Template-augmented Dialogue Representation Learning

Minsik Oh, Guoyin Wang*, Taiwoo Park, Puyang Xu

Alexa AI
Amazon

Abstract

Learning high-quality sentence embeddings from dialogue has drawn increasing attention as it is essential to solving various dialogue-oriented tasks with low annotation costs. However, directly gathering utterance relationships from conversations are difficult, while token-level annotations, *e.g.*, entities, slots, and templates, are much easier to obtain. General sentence embedding methods are based on sentence-level self-supervised frameworks and cannot utilize token-level extra knowledge. In this paper, we introduce a new dialogue utterance embedding framework, **T**emplate-**a**ugmented **D**ialogue **S**entence **E**mbedding (TaDSE). This novel method utilizes template information to learn utterance representation effectively via a self-supervised contrastive learning framework. TaDSE augments each sentence with its corresponding template and then conducts pairwise contrastive learning over both sentence and template. We evaluate TaDSE performance on two downstream benchmark datasets. The experiment results show that TaDSE achieves significant improvements over previous SOTA methods. We further analyze the representation quality and show improved alignment and boosted local structure in semantic representation hyperspace.

1 Introduction

Learning sentence embeddings from dialogue has recently attracted increasing attention (Zhou et al. 2022; Liu et al. 2021). Learning high-quality dialogue semantics (Hou et al. 2020; Krone, Zhang, and Diab 2020; Yu et al. 2021) helps solve various downstream tasks, especially in the scenarios with limited annotations (Snell, Swersky, and Zemel 2017; Vinyals et al. 2016; Kim et al. 2018; Li et al. 2021). Due to the fast development of contrastive learning (Chen et al. 2020a; He et al. 2020; Hjelm et al. 2018; Radford et al. 2021; Chen et al. 2020c), there has been a solid success in learning universal sentence representations in both supervised (Reimers and Gurevych 2019) and unsupervised manner (Gao, Yao, and Chen 2021; Chuang et al. 2022; Giorgi et al. 2021; Nishikawa et al. 2022; Jiang et al. 2022). However, universal sentence embeddings usually achieve undesirable performance in the dialogue domain (Zhou et al. 2022) since the relationship between utterances is not well utilized. This leads the universal sentence embedding models to over-index on cosmetic similarity and does not capture the cluster properties of utterances as shown in Figure 1.

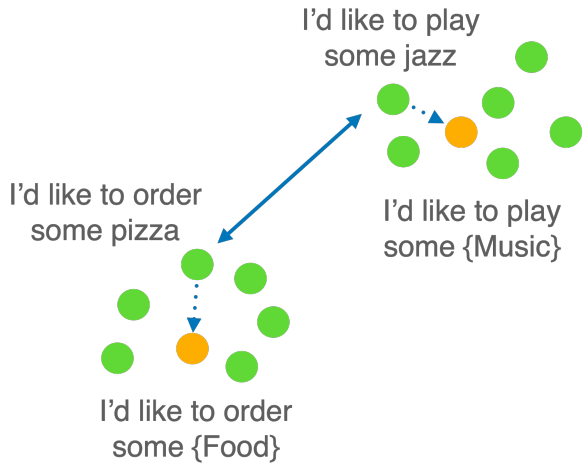

Figure 1: Illustration of the relationship between templates and corresponding utterances in semantic space. The orange dot represents a given template in the semantic space and the green dots represent its corresponding utterances. Ideally, the utterances sharing the same template should be closer and clustering over the template in the semantic space. On the contrary, utterances with different templates should be separated.

In this paper, we explore how we can create semantically relevant sentence embeddings for dialogues. Pattern templates 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 hand-annotated to be a variable representation of text structure and salient entity values. Thus, we extract salient domain information from the

*Corresponding author.
Please correspond to {ohtrent, guoyiwan}@amazon.com.

Accepted to Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI 2023.

template and its pair relation with matching utterances via representation learning methods. We present the Template-augmented Dialogue Sentence Embeddings (TaDSE) generation framework which produces superior text embeddings for dialogue understanding via both unsupervised training and inference.

Our TaDSE training method encodes auxiliary template representations and their pairwise relationships with matching utterance representations. We introduce a pair of loss terms that discriminate the templates and the pairwise relationship in a contrastive manner. Our pairwise training outperforms previous utterance-only methods, even without learning utterance representations in conjunction with SimCSE utterance-only contrastive loss. In addition, we report the surprising performance of template-only representation learning.

Our TaDSE inference method exploits the pairwise relationship between the auxiliary template and matching utterance without further training. We balance representations of the auxiliary template and matching utterances to produce an enhanced representation. This induces compression of representation hyperspace to further benefit concrete semantics stored in templates. We experiment with variations of this method, including utterance-only scenarios without corresponding templates. Our TaDSE inference method enhances the performance of TaDSE-trained models by a consistent margin.

Finally, in an effort to clarify how our approach improvements operate, we look into the properties of TaDSE representations. We start our analysis with uniformity and alignment quatization (Wang and Isola 2020). We observe that the TaDSE inference stage consistently improves alignment, while the training stage slightly improves. We also find an inverse relationship between alignment and uniformity for our models, from which we form the hyperspace skew hypothesis as a foundation of our improvements. We propose that encoding pair relationship between utterances and auxiliary templates induces local compression to hyperspace that aligns with dialogue semantics. We discover that the hypothesis is supported by representation visualizations of TaDSE training and inference methods, which shows a multitude of distinctive local clusters.

2 Related Works

Unsupervised Semantic Representation methods train with contrastive objectives effectively to learn universal sentence embeddings. For vision, methods such as SimCLR (Chen et al. 2020b,c) have effectively demonstrated the performance of contrastive representation learning with data augmentation operations. In NLP, methods such as SimCSE (Gao, Yao, and Chen 2021) show that simple augmentation such as dropout masking can be considered effective positive representation targets. DiffCSE (Chuang et al. 2022) showed that auxiliary reconstruction loss works well with contrastive representation learning schemes with random masking. EASE (Nishikawa et al. 2022) introduces Entity representation contrastive loss for better performance. DeCLUTR (Giorgi et al. 2021) mark different spans of the same document as positive pairs. Our method differs from previously studied methods since we exploit pattern templates that pair the utterances for loss design and masking during training. In addition, we introduce a novel template-based inference method.

One of the critical components of modern dialogue understanding (SLU / NLU) systems is the template and slot information. They are a good source of linguistic variability in exact comparison settings; however, they are also known to be a good auxiliary feature for a Bi-LSTM shortlister model (Kim et al. 2018) or dense re-ranking model (Li et al. 2021). They are known to be applicable for entity prediction (Bastianelli et al. 2020) with a suggestion to utilize them for machine translation (FitzGerald et al. 2022). Rather than only summarizing task-specific industry applications as in previous papers, our method employs templates as auxiliary data for sentence representation learning, and studies how their pairwise relationship with utterances can be encoded via unsupervised methods.

