Template-assisted Contrastive Learning of Task-oriented Dialogue Sentence Embeddings

Minsik Oh
Stanford University

Jiwei Li
Zhejiang University

Guoyin Wang
Alibaba Qwen Pilot

minsik@stanford.edu, guoyin.wang@alibaba-inc.com

Abstract

Learning high quality sentence embeddings from dialogues has drawn increasing attentions as it is essential to solve a variety of dialogue-oriented tasks with low annotation cost. Annotating and gathering utterance relationships in conversations are difficult, while token-level annotations, *e.g.*, entities, slots and templates, are much easier to obtain. Other sentence embedding methods are usually sentence-level self-supervised frameworks and cannot utilize token-level extra knowledge. We introduce **T**emplate-**a**ware **D**ialogue **S**entence **E**mbedding (TaDSE), a novel augmentation method that utilizes template information to learn utterance embeddings via self-supervised contrastive learning framework. We further enhance the effect with a synthetically augmented dataset that diversifies utterance-template association, in which slot-filling is a preliminary step. We evaluate TaDSE performance on five downstream benchmark dialogue datasets. The experiment results show that TaDSE achieves significant improvements over previous SOTA methods for dialogue. We further introduce a novel analytic instrument of semantic compression test, for which we discover a correlation with uniformity and alignment. Our code is available at https://github.com/minsik-ai/Template-Contrastive-Embedding.

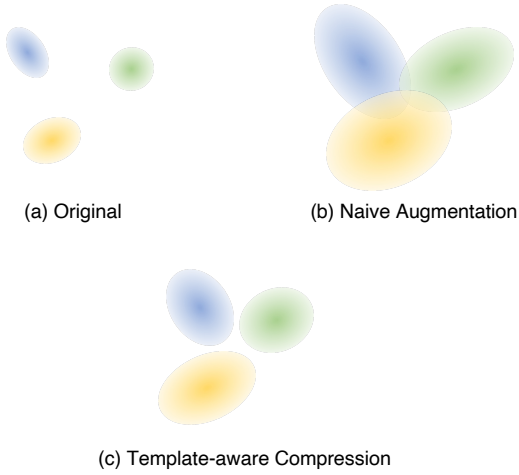

Figure 1: Embedding hyperspace changes with our method, from (a), (b) to (c). Ellipses denote sentence representations from the dataset, belonging to unique semantic groups. (a) shows the limited original data, (b) shows the effect of noisy data augmentation in which semantic clusters overlap, and (c) shows enhanced semantic group separation with our methods, with templates within each semantic group to constrain the embeddings.

1 Introduction

Learning sentence embeddings from dialogues has recently attracted increasing attentions (Zhou et al., 2022; Liu et al., 2021). [1] Learning high quality dialogue semantics (Hou et al., 2020; Krone et al., 2020; Yu et al., 2021) helps solving various downstream tasks, especially in the scenarios with limited annotations (Snell et al., 2017; Vinyals et al., 2016; Kim et al., 2018; Li et al., 2021).

Contrastive Learning (Hadsell et al., 2006) is a method to learn sentence embeddings by bringing semantically associated samples closer while pushing unrelated samples further apart. Unsupervised contrastive learning has been gaining momentum since it does not require human annotations, requiring supervision signals that augment original sentence. Some examples of supervision signals are document spans (Giorgi et al., 2021), Wikipedia entries (Nishikawa et al., 2022), consequent sentences (Zhou et al., 2022), prompt augmentations (Jiang et al., 2022) and dropout hidden weights (Gao et al., 2021).

Benefitting from the advance of contrastive learning, there has been solid success in learning universal sentence representations in both supervised (Reimers and Gurevych, 2019; Feng et al., 2022) and unsupervised manner (Gao et al., 2021; Chuang et al., 2022; Giorgi et al., 2021; Nishikawa et al., 2022; Jiang et al., 2022). However, universal

[1] Different from general embeddings (Appendix B).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22181–22198

sentence embeddings usually achieve undesirable performance in dialogue domain (Zhou et al., 2022; Wu et al., 2020a), since specific semantic relations between dialogue utterances exist (Appendix B).

In this paper, we explore how we can create semantically relevant sentence embeddings for dialogue. Templates (Kale and Rastogi, 2020) and slots are high-quality auxiliary data for dialogue understanding purposes (Kim et al., 2018; Bastianelli et al., 2020; FitzGerald et al., 2022). They are a variable representation of text structure and salient slot values. However, previous sentence embedding frameworks cannot incorporate such information. We present TaDSE, Template-aware Dialogue Sentence Embedding generation framework which produces superior text embeddings for dialogue understanding via template-aware data augmentation, training, and inference.

Our template-based data augmentation method (Section 3.1) exploits salient ingredients already present in task-oriented dialogue - templates, entities (slots), and their values. General purpose data augmentation methods, *e.g.*, rule-based methods or backtranslation (Feng et al., 2021; Wei and Zou, 2019; Zhang et al., 2022; Qu et al., 2020; Sennrich et al., 2016) are prone to semantic alterations or require a model (Wang et al., 2022). Our augmentation strategy produces consistently natural utterances and reinforces the dataset distribution in a realistic manner. We discover that our augmentation easily attain a stable performance increase, especially in combination with our training method, even if noise exists in synthetic data.

Our TaDSE training method (Section 3.2) encodes auxiliary template representations and their pairwise association with matching utterance representations. Each template is salient in regard to the semantic structure of the utterances, thus the model can improve itself by learning to distinguish between correct and mismatched utterance/template pairs. We introduce a pair of contrastive loss terms that designate the associated pairs of utterance and template as positive. Our pairwise training outperforms previous utterance-only unsupervised methods across five dialogue datasets.

Our TaDSE inference method (Section 3.3), which we define as "semantic compression test", is an instrument to inspect another conjecture of our training method, an interpretation that bringing correct utterance and template representations closer enhances representation. By enhancing specific semantics in the templates, the model can differentiate cosmetically similar utterances. Semantic compression improves the performance on augmentation-stable datasets, in addition to a noteworthy correlation with existing tools of uniformity/alignment (Wang and Isola, 2020).

Our contributions are summarized as follows:

1. We propose a special synthetic data augmentation, a novel data augmentation approach that aims to replicate *real-life* utterances.

2. We propose a novel training & inference pairwise dialogue sentence embedding learning framework, justified via SOTA performances.

3. Our experiments visibly show that the inferred utterance representations reshape the hyperspace in accordance with our expectations.

