

Probing Task-Oriented Dialogue Representation from Language Models

Chien-Sheng Wu and Caiming Xiong

Salesforce Research

[wu.jason, cxiong]@salesforce.com

Abstract

This paper investigates pre-trained language models to find out which model intrinsically carries the most informative representation for task-oriented dialogue tasks. We approach the problem from two aspects: supervised classifier probe and unsupervised mutual information probe. We fine-tune a feed-forward layer as the classifier probe on top of a fixed pre-trained language model with annotated labels in a supervised way. Meanwhile, we propose an unsupervised mutual information probe to evaluate the mutual dependence between a real clustering and a representation clustering. The goals of this empirical paper are to 1) investigate probing techniques, especially from the unsupervised mutual information aspect, 2) provide guidelines of pre-trained language model selection for the dialogue research community, 3) find insights of pre-training factors for dialogue application that may be the key to success.

1 Introduction

Task-oriented dialogue systems achieve specific user goals within a limited number of dialogue turns via natural language. They have been used in a wide range of applications, such as booking restaurants (Wen et al., 2017), providing tourist information (Budzianowski et al., 2018), ordering tickets (Schulz et al., 2017), and healthcare consultation (Wei et al., 2018). They are also crucial components of intelligent virtual assistants like Siri, Alexa, and Google Assistant.

Most of the task-oriented dialogue systems nowadays, are benefited from transfer learning (Wu et al., 2019; Lin et al., 2020), especially pre-trained language models trained on general text, such as BERT (Devlin et al., 2018) and GPT2 (Radford et al., 2019). However, previous work claims that linguistic patterns could differ between writing text

Model	Dial. Data	Parameters	Output Dim.
BERT-base	X	109.5M	768
AIBERT-base	X	11.7M	768
DistilBERT-base	X	66.4M	768
RoBERTa-based	X	124.6M	768
GPT2-small	X	124.4M	768
ELECTRA-GEN	X	33.5M	256
ELECTRA-DIS	X	108.9M	768
ConveRT	V	29M	1024
DialoGPT-small	V	124.4M	768
TOD-BERT-mlm	V	119.5M	768
TOD-BERT-jnt	V	119.5M	768
TOD-GPT2	V	124.4M	768

Table 1: An overview of selected pre-trained language models (Details in Section 2).

and human conversation, resulting in a large gap of data distributions (Bao et al., 2019; Wolf et al., 2019b). Recently, several approaches are leveraging open-domain data (Henderson et al., 2019; Zhang et al., 2019), or aggregating task-oriented data (Wu et al., 2020) to pre-train language models.

In this paper, we are interested in answering these questions: which language model has the most informative representations that is better for what task-oriented dialogue task? Does pre-training with dialogue-specific data or different objectives make any difference? We investigate how good these pre-trained representations are for a task-oriented dialogue system, ignoring the model architectures and training strategies by only probing their final representations with fine-tuning models. A good representation implies better knowledge transferring and domain generalization ability, making downstream applications easier and cheaper to be improved.

We tackle this problem with two probing solutions: supervised classifier probe and unsupervised mutual information probe. Classifier probe is commonly used in different NLP tasks such as morphology (Belinkov et al., 2017), sentence length (Adi et al., 2016), or linguistic structure (Hewitt and

Manning, 2019). In this setting, we fine-tune a simple classifier for a specific task (e.g., intent identification) on a fixed pre-trained language model. The probe uses supervision to find the best transformation for each sub-task.

In addition, we present mutual information probe to investigate these language models by directly clustering their output representations, as recent study (Pimentel et al., 2020) suggests that a simple classifier may not be able to achieve the best estimate of mutual information between features and the downstream task. We apply two clustering techniques, K-means (Lloyd, 1982) and Gaussian mixture model (Reynolds, 2009), to calculate its adjusted normalized mutual information (ANMI) (Vinh et al., 2010) between the predicted clustering and the true task-specific clustering.