Previous work showed that text data augmentation is effective in classification tasks, with random insertions or deletions (Wei and Zou 2019) and automatic compositional policy search (Ren et al. 2021) among the methods explored. Our study is different in that rather than random augmentations performed towards universal embedding improvements, we perform semantic augmentations relevant to dialogue structure, in which semantic information is provided by template and slot annotations. In terms of augmentation operations, we perform the token-level masking augmentation technique in contrast to insertions or deletions. Our method is designed to enhance concrete dialogue semantic information stored in templates and slots, rather than a variety of permutations as in previous studies.

3 Proposed Method

3.1 Template Paired Dataset

We assign Template, Intent, and Skill for each spoken dialogue utterance. Utterance and Template work as inputs to the model. Intent and Skill are labels that correspond to spoken dialogue semantics. We gather 6.7M distinct rows during a consecutive 28 weeks period. No personally identifiable information was used in experiments. Detailed description in Appendix B.

$$\begin{aligned} X_i &= (T_i, U_i) \\ Y_i &= (I_i, S_i) \\ D_i &: \{X_i, Y_i\} = \{(T_i, U_i), (I_i, S_i)\} \end{aligned} \tag{1}$$

where X_i and Y_i notate inputs and labels, D_i notates sample, U_i is utterance, T_i is paired template, I_i is intent, and S_i is skill. Only X_i is utilized for unsupervised training, while Y_i is exclusively used for evaluation. We find that skill is too granular of criteria as to determine the semantics of an utterance within an evaluation set. Thus, We experiment with intent.

3.2 Pairwise Modeling

We introduce a new concept of "pairwise anchoring", where the representation of auxiliary data (template) is trained in

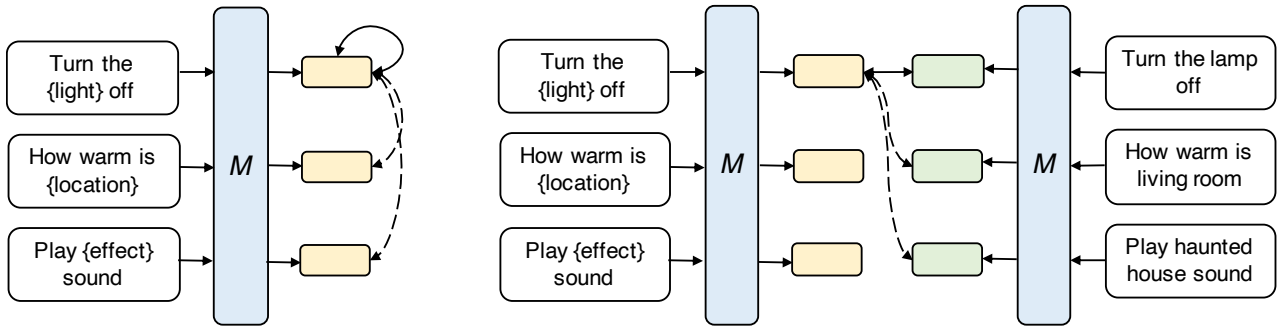

Figure 2: We show our template contrastive learning methods in this diagram. The first diagram displays template contrastive learning (L_i^t), while the second diagram displays pairwise contrastive learning with utterance negatives ($L_i^{pair_u}$). M represents the embedding generation model and yellow, and green represent template and utterance representations respectably. Solid lines designate positive pairs and dotted lines designate negative pairs when connecting representations.

tandem with the paired sentence (utterance) via an unsupervised representation learning method while teaching the capability to distinguish pairwise relationship via contrastive learning (Figure 2).

First, we define **template representation loss**, where we encourage the model to learn template representation such that we have a spoken language relevant anchor with which we further induce utterance representations. We process the template with the tokenization strategy selected from Appendix A and train with contrastive loss. We apply the dropout technique from (Gao, Yao, and Chen 2021) to obtain positive representations.

$$L_i^t = -\log \frac{e^{sim(t_i, t_i^+)/\tau_p}}{\Sigma_{j=1}^N e^{sim(t_i, t_j^+)/\tau_p}} \tag{2}$$

where t_i is template representation, t_i^+ is template representation dropout variant, τ_p is temperature hyperparameter for the template representation, and $sim(t_i, t_j)$ is cosine similarity $\frac{t_i^T t_j}{||t_i|| \cdot ||t_j||}$. While we utilize the template provided with the dataset in our experiments, such templates can be also manufactured in a heuristic manner. We leave this to future research.

Next, we compute **utterance representation loss** similarly in a contrastive manner. This is to ensure we correctly learn utterance representation without over-reliance on template representation.

$$L_i^u = -\log \frac{e^{sim(u_i, u_i^+)/\tau_u}}{\Sigma_{j=1}^N e^{sim(u_i, u_j^+)/\tau_u}} \tag{3}$$

where u_i is utterance representation, u_i^+ is utterance representation dropout variant, τ_u is temperature hyperparameter for the utterance representation, and $sim(u_i, u_j)$ is cosine similarity.

Finally, we introduce **pairwise representation loss**, where we bring utterance and template representations from the same pair closer via contrastive loss. We introduce 2 variants of the loss depending on the negative selection strategy.

1. Compare against template negatives.

$$L_i^{pair_t} = -\log \frac{e^{sim(u_i, t_i)/\tau_{pair_t}}}{\Sigma_{j=1}^N e^{sim(u_i, t_j)/\tau_{pair_t}}} \tag{4}$$

2. Compare against utterance negatives.

$$L_i^{pair_u} = -\log \frac{e^{sim(t_i, u_i)/\tau_{pair_u}}}{\Sigma_{j=1}^N e^{sim(t_i, u_j)/\tau_{pair_u}}} \tag{5}$$

We define pairwise representation loss from Eq. 4, 5:

$$L_i^{pair} = L_i^{pair_t} \quad \text{or} \quad L_i^{pair_u} \tag{6}$$

Finally, our training loss is following :

$$L_i^{train} = L_i^t + \lambda^u L_i^u + \lambda^{pair} L_i^{pair} \tag{7}$$

where λ^u and λ^{pair} are hyperparameters to scale importance of utterance and pairwise learning. L_i^t, L_i^u, L_i^{pair} are defined in Eq. 2, 3, 6 respectively. We do not introduce a hyperparameter for template loss to emphasize the pairwise anchoring effect. Empirically, we find that test performance does not always increase with bigger λ^u.

