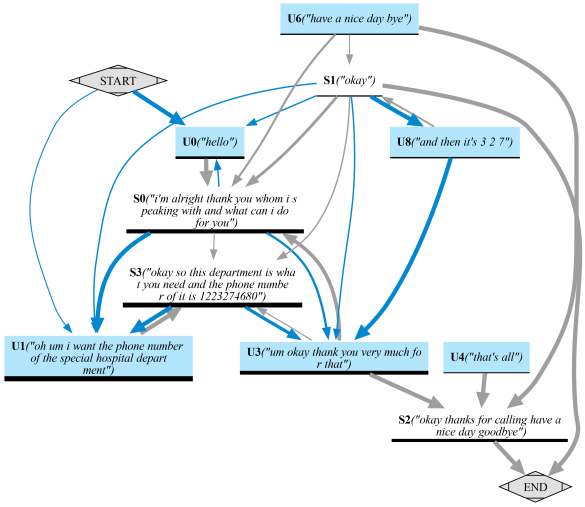


Figure A2: $\hat{G}_{hospital}$ graph obtained with DSE (12 induced actions in total). Node labels correspond to the cluster ID along a representative utterance (the closest to the cluster centroid).

524 unique slot labels, and 3,982 unique action labels (dialog act + slots). These annotations span across 52 different domains, as detailed in Table 1.

Our unified TOD dataset is a valuable resource providing a comprehensive and standardized collection of annotated utterances across diverse domains under a common format.

A.1 Dataset Format

Our unified dataset standardizes the TOD datasets into the following common JSON format with per-utterance annotations:

```
{
 "stats": {"domains": {...},
           "labels": {..}},
 "dialogs": {
  "<DIALOGUE_ID0>": [
   {
    "speaker": <SPEAKER>,
    "text": <RAW_UTTERANCE>,
```

```
      "domains": [...],
      "labels": {
        "dialog_acts": {
          "acts" : [...],
          "main_acts" : [...],
          "original_acts" : [...],
        },
        "slots": [...],
        "intents": [...]
      }
    },
    ...
  ],
  "<DIALOGUE_ID1>": [...],
  ...
 }
}
```

The JSON structure has two main parts: a "stats" header and a "dialogs" body. The "stats" field provides statistics about the labels and domains in the dataset. The "dialogs" field contains dialog IDs, each linked to a list of annotated utterance objects. Each utterance object includes its speaker, text, domains, and associated labels for dialog acts, slots, and intents. Dialog act labels contain the original labels ("original_acts") as well as their standardized values ("acts") and parent values ("main_acts") as mapped in Table A3.

B Training Details

Following the experimental setup of DSE (Zhou et al., 2022) and TOD-BERT (Wu et al., 2020), we set the contrastive head dimension to $d = 128$ and use BERT$_{base}$ as the backbone model for the encoder[12]. Additional configurations reported in Appendix C.

For the soft contrastive loss, the semantic similarity measure $\delta(y_i, y_j) = \mathbf{y}_i \cdot \mathbf{y}_j$ was computed using label embeddings $\mathbf{y}$ obtained with the best-performing pre-trained Sentence-BERT model on semantic search, namely the `multi-qa-mpnet-base-dot-v1` model. As shown in Appendix C, we also experimented with the `all-mpnet-base-v2` model, which has the best average performance among all pre-trained Sentence-BERT models. The soft label temperature parameter was set to $\tau' = 0.35$ after a preliminary study determined it to be a reasonable threshold for both joint and single training targets (Appendix E).

In line with the settings of DSE and TOD-BERT, the learning rates for the contrastive head and the encoder model were set to 3e-4 and 3e-6, respectively. The contrastive temperature parameter τ was set to 0.05. Models were trained for 15 epochs and then saved for evaluation. The maximum sequence length for the Transformer encoder was empirically set to 64 to accommodate at least 99% of the samples, as most TOD utterances are short. Finally, the batch size was set to 64 since we found that, contrary to typical self-supervised contrastive learning, larger batch sizes resulted in lower perfor-

[12]Thus, the embedding size is $n = 768$.