2 Related Works

Unsupervised Sentence Embedding methods train with contrastive objectives effectively to learn universal sentence embeddings. For vision, methods such as SimCLR (Chen et al., 2020a,b) have demonstrated the importance of data augmentation in contrastive learning. In NLP, methods such as SimCSE, DiffCSE, PromptBERT (Gao et al., 2021; Chuang et al., 2022; Jiang et al., 2022) show that simple augmentations such as dropout masking, token-wise masking, and prompt augmentation are an effective positive representation target. Our method differs from previously studied methods due to our novel application of semantically relevant token-wise templates for contrastive loss design.

Slot-filling and intent classification are major tasks for dialogue understanding purposes (Louvan and Magnini, 2020). Recent works perform the tasks jointly or in a multi-stage manner (Wu et al., 2020b; Zhang and Wang, 2016; Liu and Lane, 2016; Qin et al., 2019; Goo et al., 2018; Haihong et al., 2019). Each task has also been studied separately (Louvan and Magnini, 2020; Mesnil et al., 2013, 2015; Liu and Lane, 2015). In line with prior works, we perform slot-filling as a necessary step for representation learning in the dialogue domain, after which we perform the intent classification task. We repurpose existing tasks to model the semantic structure of utterances and assess the quality of the semantic information. This brings the benefit of both conceptually sound methodology and practical NLU applications via enhanced embeddings.

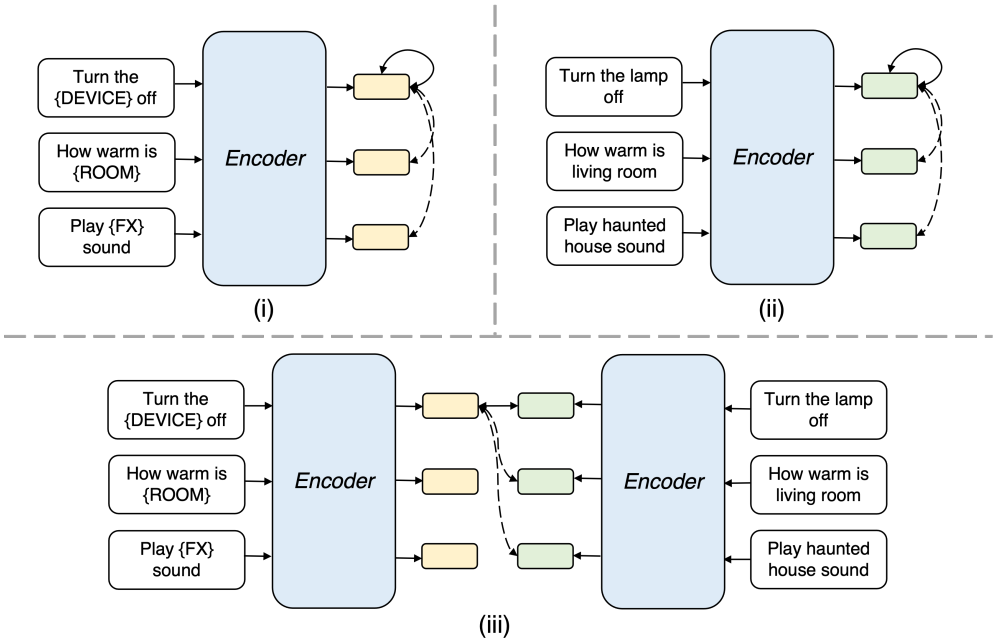

Figure 2: Our template contrastive learning methods. The first diagram displays template contrastive learning (L^t), second diagram displays utterance contrastive learning (L^u), and the third diagram displays pairwise contrastive learning (L^{pair}). *Encoder* represents the embedding generation model and yellow, and green represent template and utterance representations respectively. Solid bidirectional arrows designate positive pairs and dashed bidirectional arrows designate negative pairs.

Studying the representation space formed by learned embeddings has received influential attention, with recent work introducing uniformity/alignment to induce properties in relation to hypersphere (Wang and Isola, 2020). Anisotropy problem is also identified with language representations (Ethayarajh, 2019; Li et al., 2020; Gao et al., 2019), the problem of only narrow cone in the hyperspace being occupied by the embeddings. This behavior is also observed in multimodal setting (Liang et al., 2022). While we utilize the hyperspace analysis tools provided by previous works, we introduce a novel instrument of semantic compression which has the marked benefit of being semantically interpretable in regards to the meaning of the natural language sentences.

3 Proposed Method

3.1 Template Data Augmentation

In dialogue datasets such as SNIPS (Coucke et al., 2018) and ATIS (Hemphill et al., 1990), multiple utterances correspond to a single template, which we express as "utterance-template pairwise association". We posit that strengthening the diversity of utterance-template pairwise associations is essential for our training scheme. This variety of utterances per template will be retained in distributions from actual *real-life* scenarios.[2] We present a template-based augmentation strategy to replicate realistic usage patterns, with the added benefit of providing varied natural utterances.

We select a set of slots (*entities*) that are relevant to the dialogue domain, whether it be airlines, countries, or appliances, and categorize them to form a Slot Book. We construct permutations of the templates by filling the slot tokens with selected slot values. We select top-k frequent slot values from the training set to maintain the quality of utterance-template association. This method (Fig. 3) may be extended further (Section 8).

We utilize dialogue datasets of SNIPS (Coucke et al., 2018), ATIS (Hemphill et al., 1990), MASSIVE (FitzGerald et al., 2022), HWU64 (Liu et al., 2019) and CLINC150 (Larson et al., 2019). For MASSIVE, SNIPS, HWU64 and ATIS datasets, we utilize annotated templates and slot values already present in the dataset. For only CLINC150 dataset, we utilize a weak baseline of NER to automatically obtain slots and create templates, since no annotations are available. Slots of interest are

[2]Reasonably high percentage of customers booking an airplane ticket would tend to say "Could I book a plane ticket to {CITY}?" rather than complex variations of the template.

Src. Data	Strategy	Slots	Values	Templates	Orig. Utterances	Utterances	U/T Ratio
SNIPS	top-5	39	11.9K	7.4K	13.1K	163K	22x
ATIS	top-2	41	0.6K	3.3K	4.5K	239K	72x
MASSIVE	top-3	55	4.0K	10.3K	11.5K	78K	8x
HWU64	top-3	56	6.0K	16.6K	19.2K	133K	7x
CLINC150	top-5	17	1.7K	15.3K	15.3K	220K	14x
Total	-	208	24.2K	52.9K	63.6K	834K	16x

Table 1: Statistics of our augmented dialogue datasets. The "Slots" column is for slots (DEVICE, ROOM, etc.) while the "Values" column is values that fit the slots (television, lounge, etc.). "U/T ratio" denotes how many utterances exist per template on average after augmentation.