3.3 Inference Scaling

In addition to the training procedure in Section 3.2, we introduce a new modification for inference. Rather than just producing utterance representation as an inferred result, we introduce a scaled template representation term. We intend the inclusion of domain-adjacent anchor representation with the new representation form. We show the effect in Figure 3. We take influence from recent studies on multi-modality representations (Liang et al. 2022; Radford et al. 2021) where it is shown that representation of multiple modalities each map to distinct narrow cones in hyperspace, as such we hypothesize relatively separate utterance and template representation clusters which we balance via our method.

$$repr_i = \lambda^{infer} t_i + (1 - \lambda^{infer}) u_i \tag{8}$$

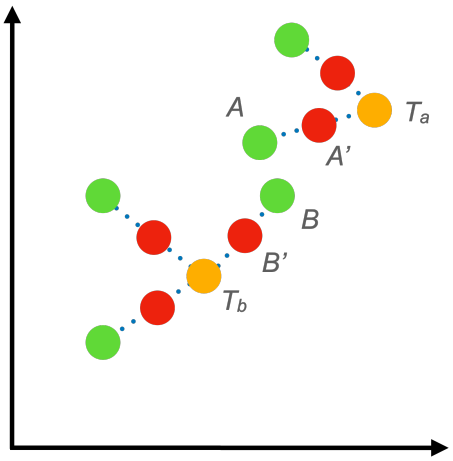

Figure 3: We demonstrate effect of inference scaling method (Section 3.3). Green, orange points represent input utterance, template representations respectably. Red points represent resulting scaled utterance representations in the hyperspace. We expect this adjustment to compress the local cluster of utterance representations at a granular level.

where λ^{infer} is relative importance of template representation with range $0 \leq \lambda^{infer} \leq 1$. A relevant template could be easily paired via heuristics or similarity computation for utterance-only scenarios, especially for the 1Slot variant (Appendix A). Other token-based augmentation methods could be also developed for such scenarios.

4 Experimental Setup

We train on NVIDIA V100 GPU and evaluate with AWS R5 instances. More details on our configurations can be found in Appendix C. We evaluate with a fine-grained NLU evaluation method for spoken dialogue semantics, which is described in Appendix E.

5 Results

In our experiments, models only trained with pairwise modeling method (Section 3.2) is marked **TaDSE-train**, while models also inferred with inference scaling method (Section 3.3) are marked **TaDSE-infer**. $L_i^t + L_i^u + L_i^{pair_u}$ is our representative model. We notate our data set (Section 3.1) as **TPD** (**T**emplate **P**aired **D**ataset).

Evaluation results are available in Table 1. We first show the importance of domain-specific training data by comparing TPD utterance-only trained models with external Wiki data trained models. The results suggest the strong effectiveness of TaDSE methods and the pairwise anchoring effect.

5.1 Pairwise Modelling

We present ablations of our pairwise contrastive learning loss in Table 2. We find that inclusion of L_i^t, L_i^{pair}, L_i^u losses each enhances performance. We also find that $L_i^{pair_u}$ is slightly more performant than $L_i^{pair_t}$ in test set, likely due to availability of more valid negatives in the batch leading to increased generalization capability.

Model	Valid	Test
$\text{SimCSE}_{\text{Wiki}}$	58.98	60.85
$\text{DiffCSE}_{\text{Wiki}}$	52.24	53.93
$\text{SimCSE}_{\text{utt-only}}$	60.24	62.10
$\text{DiffCSE}_{\text{utt-only}}$	55.98	57.46
$\text{DSE-BERT}_{\text{base}}$	15.88	16.18
TaDSE-train	**76.26**	**80.93**
TaDSE-infer	**82.04**	**82.26**

Table 1: Semantic representation performance on NLU evaluation benchmark (Section 4). "Wiki" variants are models trained with the "wiki1M" dataset provided with SimCSE according to configurations described in respective papers. "utt-only" variants are model architectures trained only using utterances in our TPD training set. We discuss low performance of $\text{DSE-BERT}_{\text{base}}$ checkpoint in Section 6.3.

Model	Valid	Test
L_i^t only (unique)	68.31	69.31
$L_i^t + L_i^{pair_t}$	76.29	77.02
$L_i^t + L_i^u + L_i^{pair_t}$	**77.04**	77.86
$L_i^t + L_i^{pair_u}$	76.16	77.23
$L_i^t + L_i^u + L_i^{pair_u}$	76.26	**80.93**

Table 2: Performance for different loss variants in Section 3.2. 'L_i^t only (unique)' model is trained with unique template data, duplicates are removed. Training configurations for each model is selected via process described in C.

Moreover, interesting fact is that training only with few-shot template data (details in Appendix F) increases the performance of the model significantly compared to SimCSE baseline only trained on utterances (Table 1). We hypothesize that it is significantly easier for the model to form a distribution about how the important keywords relate to the labels when trained on unique template data. Keywords may include "turn", "on", "off", "horoscope", "play", "sound" (template samples in Table 9). We further posit that due to salient masking available in template data, an effect akin to "blurring the image" (Park and Kim 2022) is occurring, resulting in better representation via enhancing the keywords in relation to their relative positions. We leave this interesting observation to future research.

5.2 Inference Scaling

We perform inference scaling with various backbone models and report the results in Table 3. We observe consistent performance increase across the evaluation sets. The results suggest consistent effectiveness of inference scaling method.

We further perform inference scaling with different backbone models for each utterance and template representation. The results are available in Table 4. We find that even with

Model	Valid	Test
L_i^t only (unique)	68.31	69.31
Inference Scaling	**74.40**	**74.73**
$L_i^t + L_i^{pair_t}$	76.29	77.02
Inference Scaling	**82.51**	**82.65**
$L_i^t + L_i^u + L_i^{pair_t}$	77.04	77.86
Inference Scaling	**82.14**	**82.38**
$L_i^t + L_i^{pair_u}$	76.16	77.23
Inference Scaling	**82.03**	**82.10**
$L_i^t + L_i^u + L_i^{pair_u}$	76.26	80.93
Inference Scaling	**82.04**	**82.26**

Table 3: Performance for inference scaling method (Section 3.3). We infer utterance and template representations using same model in this table. λ^{infer} for each model is selected via process described in Appendix D.

Model Pair	Test
Baseline (L_i^u)	62.10
L_i^u & L_i^t	67.40
L_i^u & $L_i^t + L_i^u + L_i^{pair_u}$	**70.45**
Baseline ($L_i^t + L_i^u + L_i^{pair_u}$)	80.93
$L_i^t + L_i^u + L_i^{pair_u}$ & L_i^t	**81.71**

Table 4: Performance for inference scaling method (Section 3.3) with different utterance and template representation models. λ^{infer} for each model is selected via process described in Appendix D.

unrelated hyperspaces resulting from different models, our effort of merging representations is constantly effective and achieves notable performance increases across variants.

6 MASSIVE Experiments

6.1 Metadata Extraction Method

As to evaluate our representation learning methods in downstream tasks such as MASSIVE (FitzGerald et al. 2022), We present a metadata extraction method that we use to extract corresponding metadata from similar training utterances (Figure 4). This is similar to REINA (Wang et al. 2022), which is a methodology that first retrieves training data that are similar to input text and provide both as input to a given task-performing model. While we do not use a task-performing model at the end, we filter retrieved training data on inference time to create a high-quality reference set. We create a neural index via inference with the TaDSE model on the training set and search it during inference time.

6.2 Experimental Setup

We train our model with MASSIVE training data and evaluate on test set data for the intent classification task. We work with the en-US locale. Because intent corresponds to spoken dialogue semantics, we consider this a suitable proxy task to evaluate the model and methodology. More details on Appendix I.