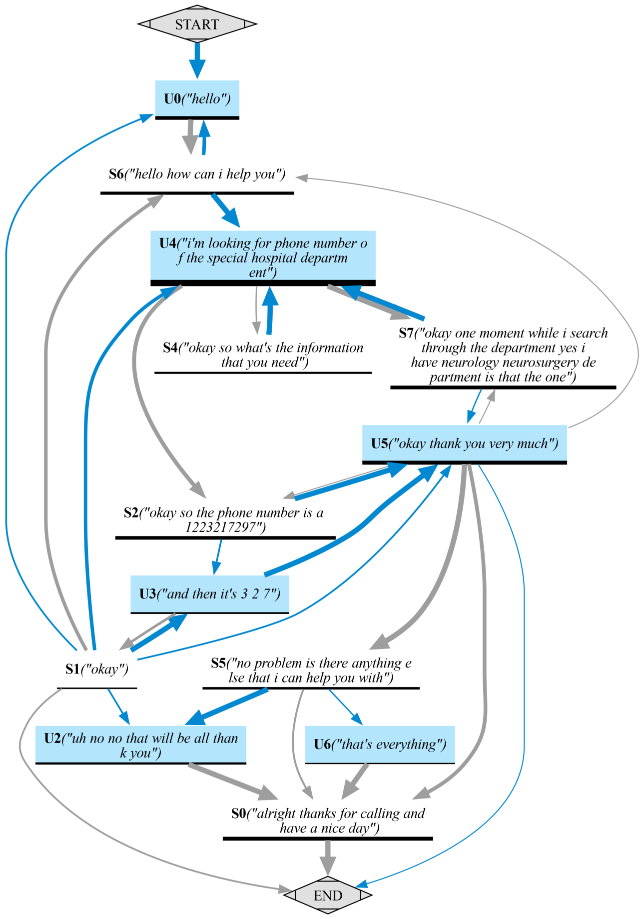


Figure A3: $\hat{G}_{hospital}$ graph obtained with D2F$_{joint}$ containing only one node less than the reference graph in Figure 2. Node labels correspond to the cluster ID along a representative utterance (the closest to the cluster centroid). Although not the exact same graph as the reference, this graph still allows us to understand the common flow of the conversations with a similar degree of detail: first, the user and system greet each other (**U0** and **S6**), then the user inform the reason of the call requesting the phone number of a department (**U4**), the agent may confirm the department (**S7**) or request more information (**S4**) before providing the phone number (**S2**). The user may then either confirm the number (**U3**) or thank the system (**U5**). Finally, the system asks if anything else is required (**S5**), to which the user may either finish the conversation (**U6**) or, more likely, thank the system (**U2**) before the system says goodbye (**S0**).

DF2 Variation	F_1 score	Δ Anisotropy (↑)
D2F-Hard$_{single}$	67.82	0.332
* *DSE Backbone*	*+2.66*	*+0.011*
+ *Self-Supervision*	*-7.41*	*-0.002*
D2F-Hard$_{joint}$	66.22	0.230
* *DSE Backbone*	*+1.97*	*+0.010*
+ *Self-Supervision*	*-6.01*	*-0.064*
D2F$_{single}$	70.89	0.597
* *DSE Backbone*	*+0.97*	*+0.012*
* *all-mpnet-base-v2 Label*	*-0.60*	*-0.038*
+ *Self-Supervision*	*-6.65*	*-0.189*
– *Contrastive Head*	*-1.13*	*-0.047*
D2F$_{joint}$	70.94	0.451
* *DSE Backbone*	*+0.65*	*+0.011*
* *all-mpnet-base-v2 Label*	*-0.34*	*-0.038*
+ *Self-Supervision*	*-8.06*	*-0.126*
– *Contrastive Head*	*-3.78*	*-0.073*

Table A2: Ablation study results for various D2F configurations. Additions, subtractions, and replacements of components are marked with **+**, **–**, and * symbols, respectively. Values show the impact on 5-shot classification F_1 score and anisotropy as reported in Table 2.

mance.[13]