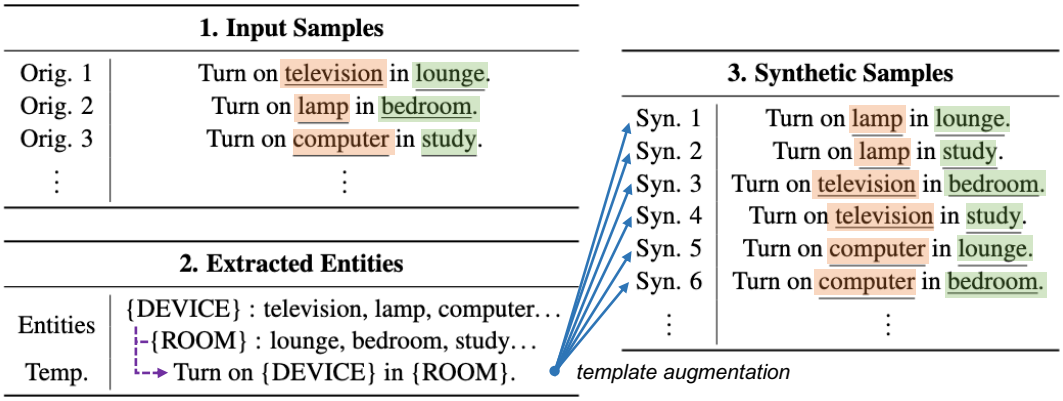

Figure 3: Our template data augmentation process in a simplified example, with a single template. In practice, thousands of templates and slot values exist per dataset (Table 1). We experiment with both manual annotations and automated slot-filling method.

cities, airlines, time, food, etc, with samples in Appendix A. The CLINC150-specific configuration is intended to verify whether the noisy baseline slot-filling system is functional with our data augmentation, training, and inference framework. We leave enhanced slot-filling techniques for future work (Section 8).

3.2 Pairwise Modeling

The effect of "anchoring" the sentence representations with auxiliary data has been studied with NLI-reliant hard-negatives (Gao et al., 2021), Wikipedia entries in multi-lingual settings (Nishikawa et al., 2022), with document spans (Giorgi et al., 2021), and co-occurring utterances in pre-training dialogue corpus (Zhou et al., 2022; Liu et al., 2021; Wu et al., 2020a). The aforementioned studies rely on incidental auxiliary data or pre-trained auxiliary models and thus are heavily reliant on distributions in a large corpus. We introduce a concept of "pairwise anchoring", in which we train with an auxiliary template generated from the utterance itself via tokenwise masking. This benefits the training procedure since it is impossible to pair the sentence with irrelevant data, and requires only a small training set for fine-tuning. We teach the model the capability to distinguish correct utterance and template pairs via contrastive learning (Fig. 2).

First, we define **template representation loss**, where we train a saliently masked anchor with which we further induce utterance representations. We train with contrastive loss and the data augmentation with dropout noise according to (Gao et al., 2021) framework. Let $\text{sim}(t_i, t_j)$ be cosine similarity $\frac{t_i^T t_j}{||t_i|| \cdot ||t_j||}$. Template representation is given as t_i and dropout-variant as t_i^+. Negative representations sampled from mini-batch are t_j. The template loss function is:

$$L_i^t = -\log \frac{e^{\text{sim}(t_i, t_i^+)/\tau_t}}{\Sigma_{j=1}^N e^{\text{sim}(t_i, t_j)/\tau_t}} \quad (1)$$

where τ_t is temperature hyperparameter for template representation. In addition, we further exper-

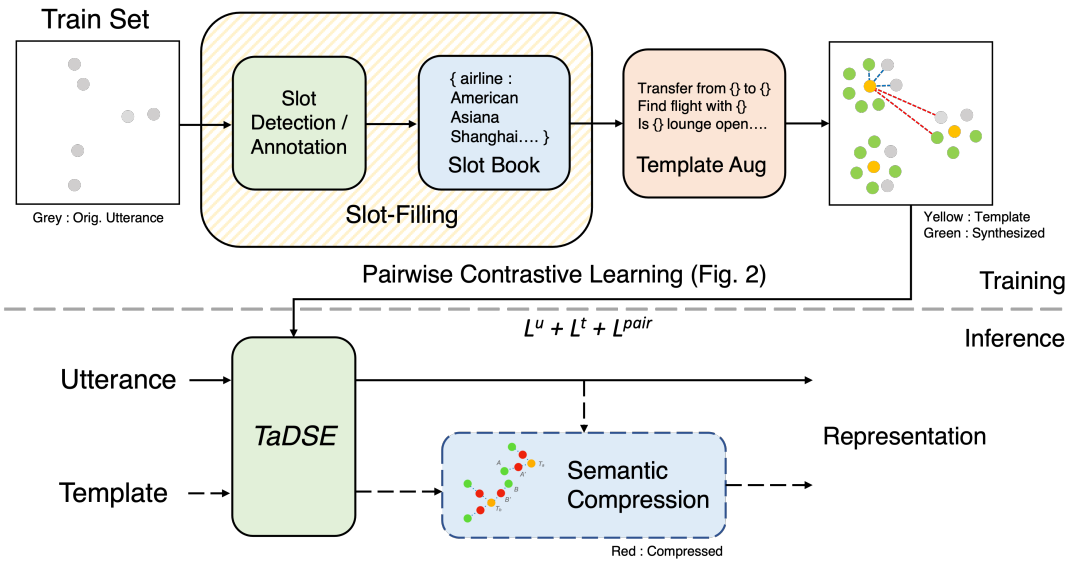

Figure 4: Our embedding generation process. Blue and red dashed lines are examples of positive and negative pairs for L^{pair} loss. Dashed arrows depict an alternative choice of semantic compression inference technique. Slot-filling baselines and template sources described in Section 3.1.

iment with a trainable MLP layer W_A to modify template representation t as $t' = W_A t$. This is an extension of pooling experiments performed on (Gao et al., 2021), the difference being that we focus on the effect of including MLP for templates only in an asymmetric configuration.

Next, we compute **utterance representation loss** similarly in a contrastive manner. This is to ensure we correctly learn utterance representation without over-reliance on templates. If we define u_i, u_i^+ and u_j similarly to t_i, t_i^+ and t_j in Eq. 1 (Gao et al., 2021), the utterance loss function is:

$$L_i^u = -\log \frac{e^{\text{sim}(u_i,u_i^+)/\tau_u}}{\Sigma_{j=1}^N e^{\text{sim}(u_i,u_j)/\tau_u}} \quad (2)$$

where τ_u is temperature for the utterance representation.