Model	Intent Accuracy (%)
Baseline (DSE)	11.60
Baseline (SimCSE)	74.14
Utterance only	76.90
Template only	76.97
$L_i^t + L_i^u + L_i^{pair_t}$	77.17
$L_i^t + L_i^u + L_i^{pair_u}$	77.30
Best ($n = 5$)	**79.52**

Table 5: Unsupervised performance on MASSIVE intent classification task with metadata extraction method and TaDSE trained models. Best model is based on $L_i^t + L_i^u + L_i^{pair_u}$ with n-best tuning. The models are trained on MASSIVE data according to the process described in Section 6.2. We discuss low performance of DSE-BERT$_{\text{base}}$ checkpoint in Section 6.3.

6.3 Results

We report the result of our experiments in Table 5. We find the good performance of the inference-only pipeline, with TaDSE models trained with template data achieving strong performance. We also report the results of ablation regarding n-best setup, in which we find that performance increase up to top-5 and stays relatively stable with higher n (Figure 5).

We compare our results with another dialogue-trained sentence representation model DSE (Zhou et al. 2022) and find surprisingly low performance. We posit that since the DSE model is trained on consecutive utterances in dialogue, it is optimized for NLI entailment or question-answer relationship. Correspondingly, the DSE paper reports lower performance of NLI supervised trained version of the model in certain dialogue tasks. We present some examples in Appendix J.

7 Analysis

7.1 Uniformity / Alignment

As to further analyze how our methods modify the representation hyperspace of utterances, we utilize key properties of uniformity and alignment (Wang and Isola 2020).

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

$$\ell_{\text{uniform}} \triangleq \log \underset{x,y \overset{i.i.d.}{\sim} p_{\text{data}}}{\mathbb{E}} e^{-2\|f(x)-f(y)\|_2^2} \tag{9}$$

Alignment measures the distance between positive normalized representations :

$$\ell_{\text{align}} \triangleq \underset{(x,x^+)\sim p_{\text{pos}}}{\mathbb{E}} \|f(x) - f(x^+)\|_2^\alpha \tag{10}$$

We compute uniformity/alignment on our test set and define p_{pos} as pairs with the same intent label, and p_{data} as

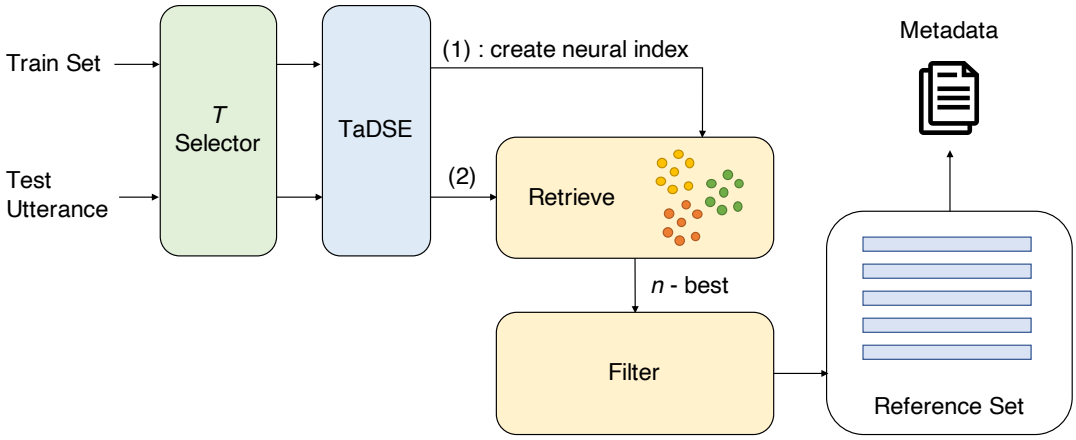

Figure 4: Metadata extraction pipeline with retrieval from training data. We select relevant reference sets via comparison with training set representations similar to the REINA method. T selector is an inference-only component to select the relevant template for the utterance, with the default configuration being the template paired with the utterance. TaDSE produces a semantic representation. The neural index is first created from the training set, while retrieve and filter steps describe how test input is utilized as a query to create a relevant reference set from the neural index.

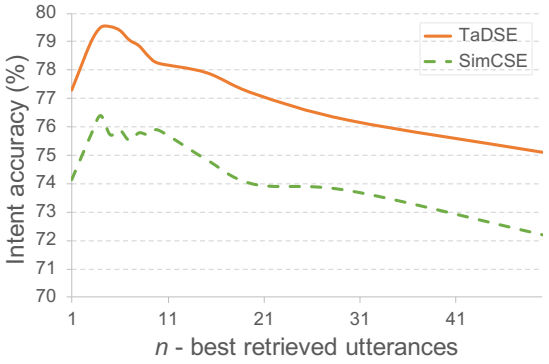

Figure 5: MASSIVE intent accuracy performance with TaDSE ($L_i^t + L_i^u + L_i^{pair_u}$) and SimCSE. The horizontal axis is the n value in n-best utterance retrieval.

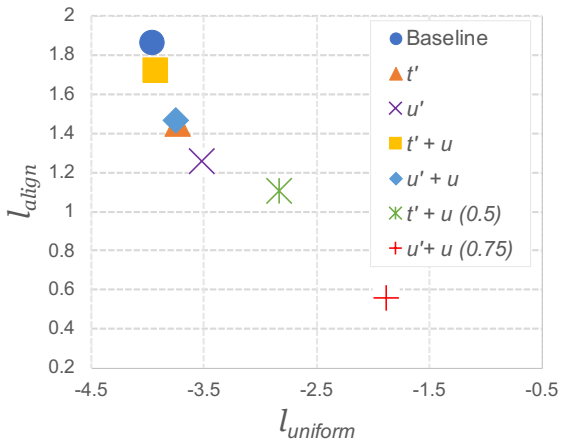

Figure 6: Uniformity & alignment of our best model variants. The smaller value is superior for both metrics. Baseline is L_i^u trained model (SimCSE$_{\text{utt-only}}$), while t', u', u are $L_i^t + L_i^{pair_t}$, $L_i^t + L_i^{pair_u}$, L_i^u losses denoting models trained via pairwise modelling methods. Floating point labels are λ^{infer} values for inference scaling variants of the model.

the full test set. We show uniformity/alignment for our best models in Figure 6.

We observe that TaDSE-infer method (rightmost in the graph) has improved alignment by almost 70% compared to the SimCSE baseline, while uniformity is inferior. We observe that models trained via the TaDSE method (t', u', $t'+u$, $u'+u$) show relatively superior alignment and slightly inferior uniformity compared to the baseline. The inclusion of L_i^u loss weakens alignment, which may mean that the models learn representations that are independent of spoken dialogue.

Interestingly, we detect that models further inferred via the TaDSE method (inference scaling with $t' + u$, $u' + u$) exhibit comparatively superior alignment that positively correlates with λ^{infer}, with inverse correlation for uniformity strength. This consistent trend could be interpreted as for TaDSE methods introducing "skew" to the hyperspace, in which successive application of pairwise anchoring methods for both training (Section 3.2) and more markedly inference (Section 3.3) redistribute the representations to a more compressed, well-aligned hyperspace. This also explains inferior uniformity. We discover that TaDSE methods obtain improved performance for spoken dialogue that roughly corresponds to superior alignment.