C Ablation study

We conducted an ablation study to evaluate the effects of different configurations on the performance of our D2F models. The following variations were tested:

- **DSE Backbone**: Replacing the original BERT encoder with the pre-trained DSE model.
- **Label Encoder**: Using the Sentence-BERT model `all-mpnet-base-v2`, which has the best reported average performance for semantic similarity.
- **Self-Supervision**: Adding the self-supervised loss from DSE ($\mathcal{L}^{self}$) trained jointly with our targets ($\mathcal{L} + \mathcal{L}^{self}$) on the same data as DSE. This was done to evaluate whether jointly training as DSE would yield better performance than using the pre-trained DSE encoder directly as the backbone.

[13]A grid search with batch sizes 64, 128, 256, and 512 was performed, training models for one epoch and evaluating the similarity-based 5-shot F_1 score on our evaluation set. Larger batch sizes consistently yielded lower scores across all models (both standard and soft supervised contrastive loss models). For instance, DFD$_{joint}$ scored 63.23, 61.64, 58.77, and 56.30 for batch sizes 64, 128, 256, and 512, respectively.

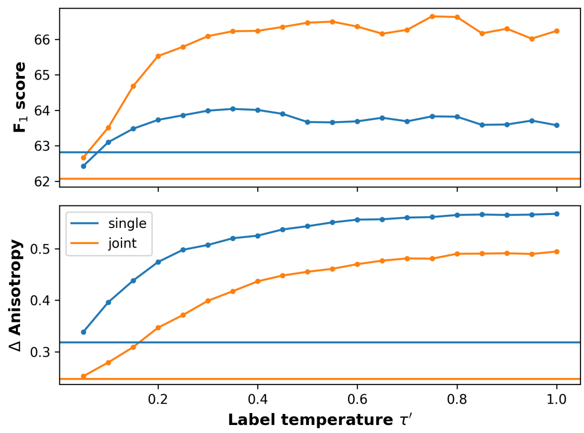


Figure A4: Change in F_1 score (top) and Δ Anisotropy (bottom) with respect to the label temperature τ' (x-axis). The blue and orange curves represent D2F$_{single}$ and D2F$_{joint}$, respectively. Horizontal lines indicate the performance of their D2F-Hard counterparts using the standard hard supervised contrastive loss.

- **Contrastive Head Removal**: Removing the contrastive head used during training.

The results of these variations are summarized in Table A2. The only configuration that consistently improved performance was the replacement of the backbone model with the pre-trained DSE model, increasing the F_1 score and anisotropy across all variations.

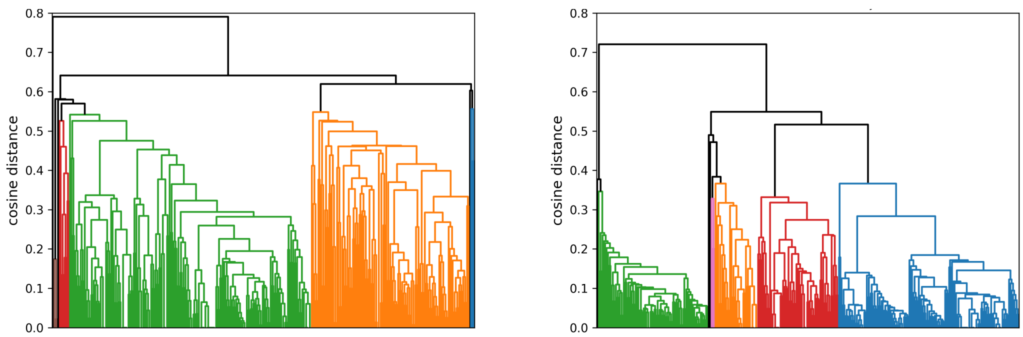
In contrast, adding self-supervision generally degraded performance, indicating that the additional DSE self-supervised loss $\mathcal{L}^{self}$ may not complement our targets effectively when trained jointly. Similarly, removing the contrastive head during training resulted in a notable performance drop, highlighting its importance.[14]

D Supervised Soft Contrastive Loss Explanation

Let $p(pos = j \mid x_i)$ be the probability of j-th sample in the batch being positive given the i-th anchor. Then, the loss in Equation 1 is equivalent to the categorical cross-entropy of correctly classifying the positions in the batch with positive samples for the given x_i anchor:

$$-\sum_{j=1}^{N} p(pos = j \mid x_i) \log \hat{p}(pos = j \mid x_i) \quad (2)$$

[14]Each different configuration required re-training the model for 15 epochs, a process that takes approximately 5 days on a single GeForce RTX 3090 GPU.