Lastly, we introduce **pairwise representation loss**, where we distinguish between correct and negative utterance-template pairs via contrastive learning to teach how certain semantically similar representations should group together (Section 5.4). We compare within utterances instead of templates as to ensure unique negatives in relation to the template augmented data in Section 3.1. Defining t_i, u_i, u_j same as Eq. 1, 2, (note we draw negative mini-batch representations from utterances), the pairwise loss function is defined as:

$$L_i^{pair} = -\log \frac{e^{\text{sim}(t_i,u_i)/\tau_{pair}}}{\Sigma_{j=1}^N e^{\text{sim}(t_i,u_j)/\tau_{pair}}} \quad (3)$$

where τ_{pair} is temperature for the pairwise representation.

Finally, combining losses L_i^t, L_i^u, L_i^{pair} defined in Eq. 1, 2, 3, our training loss is the following :

$$L_i^{train} = L_i^t + \lambda^u L_i^u + \lambda^{pair} L_i^{pair} \quad (4)$$

where λ^u, λ^{pair} are hyperparameters to scale the importance of utterance and pairwise learning.

3.3 Semantic Compression

In addition to the training procedure in Section 3.2, we introduce a new modification for inference as an instrument to examine our hypothesis about the semantic correlation between templates and utterances. Specifically, we measure how much it is possible to compress the hyperspace towards superior performance in a semantically interpretable process. The optimal value of compression coefficient λ^{comp} denotes the semantic well-formedness of the representations.

Our inference method of semantic compression is as follows: rather than just producing utterance representation as an inferred result, we introduce a scaled template representation term. This method augments the performance of the model, in addition

Model Type	SNIPS	ATIS	MASS.	HWU64	Clinc150	Average
BERT	80.00	78.05	41.86	50.84	33.35	56.82
SimCSE	91.71	85.67	76.77	81.08	71.00	81.25
SimCSE (ours)	92.00	86.56	77.27	80.24	71.05	81.42
TOD-BERT	90.71	81.75	58.47	63.25	50.60	68.96
TOD-BERT (ours)	91.00	81.63	59.92	61.33	51.11	69.00
DSE	95.86	87.01	76.77	79.28	70.16	81.82
DSE (ours)	95.86	84.66	73.50	76.75	68.51	79.86
TaDSE	**97.00**	**89.70**	78.18	**82.77**	70.56	83.64
TaDSE w/ MLP	96.29	89.14	**79.15**	82.29	**72.49**	**83.87**

Table 2: Unsupervised sentence embedding performance on intent classification task. "ours" models are trained with our augmented training set. More comments on the evaluations with dialogue datasets are in Appendix B. Comparison with supervised black-box embeddings in Section 5.5.

to functioning as a tool to find optimal λ^{comp} on separate validation set. This results in the explicit inclusion of salient anchor representation with the new representation form, via which we enhance specific semantics in the templates. We show the effect in Fig. 4. New compressed representation $repr_i$ is as follows :

$$repr_i = \lambda^{comp} t_i + (1 - \lambda^{comp}) u_i \quad (5)$$

where λ^{comp} is relative importance of template representation with range $0 \leq \lambda^{comp} \leq 1$.

4 Experimental Setup

We experiment with transfer learning on top of SimCSE (Gao et al., 2021) BERT-base model, as to influence expected TaDSE properties on a representation model. We utilize kNN with the training set to select relevant reference vectors (Fig. 7) and compute intent detection accuracy (Section 2). Baselines of TOD-BERT (Wu et al., 2020a) and DSE (Zhou et al., 2022) are dialogue embedding models utilizing utterance-only contrastive learning. We do not perform STS evaluation due to domain mismatch and lack of context-aware semantics (Appendix B). More details in Appendix C.

5 Results

5.1 Main Results

We report the results of unsupervised learning evaluation in Table 2 and Table 4. We discover that our models consistently outperform other unsupervised learning embeddings. In particular, we observe a 5 - 6% performance increase for SNIPS and ATIS datasets over the baseline. This is in line with the observation regarding augmentation stability in Section 5.2. In addition, we find that augmenting template representation with a trainable MLP layer achieves similar performance. This is in line with observation in (Gao et al., 2021) where inference with or without MLP achieve comparable performance (more experiments in Table 4, Fig. 8).

Data	Loss	Orig.	top-3	top-4	top-5
SNIPS	L^u	91.71	93.29	93.00	93.29
	L^{pair}	91.71	93.71	95.14	**96.14**
ATIS	L^u	85.67	86.00	N/A	N/A
	L^{pair}	85.55	**89.59**	N/A	N/A
MASS.	L^u	77.00	77.37	77.23	76.36
	L^{pair}	77.30	**79.39**	79.29	78.41
CLINC	L^u	71.05	70.98	70.47	69.62
	L^{pair}	71.27	72.25	72.73	**72.98**

Table 3: Template augmentation performance with single-source data. Note that ATIS reports top-2 instead of top-3, as we do not perform augmentations of higher order due to a large utterance count (Table 1).

5.2 Augmentation Stability

We experiment with increased k value in regards to the template-based augmentation process described in Section 3.1 (Table 3). Each source datasets exhibit different characteristics in regard to augmentation - for example, the performance of SNIPS, ATIS models increases substantially with the higher order of augmentation (augmentation-stable), while MASSIVE models decrease after 3. We detail this behavior in terms of stability regarding slot augmentations - while augmenting the

templates with different entities assists in creating new salient utterances, the process may be compromised if sample-specific slot values are configured in the non-relevant templates. Thus, we assert that template and slot quality is important for token-based augmentation methods.

An interesting observation here is CLINC150 dataset, which we augment via a simple NER-based automatic slot-filling method instead of manual annotations. The process results in a highly noisy Slot Book specific to the dataset as described in Appendix A, thus as expected the baseline utterance-only performance drops with higher-order augmentations. Interestingly, in contrast, L^{pair} models seem augmentation stable. This outcome inspires us to independently judge *"slot correctness"*[3] and *"template quality"*[4] - TaDSE is able to consider template quality in addition to slot correctness, while the utterance-only method would be greatly affected by low slot correctness and subsequent unnatural utterances. We leave further quantification of the observed behavior to future work.

5.3 Pairwise Training

To study the effect of different losses introduced in Section 3.2, we perform experiments with single-source augmented datasets and report ablation results per selected losses (Table 4). Interestingly, the inclusion of template loss itself enhances the performance of the representations, showing the importance of salient semantic information stored in templates. The inclusion of pairwise loss further enhances performance, showing that training the models to distinguish utterance-template pairs enables models to learn superior representations.