7.2 Qualitative Analysis

We analyze distributions of our representations via the 2D T-SNE method. We employ the default T-SNE configuration from the Scikit-learn library with a perplexity of 30.0.

T-SNE diagrams with color-coded intents are shown in

Figure 7, 8 with more intent cluster diagrams in Appendix H. While we can observe global intent clusters in all diagrams, TaDSE models reveal a clear local structure with a multitude of clusters per label. This observation aligns with an in-depth discussion about "hyperspace skew" in Section 7.1.

T-SNE diagrams with auxiliary template representations are shown in Figure 10, 11 and 12 of Appendix G. Figure 10 exhibits sparse template representation groupings learned via L_i^t loss, while utterance representations are roughly gathered together. TaDSE models (Figure 11, 12) display an abundance of local utterance representation clusters along with dense template representation clusters. Global utterance representation structure is observed to have also expanded. However, we observe pronounced representation clusters with separate utterance and template representations. This suggests a possible interpretation as multimodality (Liang et al. 2022).

8 Conclusions

In this work, we propose TaDSE, a novel unsupervised representation learning method that produces semantic representations for dialogue. We present methods of encoding pair relationships between templates and matching utterances via a new training scheme and adjusting the inferred representations. We achieve a strong performance increase on the NLU task and MASSIVE downstream task in contrast to utterance-only representation models and methodology. We further explain the inner workings of our methods via uniformity/alignment analysis and representation visualization, in which we report that our methods induce informed compression of semantic hyperspace as per the intended effect. Correspondingly, we report distinct local structures in resulting hyperspace that is consistent with our expectation of semantic representations for spoken dialogue. We believe that problem of producing correct representations for non-universal domains is of immediate importance in enriching our understanding of language usage patterns in a particular domain, of which spoken dialogue is an interesting example. Some examples of this could be sentences in legal or medical documents. We leave it to future work.

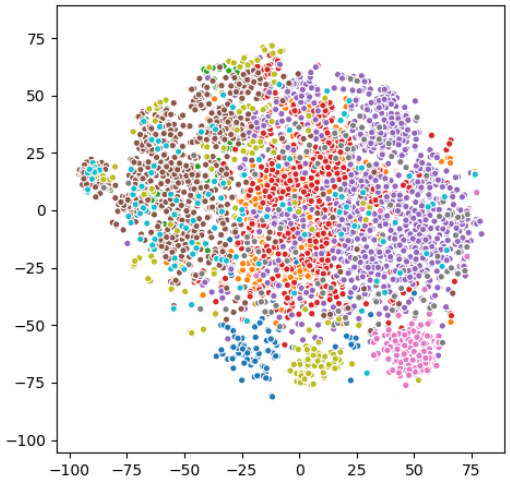

Figure 7: T-SNE diagram with intent label distribution of $\text{SimCSE}_{\text{utt-only}}$ baseline model.

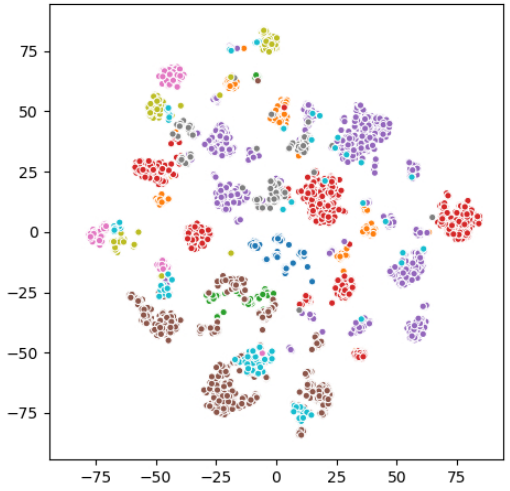

Figure 8: T-SNE diagram with intent label distribution of TaDSE-train model.

References

Bastianelli, E.; Vanzo, A.; Swietojanski, P.; and Rieser, V. 2020. SLURP: A Spoken Language Understanding Resource Package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7252–7262. Online: Association for Computational Linguistics.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. 2020c. Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv preprint arXiv:2006.10029*.

Chuang, Y.-S.; Dangovski, R.; Luo, H.; Zhang, Y.; Chang, S.; Soljacic, M.; Li, S.-W.; Yih, S.; Kim, Y.; and Glass, J. 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*, 4207–4218. Seattle, United States: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

FitzGerald, J.; Hench, C.; Peris, C.; Mackie, S.; Rottmann, K.; Sanchez, A.; Nash, A.; Urbach, L.; Kakarala, V.; Singh, R.; Ranganath, S.; Crist, L.; Britan, M.; Leeuwis, W.; Tur, G.; and Natarajan, P. 2022. MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. arXiv:2204.08582.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Giorgi, J.; Nitski, O.; Wang, B.; and Bader, G. 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)*, 879–895. Online: Association for Computational Linguistics.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hou, Y.; Che, W.; Lai, Y.; Zhou, Z.; Liu, Y.; Liu, H.; and Liu, T. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*.

Jiang, T.; Jiao, J.; Huang, S.; Zhang, Z.; Wang, D.; Zhuang, F.; Wei, F.; Huang, H.; Deng, D.; and Zhang, Q. 2022. PromptBERT: Improving BERT Sentence Embeddings with Prompts.

Kim, Y.-B.; Kim, D.; Kim, J.-K.; and Sarikaya, R. 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)*, 16–24. New Orleans - Louisiana: Association for Computational Linguistics.

Krone, J.; Zhang, Y.; and Diab, M. 2020. Learning to classify intents and slot labels given a handful of examples. *arXiv preprint arXiv:2004.10793*.

Li, H.; Park, S.; Dara, A.; Nam, J.; Lee, S.; Kim, Y.-B.; Matsoukas, S.; and Sarikaya, R. 2021. Neural model robustness for skill routing in large-scale conversational AI systems: A design choice exploration.

Liang, W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. 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*.

Liu, C.; Wang, R.; Liu, J.; Sun, J.; Huang, F.; and Si, L. 2021. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. *arXiv preprint arXiv:2109.12599*.

Nishikawa, S.; Ri, R.; Yamada, I.; Tsuruoka, Y.; and Echizen, I. 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*, 3870–3885. Seattle, United States: Association for Computational Linguistics.

Park, N.; and Kim, S. 2022. Blurs Behave Like Ensembles: Spatial Smoothings to Improve Accuracy, Uncertainty, and Robustness. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 17390–17419. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 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)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Ren, S.; Zhang, J.; Li, L.; Sun, X.; and Zhou, J. 2021. Text AutoAugment: Learning Compositional Augmentation Policy for Text Classification.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wang, S.; Xu, Y.; Fang, Y.; Liu, Y.; Sun, S.; Xu, R.; Zhu, C.; and Zeng, M. 2022. Training Data is More Valuable than You Think: A Simple and Effective Method by Retrieving from Training Data.