(a) **Sentence-BERT**

(b) **D2F**$_{joint}$

Figure A5: Dendrograms obtained by hierarchically clustering all user utterances in the `hospital` domain using Sentence-BERT embeddings (left) and D2F$_{joint}$ embeddings (right). The clustering and the plots were obtained using the `AgglomerativeClustering` class from `scikit-learn`, with the number of clusters set to 4 (indicated by different colors).

where the true/target distribution p is defined as

$$p(pos\!=\!j\,|\,x_i) = \begin{cases} \frac{1}{|\mathcal{P}_i|}, & \text{if } y_i = y_j \\ 0, & \text{if } y_i \neq y_j \end{cases} \qquad (3)$$

and the predicted distribution $\hat{p}$ is an N-way softmax-based distribution proportional to the alignment/similarity between (the vectors of) the given x_i anchor and each x_j^+ sample:

$$\hat{p}(pos\!=\!j\,|\,x_i) = \frac{e^{\mathbf{z}_i \cdot \mathbf{z}_j^+/\tau}}{\sum_{k=1}^{N} e^{\mathbf{z}_i \cdot \mathbf{z}_k^+/\tau}}$$

Note that the target distribution in Equation 3 treats all samples with different labels as equally negative, independently of the semantics of the labels. However, we hypothesize that better representations can be obtained by taking advantage of the semantics of the labels to model more nuanced relationships. More precisely, let $\delta(y_i, y_j)$ be a semantic similarity measure between both labels, we define a new target distribution $p(pos\!=\!j\,|\,x_i) \propto \delta(y_i, y_j)$ as:

$$p(pos\!=\!j\,|\,x_i) = \frac{e^{\delta(y_i,y_j)/\tau'}}{\sum_{k=1}^{N} e^{\delta(y_i,y_k)/\tau'}} \qquad (4)$$

where τ' is the temperature parameter to control how soft/hard the negative labels are (Appendix E).[15] Note that unlike Equation 3,[16] this equation allows searching for an encoder that tries to separate anchors and negatives by *degrees proportional to how semantically similar their labels are*. Therefore, by replacing Equation 4 in Equation 2, our soft contrastive loss is finally defines as:

$$\ell_i^{soft} = -\sum_{j=1}^{N} \frac{e^{\delta(y_i,y_j)/\tau'}}{\sum_{k=1}^{N} e^{\frac{\delta(y_i,y_k)}{\tau'}}} \log \frac{e^{\mathbf{z}_i \cdot \mathbf{z}_j^+/\tau}}{\sum_{k=1}^{N} e^{\frac{\mathbf{z}_i \cdot \mathbf{z}_k^+}{\tau}}}$$

[15]On both extremes, sufficiently small τ' will resemble the original distribution in Equation 3 while sufficiently large τ' will resemble a uniform distribution leading to no contrast between positive and negative samples.

[16]Equation 3 encourages the encoder to separate all negatives 180° away from their anchors: if $y_i \neq y_j$, $\hat{p}(pos\!=\!j\,|\,x_i) \to 0 \Rightarrow e^{(\cdot)} \to 0 \Rightarrow \mathbf{z}_i \cdot \mathbf{z}_j^+ \to -1$.

E Soft Contrastive Loss Temperature

To understand the benefits of the "softness" introduced by our proposed contrastive loss compared to the conventional hard supervised contrastive loss, we conducted a preliminary study examining the impact of the label temperature parameter τ'. We trained models over three epochs, varying the temperature τ' across a range of values from 0.05 to 1.0 in increments of 0.05. This resulted in 42 different model variants: 20 each for D2F$_{single}$ and D2F$_{joint}$, and one for each D2F-Hard counterpart.