We emphasize how augmenting templates with plausible utterance values unlocks TaDSE training, as augmented synthetic data increases utterances per template. The extra utterance-template pairs assist in the learning of discrimination capability. Results in Table 3 show that the performance gap between TaDSE and the baseline method appears consistently with higher-order data augmentations.

5.4 Semantic Structure

We propose a semantic structure interpretation of the experimental results presented in Section 5.3. We assert that pairwise representation loss (Eq. 3) brings utterance and template representations closer, enhancing semantic distances within utterance sub-cluster correlated with a template.

To examine the hypothesis, our semantic compression test (Section 3.3) estimates how much we can enhance the aforementioned cluster properties by adjusting the hyperspace in a semantically interpretable way. Table 5 reports that the test succeeds with augmentation-stable (Section 5.2) datasets of SNIPS and ATIS[5], while non-stable datasets of MASSIVE and CLINC150 yield inconclusive results. This outcome shows the value of semantic structure interpretation and augmentation stability.

5.5 Supervised Embeddings

We compare TaDSE with state-of-the-art commercial black-box and open-source embedding models (Table 6) to contextualize our results beyond dialogue-specific baselines. We evaluate OpenAI `text-embedding-3-small`

Model	SNIPS	ATIS	MASS.	CLINC
w/o aug	91.71	85.67	77.00	71.05
aug	93.29	86.00	77.37	70.98
$+L^t$	95.29	88.47	78.58	71.53
$+L^t$, L^{pair}	96.14	**89.59**	79.39	72.98
$+L^{t'}$, L^{pair}	**97.00**	88.69	**79.83**	**73.45**

Table 4: Pairwise training experiments for single-source models. The baseline is non-augmented original data trained via SimCSE method (L^u). Other models are trained with our augmented data, with model column depicting new losses.

Model	SNIPS	ATIS	MASS.	CLINC
$+L^t$	95.29	88.47	78.58	71.53
S.Comp.	**95.86**	88.47	77.61	**71.62**
$+L^t$, L^{pair}	96.14	89.57	79.39	72.98
S.Comp.	**96.43**	**90.03**	78.24	72.25
$+L^{t'}$, L^{pair}	97.00	88.69	79.83	73.45
S.Comp.	**97.29**	**89.36**	77.47	72.71

Table 5: Semantic compression (S.Comp.) test for single-source data models, with model column depicting new losses with regards to L^u. The test succeeds on augmentation-stable datasets of SNIPS and ATIS.

[3]Criteria for slot correctness would be: How granular are slots? Are right values assigned to correct slots?

[4]Criteria for template quality would be: How many natural utterances would share templates? Are all salient entities identified and replaced?

[5]The test succeeds with $\lambda^{comp} = 0.1$ or 0.2, and $\lambda^{comp} = 0.5$ is not selected for any of datasets. We leave experiments with continuous λ^{comp} values to future work.

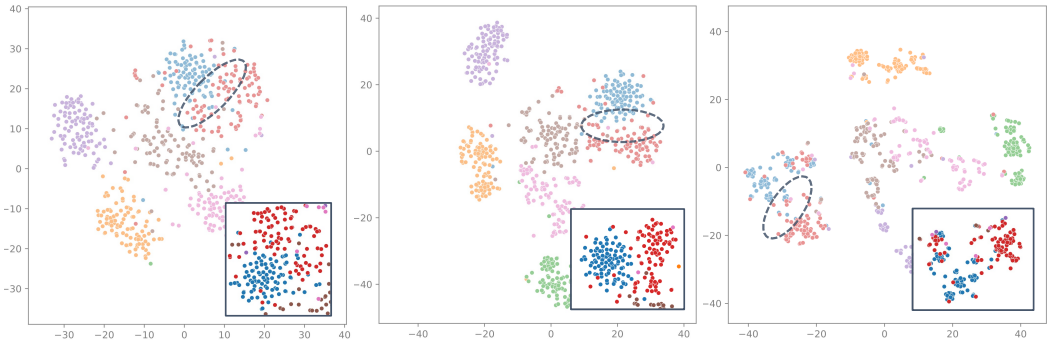

Figure 5: T-SNE diagram for SNIPS models, left : SimCSE, middle : TaDSE, right : TaDSE-compressed 0.5. Embeddings are color-coded according to their labels, with red, blue colored embeddings being representations with PlayMusic, AddToPlaylist labels. We circle the increased sparcity near the effective decision boundaries and show a magnified view at lower right. Note that more compression does not always result in better performance. ATIS diagrams in Fig. 9, 10, 11.

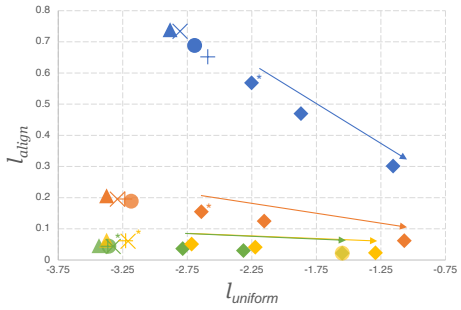

Figure 6: Uniformity / Alignment plot. Blue, orange, green, yellow are each ATIS, SNIPS, MASSIVE, CLINC models. Models for symbols are + : SimCSE, × : TaDSE w/ MLP, △ : utterance-only (L^u), ○ : TaDSE, ◇ : TaDSE-compressed. The arrows depict increasing λ^{comp}. Most performant models marked with an asterisk (*). Lower values are superior.

Model	SNIPS	ATIS	Avg
Qwen3-0.6B	92.14	89.14	90.64
OpenAI-small	97.86	84.88	91.37
OpenAI-large	**98.57**	84.77	91.67
Gemini-001	98.29	86.00	92.15
TaDSE	97.00	**89.70**	**93.35**

Table 6: Comparison with supervised black-box and open-source embeddings. Only TaDSE embeddings are unsupervised.

and `text-embedding-3-large`, Google `gemini-embedding-001` (Lee et al., 2025), and Qwen `Qwen3-Embedding-0.6B` (Zhang et al., 2025) on the SNIPS and ATIS benchmarks with dimension size of 768. Importantly, these models are known or likely to be *supervised* sentence embeddings trained on large-scale similarity-labeled datasets, in contrast to TaDSE which requires no supervision labels.

TaDSE achieves the highest average accuracy despite being the smallest model (110M parameters) and requiring no supervision labels, outperforming the best commercial black-box and open-source embedding models.