Wang, T.; and Isola, P. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks.

Yu, D.; He, L.; Zhang, Y.; Du, X.; Pasupat, P.; and Li, Q. 2021. Few-shot intent classification and slot filling with retrieved examples. *arXiv preprint arXiv:2104.05763*.

Zhou, Z.; Zhang, D.; Xiao, W.; Dingwall, N.; Ma, X.; Arnold, A. O.; and Xiang, B. 2022. Learning Dialogue Representations from Consecutive Utterances. *NAACL*.

A Tokenization Strategy Ablations

We experiment with 3 tokenization strategies to evaluate whether template structure or slot values are more important. We start with 1Slot configuration, where all slots are uniformly replaced with {SLOT} token, and keep top 300 / 1000 occurring slots in the dataset as unique tokens in 300Slot / 1000Slot variants respectively. Sample in Table 6.

$$\begin{aligned} &\text{1Slot} : \forall s_j \rightarrow \{\text{SLOT}\} \\ &k\text{Slot} : s_j \rightarrow \begin{cases} s_j, & \text{if } s_j \text{ in top } k \\ \{\text{SLOT}\}, & \text{otherwise} \end{cases} \end{aligned} \tag{11}$$

where s_j notates each slot value and {SLOT} notates replacement token.

Raw Input Example	
Utterance	Turn on the lamp in study.
Template	Turn on the {LIGHT} in {ROOM}.
Slot Rank	{LIGHT}:750th, {ROOM}:52th

Tokenization Output	
1Slot	Turn on the {SLOT} in {SLOT}.
300Slot	Turn on the {SLOT} in {ROOM}.
1000Slot	Turn on the {LIGHT} in {ROOM}.

Table 6: Tokenization schemes.

We find that 1Slot, 300Slot, and 1000Slot variants each achieve similar performance on a separate large validation set (2.6M rows) when trained with L_i^t loss. We posit that L_i^t performance comes from masking and remaining keywords, thus the number of slot tokens could be irrelevant.

Model	Hit Rate		MRR		MAP		Average
	Skill	Intent	Skill	Intent	Skill	Intent	
L_i^t (1Slot)	80.35	75.85	88.10	85.57	86.72	83.86	83.41
L_i^t (300Slot)	80.26	74.48	88.43	84.90	86.97	83.13	83.03
L_i^t (1000Slot)	80.85	75.12	88.56	85.02	87.28	83.38	83.34

Table 7: Evaluation of tokenization variants.

B Template Paired Dataset

We define training set from random 80% of unique rows (5.3M). We further randomly sample 0.5% of the remaining data each to create valid and test sets (6K each). Templates and slots are defined with human annotations, which are subsequently applied to utterances. We show the dataset sizes in Table 8. Dataset samples are shown in Table 9.

Data Split	Train	Valid	Test
Data Size	5.3M	6K	6K

Table 8: Dataset size.

Utterance	Template
turn my bedside lamps on	`turn my {device} {lights} on`
turn porch side lights off	`turn {device} {lights} off`
are the garage doors open	`are the {doors} {state}`
how warm is the house	`how {temp} is {location}`
what is virgo horoscope	`what is {zodiac} horoscope`
play haunted house sound	`play {sound} sound`
play playboy radio	`play {channel} radio`

Table 9: Utterance and template pairs. Cosmetically similar utterances with dissimilar intents are as follows: "turn my bedside lamps on" & "turn porch side lights off", "play haunted house sound" & "play playboy radio".

C Training Details

We work with BERT-base (Devlin et al. 2019) model. We train the models using a modification of SimCSE (Gao, Yao, and Chen 2021) code-base. The learning rate is $3e-5$ and we use 'cls' pooler in our experiments. We select the resulting model after 1 epoch. We mostly experiment with the 1Slot variant as similar performance improvements are observed (Appendix A). We iterate over loss ratio $\lambda^u, \lambda^{pair} \in \{0, 0, 0.1, 0.5, 1.0\}$. We experiment with different temperatures $\tau_u, \tau_p, \tau_{pair} \in \{0.05, 0.5, 5, 100\}$, and we find that 0.05 works best. For inference scaling ratio $\lambda^{infer} \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, we ablate on the validation set and choose the best value for each configuration. The best values are found to be 0.5 or 0.75 (experiments in Appendix D).

The following hyperparameters are selected via the above process:

1. $L_i^t + L_i^u + L_i^{pair_t}$, $L_i^t + L_i^{pair_t}$: $\lambda^{pair_t} = 0.5$ & $\lambda^u = 0.5, 0.0$
2. $L_i^t + L_i^u + L_i^{pair_u}$, $L_i^t + L_i^{pair_u}$: $\lambda^{pair_u} = 1.0$ & $\lambda^u = 0.1, 0.0$

D Inference Scaling Experiments

We display inference scaling ablation results on the validation set. We tune hyperparameter λ^{infer} according to the validation set performance and select the best configuration for test set evaluations in Section 5.2.

Model	$\lambda^{infer} = 0.0$	$\lambda^{infer} = 0.25$	$\lambda^{infer} = 0.5$	$\lambda^{infer} = 0.75$	$\lambda^{infer} = 1.0$
$L_i^t + L_i^{pair_t}$	76.29	80.38	82.43	**82.51**	43.31
$L_i^t + L_i^u + L_i^{pair_t}$	77.04	79.23	**82.14**	81.96	42.98
$L_i^t + L_i^{pair_u}$	76.16	80.43	**82.03**	81.87	42.73
$L_i^t + L_i^u + L_i^{pair_u}$	76.26	79.77	81.94	**82.04**	42.83

Table 10: TaDSE inference scaling method with different λ^{infer} values, evaluated on validation set. Selected models are highlighted in bold.

E NLU Metrics per Evaluation Set

We develop new evaluation metrics as to better reflect spoken dialogue semantics. For this, we utilize intent labels already present in NLU datasets (FitzGerald et al. 2022; Bastianelli et al. 2020). To enable fine-grained evaluation of semantic relevance, we borrow ranking methodology. After computing representations for all utterances in selected evaluation set, we iterate over the utterances to each select their top-k ($k = 5$) most similar utterance candidates from the evaluation set from which we evaluate based on selected label.

We average 3 performance metrics :

1. **Hit Rate (**$HR@K$**)** : Number of positive pairs within candidates.
2. **Mean Reciprocal Rank (**$MRR@K$**)** : We select first positive candidate and calculate its reciprocal rank.
3. **Mean Average Precision(**$MAP@K$**)** : We find all positive candidates and calculate average precision.

We compute intent ranking metric on validation and test set via designating average of all ranking metrics per data set as its representative metric.

We display raw results for each of the evaluation sets, Valid and Test. We find that individual metrics correlate well with each other in terms of relative performance.

E.1 Valid

Model	Hit Rate (Intent)	MRR (Intent)	MAP (Intent)	Average
SimCSE, Wiki	50.24	64.66	62.05	58.93
SimCSE, TPD	55.13	63.43	62.17	60.24
TaDSE-train	70.06	80.14	78.59	76.26
TaDSE-infer	**76.47**	**85.42**	**84.24**	**82.04**

Table 11: Evaluation of major models on Valid set.