We investigate 12 language models, as shown in Table 1, where five of them have been pre-trained with dialogue data. We evaluate four core task-oriented dialogue tasks, domain identification, intent detection, slot tagging, and dialogue act prediction. They correspond to the commonly defined natural language understanding, dialogue state tracking, and dialogue management modules (Wen et al., 2017). We hope our probing analysis can provide insights to facilitate future task-oriented dialogue research. Some of the key observations in this work are summarized here (More discussion in Section 4.4):

- No matter the open-domain or close-domain, pre-training with dialogue data helps learning better representations for task-oriented dialogue.
- Pre-trained language models intrinsically contain more information about intents and dialogue acts but less for slots.
- ConveRT (Henderson et al., 2019) and TOD-BERT-jnt (Wu et al., 2020) have the highest classification accuracy and mutual information score, suggesting that response selection is useful for dialogue pre-training, especially when we compare TOD-BERT-jnt to TOD-BERT-mlm.
- Top models also include TOD-GPT2 and DistilBERT (Sanh et al., 2019). The distilled version of BERT surprisingly outperforms BERT and other strong baselines such as RoBERTa (Liu et al., 2019).
- DialoGPT and GPT2 do not perform well on mutual information evaluation but have a middle-

ranking classification accuracy, implying that their representations are informative but not suitable for unsupervised clustering.

- Models such as AIBERT (Lan et al., 2019) and ELECTRA (Clark et al., 2020) have low classification accuracy and mutual information, showing the least useful information on task-oriented dialogue tasks.

2 Pre-Trained Language Models

We can roughly divide pre-trained language models into two categories: uni-directional and bi-directional. BERT-based systems are bi-directional language models and usually trained with the masked language modeling (MLM) objective, i.e., given the left and right context to predict the current masked token. GPT-based models, on the other hand, are uni-directional language models trained always to predict the next token in an auto-regressive way.

For a BERT-based model, we use the final-layer hidden state of its first token, [CLS], to represent an input sequence. This built-in token is originally designed to aggregate the information. Since GPT-based models are uni-directional and do not have a similar design as the [CLS] token, we use the mean pooling of its output hidden states to represent the input sequence, which is better than only using the last hidden state in our experiments.

BERT-based BERT is a Transformer (Vaswani et al., 2017) encoder with a self-attention mechanism, which is trained on Wikipedia and BookCorpus using the MLM and next sentence prediction objectives. Liu et al. (2019) proposed a robustly optimized approach for BERT, call RoBERTa, where they improved it by training the model longer with bigger batches over more data and longer sequences, and removing the next sentence prediction objective. Lan et al. (2019) proposed a lite BERT (AIBERT) that trained with MLM and inter-sentence coherence losses, and aimed to lower memory consumption and increase the training speed. With similar motivation, Sanh et al. (2019) trained a DistilBERT that reduce 40% of parameters with a triple loss, including MLM, distillation, and cosine-distance losses. Clark et al. (2020) proposed ELECTRA using a sample-efficient pre-training task called replaced token detection. They used a generator network (ELECTRA-GEN) to replace tokens with plausible alternative tokens and

trained a discriminative model (ELECTRA-DIS) to predict whether the generator replaced each token in the input.

Most of the pre-trained models above are trained on general text corpora with language modeling objectives. Henderson et al. (2019), on the other hand, used social media conversational data to train the ConveRT model. It is a Transformer-based dual-encoder model pre-trained on a dialogue response selection task using 727M Reddit (input, response) pairs. Very recently, Wu et al. (2020) proposed task-oriented dialogue BERT (TOD-BERT), which is initialized by BERT and further pre-trained on nine publicly available task-oriented dialogue corpora. They have one version with only MLM objective (TOD-BERT-mlm) and another with both MLM and contrastive learning objectives of response selection (TOD-BERT-jnt). TOD-BERT has shown good performance on several task-oriented downstream tasks, especially in the few-shot setting.