The per-dataset breakdown reveals a consistent pattern: commercial embeddings achieve near-ceiling performance on SNIPS, which contains only 7 intent classes (Appendix C Table 12) and relatively simple utterance structures, while TaDSE leads on ATIS by a substantial margin over the commercial baselines. ATIS utterances are characterized by compositional queries with complex syntactic structures.[6] We hypothesize that templates capture the compositional skeleton of such utterances in a way that surface-level supervised similarity training does not. This aligns with our augmentation stability findings (Section 5.2): both SNIPS and ATIS are augmentation-stable, yet ATIS is most sensitive to template-aware modeling, only requiring limited top-2 template augmentation compared to top-5 required by SNIPS to achieve similar performance increase of 5 - 6 %. This may be because its structural complexity enables richer semantic anchoring.

TaDSE operates at 110M parameters, approximately 5× smaller than Qwen3-Embedding-0.6B and orders of magnitude smaller than the infras-

[6] e.g.
1. *find me the earliest boston departure for atlanta and the lastest return trip from atlanta so that i can be in atlanta the longest amount of time but return to boston the same day*
2. *show me all flights from pittsburgh to boston both direct and connecting that depart pittsburgh after 7 pm*

tructure behind the OpenAI and Gemini embedding APIs. The result suggests that domain-specific structural priors of utterance-template pairwise associations can effectively substitute for large-scale supervised training data and model capacity when the target domain exhibits complex compositional patterns.

6 Analysis

6.1 Uniformity / Alignment

To identify inner workings of our methods, we utilize uniformity and alignment (Wang and Isola, 2020) to uncover how semantic structure of representations is altered. Definitions in Appendix E.

We display uniformity/alignment for our models in Fig. 6. Surprisingly, we find that uniformity and alignment for TaDSE models have an inverse correlation.[7] We also find that utterance-only (L^u) models have the worst alignment, which relates to the naive augmentation step (Fig. 1). Consequently, we report superior alignment and inferior uniformity for TaDSE models. This trade-off suggests that the performance increase may relate to superior alignment, especially in augmentation-stable (Section 5.2) datasets of SNIPS and ATIS.

Importantly, we identify that TaDSE models from semantic compression test (Section 3.3) obtain superior alignment correlated with higher λ^{comp} values. The results support our semantic structure interpretation in Section 5.4. We conclude that our semantic compression test correlate with existing tools of uniformity/alignment.

6.2 Qualitative Analysis

We graph a set of T-SNE diagrams[8] for our representations (Fig. 5, 9, 10, 11). We observe a clearer separation between music-associated clusters and a set of pronounced sub-clusters that correspond to semantic structure interpretation (Section 5.4, 6.1).

7 Conclusions

In this work, we propose TaDSE, a novel unsupervised representation learning method that produces high-quality semantic representations of task-oriented dialogue. We develop a template-based data augmentation strategy that synthetically supplies diverse utterance-template pairs. We present methods of learning utterance-template discriminative capability and pairwise association via a new training scheme. We further justify the inner workings of our methods by applying a novel inference instrument that aligns well with uniformity/alignment analysis and visualization of representations. Our paper is first to employ semantic information in dialogue templates toward dialogue embeddings. Our template-aware data augmentation strategy and training losses enrich any dialogue datasets' semantic content. We believe that TaDSE is a reinforced text encoder for dialogue system applications.

8 Limitations

An enhanced slot-filling method for Clinc150 could be applied, rather than the current functional baseline method (the other 4 datasets use provided annotations). For example, slot-filling could suffer from a disambiguation problem, of slot values in different dialogue scenarios classified as the same slots. We leave enhanced slot-filling method experiments to future work. However, we emphasize that our methods work with automatically generated noisy Slot Book on Clinc150 dataset, which we describe in detail (Section 5.2) in terms of *template quality*. It will be interesting to observe how other slot-filling methods and hand annotations differ in terms of representation quality.

Our methods, in line with existing literature (Zhou et al., 2022; Wu et al., 2020a), works with dialogue datasets of SNIPS (Coucke et al., 2018), ATIS (Hemphill et al., 1990), MASSIVE (FitzGerald et al., 2022), HWU64 (Liu et al., 2019) and CLINC150 (Larson et al., 2019). We actively avoid general datasets and evaluations such as STS (Appendix B). Future works may take inspiration from recent works in general embeddings (Zhang et al., 2023; Cheng et al., 2023; Wang et al., 2024a,b; Izacard et al., 2022) and apply to dialogue domain. See also Appendix B.

The data augmentation process could be further improved with regard to the diversity of resulting utterances. Our baseline augmentation method improves performance (Table 3). However, the number of slot values for each slot is limited in the current setting. A smart slot value selection strategy such as one incorporating retrieval or randomization could be implemented in future works.

While preprocessing model input[9], we did not

[7] This is a viable outcome considering that both uniformity and alignment are asymptotic of the same order to $\|f(a) - f(b)\|_2^2$, with distinct eligible representation pairs (a, b). (Gao et al., 2021) report similar trends with certain variations.

[8] Per default Scikit-learn configuration of 30.0 perplexity.

[9] After data augmentation.

utilize distinct slot tokens to emphasize the semantic structure aspect. Instead, we replaced them as one token "{SLOT}" in templates. We leave it to future work to identify how discernible slot tokens in model inputs relate to representations.

It will be interesting to perform contrastive learning with token-level annotated "anchors" (Section 3.2) in other domains, such as legal or medical documents, and evaluate on such domains. Cross-domain evaluations can also be performed, with the model tuned on a domain that's different from selected evaluation dataset. We leave it to future work.

We experiment with discrete λ^{comp} values for semantic compression. While we observed sufficient trends for datasets with certain characteristics (Table 5, Fig. 6), we leave it to future work to perform semantic compression with continuous λ^{comp} values for follow-up evaluation. The test could also be theoretically expanded upon.

Our work only experiments with English, thus there is a potential risk of enhancing overexposure to the English language and its token-wise semantic characteristics, especially since we perform token-wise replacements and augmentations to the datasets. Templates that correspond to specific tokenization strategies might be necessary for experiments in other languages.

References

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Sergio Burdisso, Srikanth Madikeri, and Petr Motlicek. 2024. Dialog2Flow: Pre-training soft-contrastive action-driven sentence embeddings for automatic dialog flow extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5421–5440, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

Qinyuan Cheng, Xiaogui Yang, Tianxiang Sun, Linyang Li, and Xipeng Qiu. 2023. Improving contrastive learning of sentence embeddings from AI feedback. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11122–11138, Toronto, Canada. Association for Computational Linguistics.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan,

Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

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.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Chih-Wen Goo, Guang-Lai Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung (Vivian) Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *North American Chapter of the Association for Computational Linguistics*.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

E. Haihong, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Annual Meeting of the Association for Computational Linguistics*.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning.