E.2 Test

Model	Hit Rate (Intent)	MRR (Intent)	MAP (Intent)	Average
SimCSE, Wiki	52.46	66.34	63.76	60.85
SimCSE, TPD	56.88	65.19	64.23	62.10
TaDSE-train	72.00	81.04	89.75	80.93
TaDSE-infer	**76.71**	**85.50**	**84.57**	**82.26**

Table 12: Evaluation of major models on Test set. We find that Valid and Test performance mostly align.

F Template Data Distribution

We display number of utterances per representative templates in graph below. In terms of unique templates in training data, there are 10K templates with slot values and 6K utterances without such slots.

F.1 Top 20 Templates

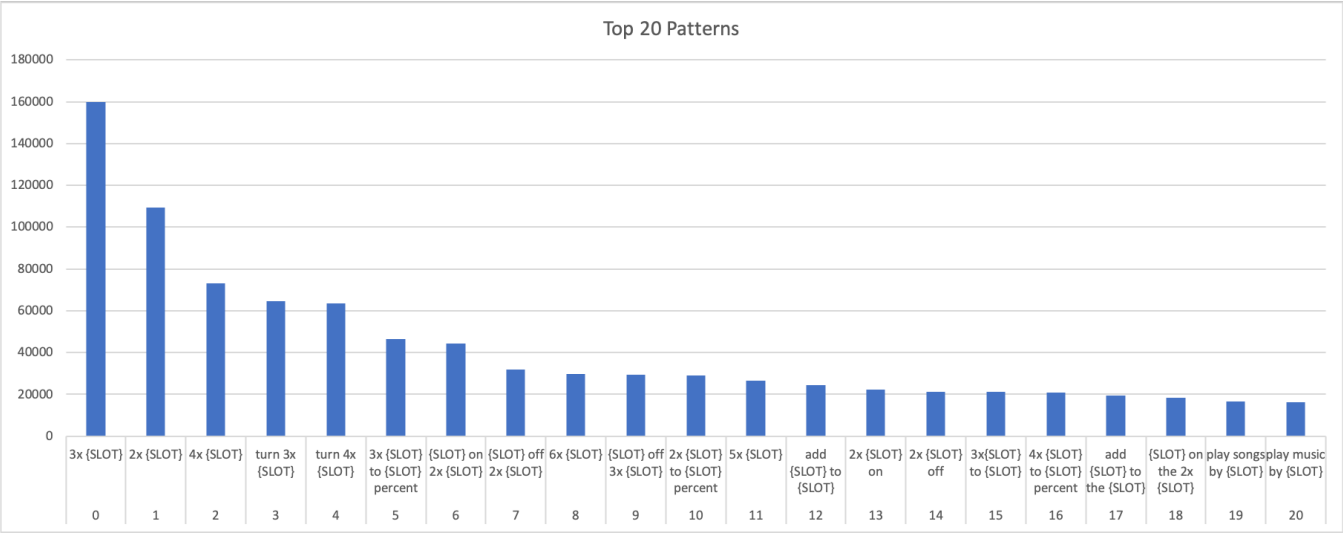

Figure 9: We display top 20 templates in terms of paired utterance count. Refer to Table 11 for tokenization strategy.

F.2 Few-shot Utterances

We enumerate randomly selected utterances that does not have matching templates with slot values in Table 13. They function as few-shot utterance data for training L_i^t model.

Utterances
force a comms channel
tell me the status
go to sleep girl
what is latent defect
fight another knight
zapdos
bring me my coffee
end my party
cancel the mission
how did the stock market go today
any good restaurants near me
we're ready to resume

Table 13: Utterances that do not have templates with slot values.

G Visualization of Template Representations

T-SNE diagrams for template representation distribution is available in Figure 10, 11 and 12. Discussion in Section 7.2.

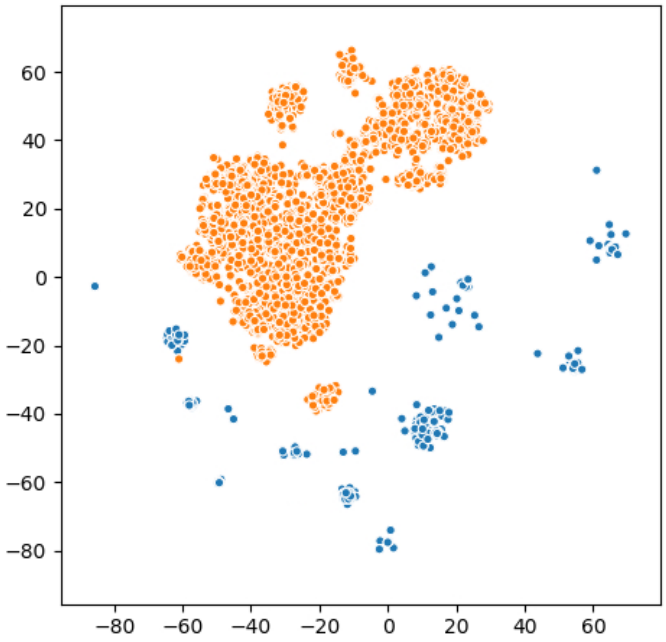

Figure 10: T-SNE diagram with utterance (orange) / template (blue) representations from L_i^t only model.

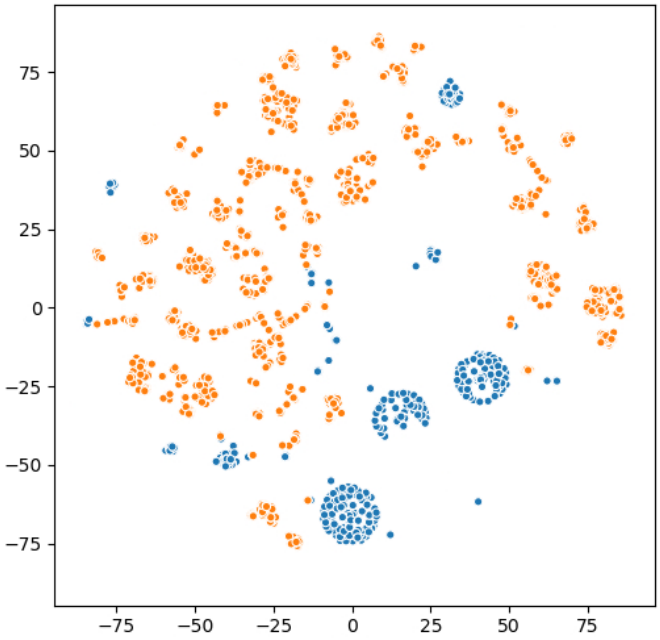

Figure 11: T-SNE diagram with utterance (orange) / template (blue) representations from $L_i^t + L_i^u + L_i^{pair_u}$ model.