GPT-based GPT2 (Radford et al., 2019) is the representative of uni-directional language models using a Transformer decoder, where the objective is to maximize left-to-right generation likelihood. To ensure diverse and nearly unlimited text sources, they use Common Crawl to obtain 8M documents as its training data. Budzianowski and Vulić (2019) trained GPT2 on task-oriented response generation task, taking system belief, database result, and last dialogue turn as inputs. It only uses one dataset to train its model because few public datasets have database information available for pre-training. Zhang et al. (2019) pre-trained GPT2 on 147M open-domain Reddit data for response generation and called it DialoGPT. It aims to generate more relevant, contentful, and consistent responses for chit-chat dialogue systems. In this paper, following TOD-BERT’s idea, we train a task-oriented GPT2 model (TOD-GPT2) built on the GPT2 model and further pre-trained with task-oriented datasets. We use the same dataset collection, which contains nine datasets in total, as shown in Wu et al. (2020), to pre-train the model as a reference.

3 Method

We define a dialogue corpus $D = \{D_1, \dots, D_M\}$ has M dialogue samples, and each dialogue sample D_m has T turns of conversational exchange $\{U_1, S_1, \dots, U_T, S_T\}$ between a user and a system. For every utterance U_t or S_t , we have human-

annotated domain, user intent, slot, and dialogue act labels. We first feed all the utterances to a pre-trained model and obtain user and system representations. In this section, we first discuss how we design our classifier probe and then introduce our mutual information probe’s background and usage.

3.1 Classifier Probe

We use a simple classifier to transform those representations for a specific task and optimize it with annotated data.

$$V_i = A(FFN(E_i)), \quad (1)$$

where $E_i \in \mathbb{R}^{d_B}$ is the output representation with dimension d_B from a pre-trained model, $FFN \in \mathbb{R}^{N \times d_B}$ is a feed-forward layer that maps from dimension d_B to a prediction with N classes, and A is an activation layer. For domain identification and intent detection, we use a Softmax layer and backpropagate with the cross-entropy loss. For dialogue slot and act prediction, we use a Sigmoid layer and the binary cross-entropy loss since they are multi-label classification tasks.

3.2 Mutual Information Probe

We first cluster utterances in an unsupervised fashion using either K-means (Lloyd, 1982) or Gaussian mixture model (GMM) (Reynolds, 2009) with K clusters. Then we compute the adjusted mutual information score (Vinh et al., 2010) between the predicted clustering and each of the true clusterings (e.g., domain and intent) for different hyperparameters K . Note that the predicted clustering is not dependent on any particular labels.

3.2.1 Utterance Clustering

K-means is a common clustering algorithm that aims to partition N samples into K clusters $A = \{A_1, \dots, A_K\}$ in which each sample is assigned to a cluster centroid with the nearest mean.

$$\arg \max_A \sum_{i=1}^K \sum_{x \in A_i} \|x - \mu_i\|^2, \quad (2)$$

where μ_i is the centroid of the A_i cluster and the algorithm is updated in an iterative manner.

On the other hand, GMM assumes a certain number of Gaussian distributions (K mixture components). It takes both mean and variance of the data into account, while K-means only consider the

data's mean. By the Expectation-Maximization algorithm, GMM first calculates each sample's probability belongs to a cluster A_i during the E-step, then updates its density function to compute new mean and variance during the M-step.

In our experiments, we cluster separately for user utterances U and system response S . Note that K is a hyper-parameter since we may not know the true distribution in a real scenario. To avoid the local minimum issue, we run multiple times (typically ten runs) and use the best clustering result for mutual information evaluation.

3.2.2 ANMI

To evaluate two clusterings' quality, we compute the ANMI score between a clustering and its ground-truth annotation. ANMI is adjusted for randomness, which accounts for the bias in mutual information, giving high values to the clustering with a larger number of clusters. ANMI has a value of 1 when two partitions are identical, and an expected value of 0 for random (independent) partitions.

More specifically, we assume two label clusterings, A and B , that have the same N objects. The mutual information (MI) between A and B is defined by

$$MI(A, B) = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} P(i, j) \log\left(\frac{P(i, j)}{P(i)P(j)}\right), \quad (3)$$

where $P(i, j) = |A_i \cap B_j|/N$ is the probability that a randomly picked sample falls into both A_i and B_j classes. Similarly, $P(i) = |A_i|/N$ and $P(j) = |B_j|/N$ are the probabilities that the sample falls into either the A_i or B_j class.