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.

Mihir Kale and Abhinav Rastogi. 2020. Template guided text generation for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6505–6520, Online. Association for Computational Linguistics.

Young-Bum Kim, Dongchan Kim, Joo-Kyung Kim, and Ruhi Sarikaya. 2018. A scalable neural shortlisting-reranking approach for large-scale domain classification in natural language understanding. 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 16–24, New Orleans - Louisiana. Association for Computational Linguistics.

Jason Krone, Yi Zhang, and Mona Diab. 2020. Learning to classify intents and slot labels given a handful of examples. *arXiv preprint arXiv:2004.10793*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. 2025. Gemini embedding: Generalizable embeddings from gemini.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Han Li, Sunghyun Park, Aswarth Dara, Jinseok Nam, Sungjin Lee, Young-Bum Kim, Spyros Matsoukas, and Ruhi Sarikaya. 2021. Neural model robustness for skill routing in large-scale conversational ai systems: A design choice exploration.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Thirty-sixth Conference on Neural Information Processing Systems, NeurIPS 2022*.

Bing Liu and Ian R. Lane. 2015. Recurrent neural network structured output prediction for spoken language understanding.

Bing Liu and Ian R. Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *ArXiv*, abs/1609.01454.

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

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online). International Committee on Computational Linguistics.

G. Mesnil, Xiaodong He, L. Deng, and Y. Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. pages 3771–3775.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-aware contrastive learning of sentence embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3870–3885, Seattle, United States. Association for Computational Linguistics.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Conference on Empirical Methods in Natural Language Processing*.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. 2020. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

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

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models.

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.

Wei Wang, Liangzhu Ge, Jingqiao Zhang, and Cheng Yang. 2022. Improving contrastive learning of sentence embeddings with case-augmented positives and retrieved negatives. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2159–2165, New York, NY, USA. Association for Computing Machinery.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

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

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020b. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling.

Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. Few-shot intent classification and slot filling with retrieved examples. *arXiv preprint arXiv:2104.05763*.

Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew O. Arnold. 2022. Virtual augmentation supported contrastive learning of sentence representations.

Junlei Zhang, Zhenzhong Lan, and Junxian He. 2023. Contrastive learning of sentence embeddings from scratch. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3932, Singapore. Association for Computational Linguistics.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *International Joint Conference on Artificial Intelligence*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew O Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. *NAACL*.

A CLINC150 Slot Book

We only report top-5 occurrences for the allotted space. We experiment with the SpaCy NER model "*en_web_core_lg*" on the Clinc150 dataset. Please note that this is a noisy but effective baseline only for the Clinc150 dataset (discussion in Section 8), required due to lack of annotations. The first table denotes slots. The second and 3rd tables are examples of well-formed slot books. The 4th and 5th tables are examples of noisy slot books - the 4th one is a combination of card companies, retirement funds, and bank names, and the 5th one is a combination of airlines, continents, and sightseeing locations. All categories should be separated for good *slot correctness* (Section 5.2). Note that even if the wrong slot is not in top-k frequency, it still causes noisiness due to associated slots filled by incorrect values.

Slot	Count
GPE	1397
DATE	1194
ORG	844
CARDINAL	594
TIME	443

Table 7: Slots.

Slot Value	Count
french	49
italian	28
spanish	23
british	23
mexican	14

Table 8: NORP slot values.

Slot Value	Count
first	21
5th	11
second	8
3rd	8
4th	7

Table 9: ORDINAL slot values.

Slot Value	Count
mastercard	58
401k	49
bank of america	39
chase	39
american express	37

Table 10: ORG slot values.

Slot Value	Count
delta	19
africa	14
europe	7
asia	6
the grand canyon	3

Table 11: LOC slot values.

B Dialogue Embedding Evaluations

In line with prior publications on dialogue sentence embeddings (Zhou et al., 2022; Wu et al., 2020a; Liu et al., 2021; Burdisso et al., 2024), we do not evaluate on general benchmarks such as (Muennighoff et al., 2023; Conneau and Kiela, 2018; Thakur et al., 2021) or STS (Cer et al., 2017). This is in contrast to general sentence embeddings (Zhang et al., 2023; Cheng et al., 2023; Wang et al., 2024a,b; Izacard et al., 2022). The reason being:

"Here we do not adopt the standard semantic textual similarity (STS) task for two reasons: (1) The sentence embedding performance varies greatly as the domain of the training data changes. As a dialogue dataset is always about several certain domains, evaluating on the STS benchmark may mislead the evaluation of the model. (2) The dialogue-based sentence embeddings focus on context-aware rather than context-free semantic meanings, which may not be suitable to be evaluated through the context-free benchmarks." (Liu et al., 2021)

C Configuration

Src. Data	Test	Slots	Intents
ATIS	893	129	26
HWU64	1076	54	64
SNIPS	700	53	7
MASSIVE	2974	55	60
CLINC150	5500	17 (aug.)	150

Table 12: Statistics of original source dialogue datasets. Note that the slot count in CLINC150 is from our slot-filling augmentation.

We perform transfer learning on top of the SimCSE (Gao et al., 2021) BERT-base (110M params) model (unsup-simcse-bert-base-uncased), as our purpose is to evaluate dialogue-specific effects. Table 2 uses augmented data from all sources (SNIPS / ATIS / MASSIVE / HWU64 / CLINC150) while other tables experiment with single-source data. For "ours" models in Table 2, transfer learning is performed on corresponding checkpoints from each publication using our augmented data and SimCSE loss (L^u-only). For baselines in other tables, we experiment with the same BERT-base variants. We utilize a low learning rate of $1e-8$ and train for 2 epochs for all models. The only exception is considering 8 epochs for single-source MLP pooled models to learn the W_A properly (Fig. 8). Batch size 16. We experiment with $\lambda^t \in \{1.0, 0.0\}$, $\lambda^u \in \{1.0, 0.0\}$ and $\lambda^{pair} \in \{0.5, 0.0\}$ and $\lambda^{comp} \in \{0.1, 0.2, 0.5\}$. We select best λ^{comp} per configuration according to the validation set. We perform experiments on RTX 3090. We only use NLP research tools as they are intended for research purposes. Public dialogue datasets do not contain identifiable information.

We evaluate with kNN to select most relevant reference vector to extract intent information ($k = 1$, experiment in Fig. 7). We assert that our choice of evaluation method (kNN) emphasizes the local structure of the representation space in contrast to the global structure. This is fitting for our approach as a template representation may reside close to utterance representation, affecting local structure more. We evaluate with full utterance/template training set representations and we include CLINC150 OOS labels.