For each τ' value, we recorded the 5-shot classification F_1 score and Δ anisotropy values as outlined in Section 6. The results are depicted in Figure A4.

The plots reveal that as the temperature τ' increases from 0, indicating a transition from hard to softer negative labels, both F_1 scores and Δ anisotropy values improve beyond those obtained with the standard supervised contrastive loss. For both D2F$_{single}$ and D2F$_{joint}$ models, increasing the temperature leads to greater separation between intra-class and inter-class embeddings, as indicated by higher Δ anisotropy values.

The performance metrics exhibit a steady rise up to a temperature around between 0.35 and 0.4, beyond which Δ anisotropy values begin to plateau and F_1 scores become less stable. The advantage of using softer contrast is more pronounced for the joint target (D2F$_{joint}$, represented by the orange line), as evidenced by the larger gap between the orange curve and its corresponding horizontal line (D2F-Hard$_{joint}$).

However, it's important to note that these improvements diminish with additional training epochs. The final difference in performance metrics between soft and hard labels narrows after extended training, as reflected in the results reported in Table 2, where models were trained for 15 epochs.

F How Many Actions to Cluster?

In practice, determining the optimal number of clusters (actions) in dialog flow extraction is challenging because it directly affects the granularity of the extracted flows. Hierarchical clustering algorithms, such as agglomerative clustering, are preferred over centroid-based methods like k-means because they provide a visual representation of the data's hierarchical structure, which can be examined to decide the number of clusters or set a distance threshold.

Figure A5 illustrates dendrograms obtained by hierarchically clustering user utterances in the `hospital` domain using Sentence-BERT embeddings and D2F$_{joint}$ embeddings. The clustering and plotting were performed using the `AgglomerativeClustering` class from `scikit-learn`, with the number of clusters set to 4, represented by different colors.

The dendrograms reveal notable differences between the embeddings. The Sentence-BERT dendrogram (left) shows a structure with two main (semantic) groups with low variability in the distances between child and parent nodes, resulting in a more stretched plot. In contrast, the D2F$_{joint}$ dendrogram (right) displays a clearer separation into four main groups, with larger gaps between child and parent nodes at a certain level of the hierarchy, indicating distinct clusters. D2F$_{joint}$ embeddings were trained to minimize intra-action distances (pushing them towards the bottom of the dendrogram) and maximize inter-action distances (pushing parent nodes towards the top) facilitating easier identification of clusters. For instance, in the D2F$_{joint}$ dendrogram, the number of actions could be estimated to be between 4 and 7, or a distance threshold around 0.4 could be used to form the clusters.

In our experiments (Section 6), we used the ground truth number of clusters from annotations to ensure consistency in evaluation across the different embeddings. However, agglomerative clustering was employed to mimic closer a realistic scenario where the number of actions is not predefined.

Thus, hierarchical clustering methods provide a practical approach for approximating the number of actions in practice when such number is unknown.

Original	Standardized	Parent
inform	**inform** (slots)	inform
notify_fail notify_failure no_result nobook nooffer sorry cant_understand canthelp reject	**inform_failure**	
book offerbooked notify_success	**inform_success**	
request request_alt request_compare request_update	**request** (slots) **request_alternative** **request_compare** **request_update**	request
req_more request_more moreinfo hearmore	**request_more**	
confirm confirm_answer confirm_question	**confirm** (slots) **confirm_answer** **confirm_question**	confirmation
affirm affirm_intent	**agreement**	agreement
negate negate_intent deny	**disagreement**	disagreement
offer select multiple_choice offerbook	**offer**	offer
suggest recommend	**recommendation**	recommendation
greeting welcome	**greeting**	greeting
thank_you thanks thankyou	**thank_you**	thank_you
good_bye goodbye closing	**good_bye**	good_bye

Table A3: The original 44 dialog acts with their respective 18 standardized names used to unify all the datasets, along with a parent category grouping them further into 10 parent acts.