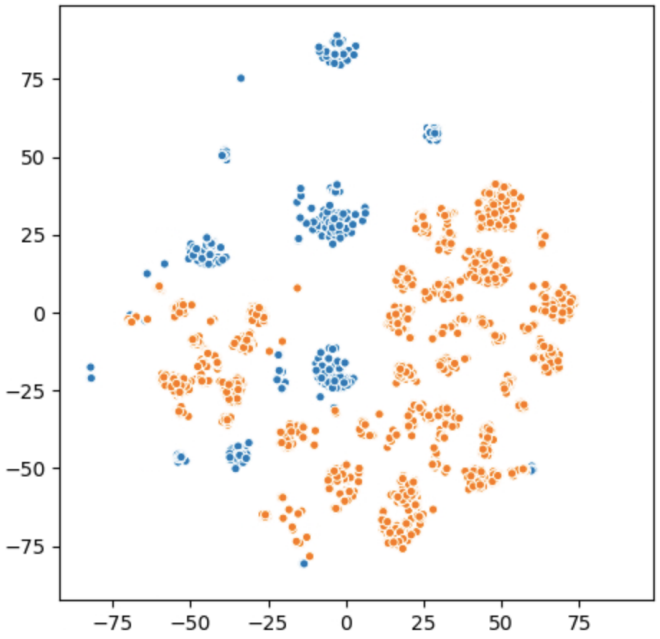

Figure 12: T-SNE diagram with utterance (orange) / template (blue) representations from $L_i^t + L_i^u + L_i^{pair_t}$ model.

H Visualization of Intent Labels

Additional T-SNE diagrams for intent label distribution are available in Figure 13, 14 and 15. Discussion in Section 7.2.

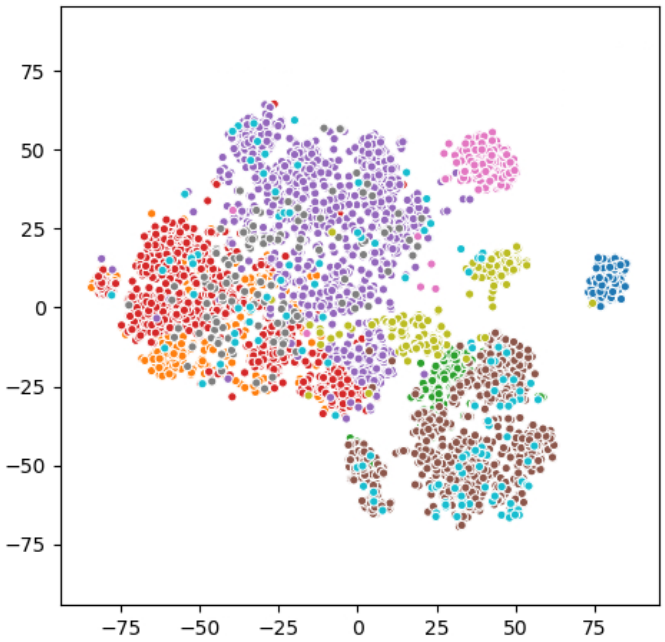

Figure 13: T-SNE diagram with intent label distribution of L_i^t only model.

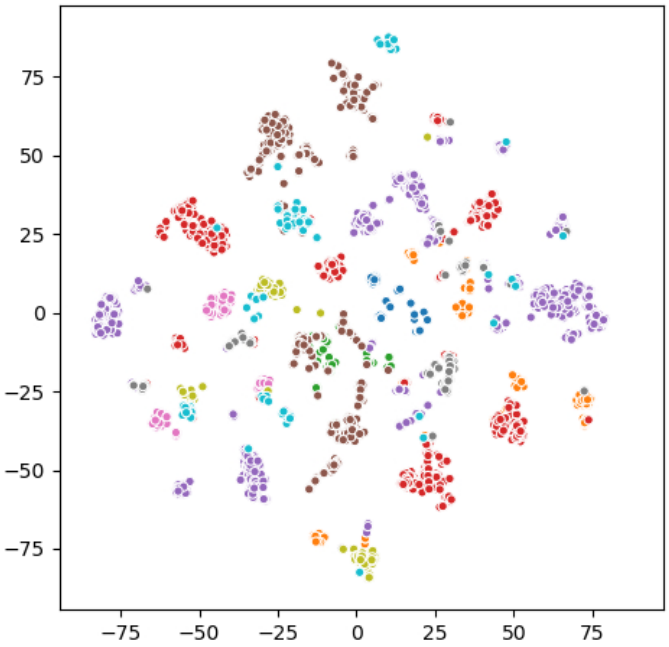

Figure 14: T-SNE diagram with intent label distribution of $L_i^t + L_i^{pair_u}$ model.

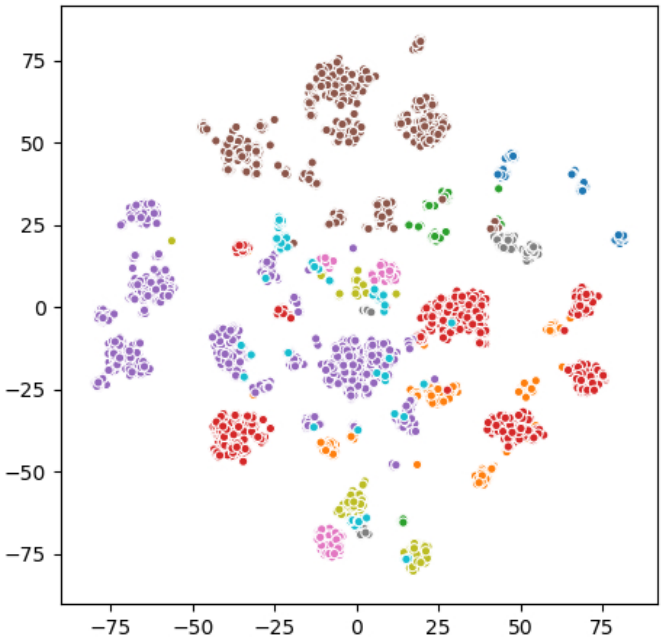

Figure 15: T-SNE diagram with intent label distribution of TaDSE-infer ($L_i^t + L_i^u + L_i^{pair_u}$) model.

I MASSIVE Experiment Setup

We follow the TaDSE methodology for training and perform experiments with a learning rate of $3e-5$. We select the resulting model after 1 epoch. Our MASSIVE models have $\lambda^u = 1, \lambda^{pair} = 0.01$. We further experiment with tuning the reference set by changing n-best criteria, in which we find $n = 5$ best. We obtain a reference set by filtering the retrieved training data based on the overall frequency of intent label and similarity score. Training data from the dominant cluster near the test utterance will be selected as a reference set.

J DSE MASSIVE Samples

Query	Top-1 Result
could you please help me in listening to the radio	the radio should play only on nine hundred and ninety nine f. m.
i would be happy if you update me the events going on our area	go silent until three p. m.
please describe that object for me	open internet
book me a cab going to location	book a taxi uber
send mom an email now	start a new email to
i got promoted today it feels so good	how busy am i this week

Table 14: Samples from DSE-BERT$_{\text{large}}$. This model achieves lower intent accuracy of 9.99%.