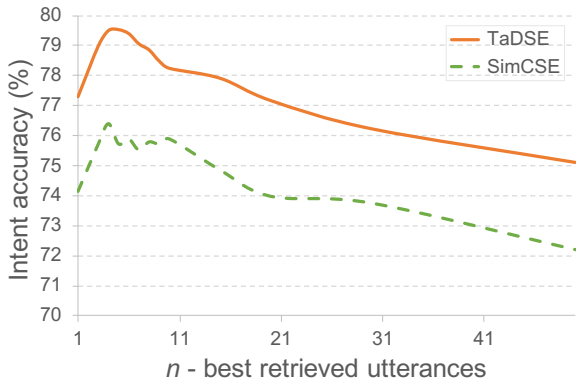

Figure 7: MASSIVE performance with TaDSE and SimCSE. The horizontal axis is the k value in kNN.

D Training Stability

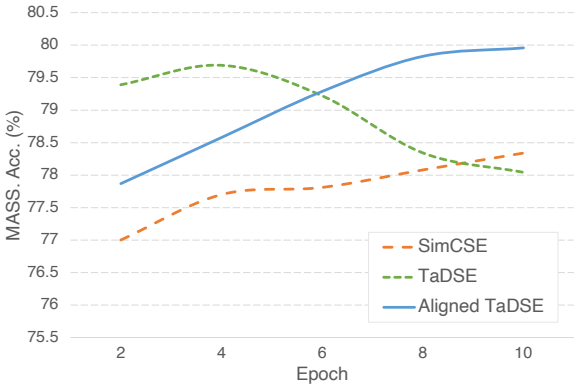

Figure 8: Performance on test set during 10 training epochs. 'Aligned' model is the 'w/ MLP' variant.

We report separate evaluations with the inclusion of MLP layer W_A, which improve performance on certain datasets (Table 2, $L^{t'}$ in Table 4). As the MLP layer needs to be trained from scratch, we empirically require more training epochs than non-MLP TaDSE in our experiments (Fig. 8, Section C). In contrast, non-MLP TaDSE performance is optimal at a lower epoch.

E Uniformity / Alignment Definition

We compute uniformity/alignment per test set of each source data and define p_{pos} as representations within the same label, and p_{data} as the sentences from each original non-augmented source dataset.

Uniformity is a measurement of the degree of uniformness of the representations :

$$\ell_{uniform} \triangleq \log \underset{x,y \overset{i.i.d.}{\sim} p_{data}}{\mathbb{E}} e^{-2\|f(x)-f(y)\|_2^2} \tag{6}$$

Alignment measures the distance between positive representations :

$$\ell_{align} \triangleq \underset{(x,x^+)\sim p_{pos}}{\mathbb{E}} \|f(x) - f(x^+)\|_2^2 \tag{7}$$

F ATIS T-SNE Diagrams

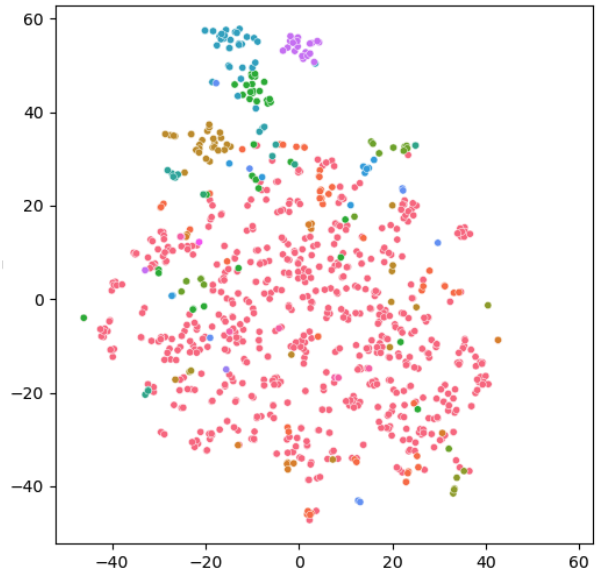

Figure 9: T-SNE diagram for ATIS representation hyperspace from SimCSE model, trained with our data.

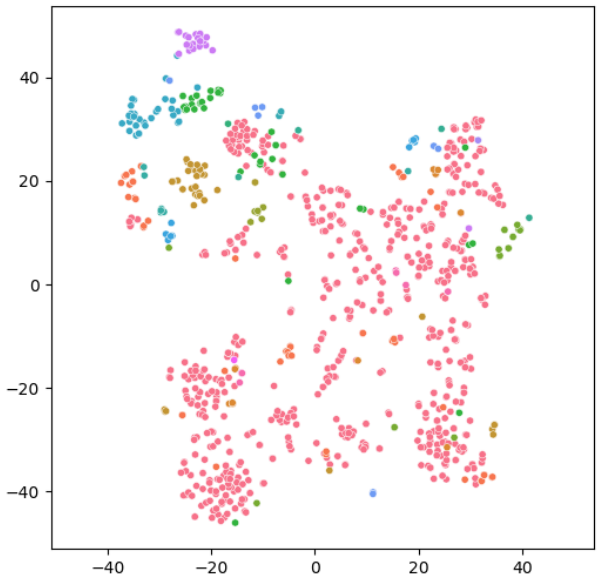

Figure 10: T-SNE diagram for ATIS representation hyperspace from TaDSE model, trained with our data.

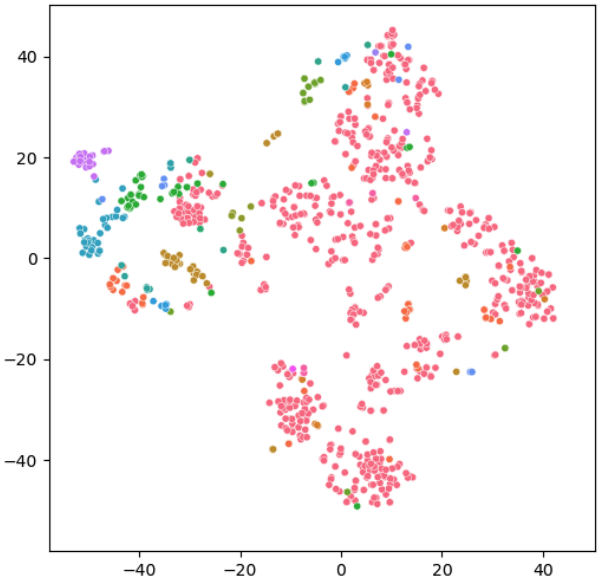

Figure 11: T-SNE diagram for ATIS representation hyperspace from our optimal TaDSE model ($\lambda^{comp} = 0.2$).