The normalized mutual information (NMI) normalizes MI with the mean of entropy, which is defined as

$$NMI(A, B) = \frac{MI(A, B)}{\text{mean}(H(A), H(B))}, \quad (4)$$

where $H(A) = -\sum_{i=1}^{|A|} P(i) \log(P(i))$ is the entropy of the A clustering, which measures the amount of uncertainty for the partition set.

MI and NMI are not adjusted for chance and will tend to increase as the number of cluster increases, regardless of the actual amount of "mutual informaton" between the label assignments. Therefore, adjusted normalized mutual information (ANMI) is designed to modify NMI score with its expectation, which is defined by

$$ANMI = \frac{MI - \mathbb{E}[MI]}{\text{mean}(H(A), H(B)) - \mathbb{E}[MI]}, \quad (5)$$

MWOZ		
Domain	Dialogue Act	Slot
restaurant	nobook	type
hotel	bye	book day
attraction	request	book people
train	recommend	day
taxi	welcome	pricerange
	book	leaveat
	greet	arriveby
	nooffer	parking
	reqmore	book time
	offerbooked	name
	select	destination
	inform	internet
	offerbook	stars
		book stay
		departure
		area
		food
		department

Table 2: Labels classes in the MWOZ Data.

where the expectation $\mathbb{E}[MI]$ can be calculated using the equation in [Vinh et al. \(2010\)](#).

4 Experiments

4.1 Datasets

The multi-domain Wizard-of-Oz (MWOZ) dataset ([Budzianowski et al., 2018](#)) is one of the most common benchmark datasets for task-oriented dialogue systems. We use MWOZ to evaluate domain identification, dialogue slot tagging, and dialogue act prediction tasks. It contains 8420/1000/1000 dialogues for training, validation, and testing sets, respectively. There are seven domains in the training set and five domains in the others. There are 13 unique system dialogue acts and 18 unique slots as shown in Table 2.

Besides, we use the out-of-scope intent (OOS) dataset ([Larson et al., 2019](#)) for our intent detection experiment. The OOS dataset is one of the largest annotated intent datasets, including 15,100/3,100/5,500 samples for the train, validation, and test sets, respectively. It has 150 intent classes over ten domains and an additional out-of-scope intent class, a user utterance that does not fall into any of the predefined intents. The whole intent list is shown in the Appendix.

4.2 Training Details

We first process user utterance and system response using the tokenizer corresponding to each per-trained model. To obtain each representation, we run most of the pre-trained models using the

HuggingFace (Wolf et al., 2019a) library, except the ConveRT¹ and TOD-BERT². We fine-tune GPT2 using its default hyper-parameters and the same nine datasets as shown in Wu et al. (2020) to train for TOD-GPT2 model. For classifier probing, we fine-tune the top layer with a consistent hyper-parameter setting. We apply AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate $5e^{-5}$ and gradient clipping 1.0. We use $K = 4, 8, 16, 32, 64, 128, 256$ with 50 iterations each, and report the moving trend for MI probing. We use GMM clustering from the scikit-learn library, and we adopt the K-means implementation from the faiss library (Johnson et al., 2017). Experiments were conducted on a single NVIDIA Tesla V100 GPU.

4.3 Evaluation

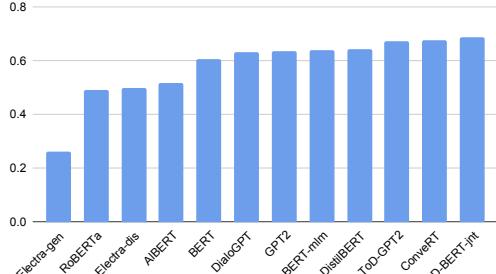
Domain identification and intent detection tasks are multi-class classification problems. Therefore, we can directly use their annotated domain and intent labels to compute the ANMI scores. Slot tagging and dialogue act prediction tasks, meanwhile, are multi-label classification problems. For example, each utterance can include multiple slots mentioned (<food> and <price> slots) and various actions triggered (<greeting> and <inform> acts). In our experiment, we use a naive way that is viewing a different set of slot or act combination as different labels, e.g., three slot sets <food>, <food, price>, and <price, location> belong to three different clusters.

4.4 Results

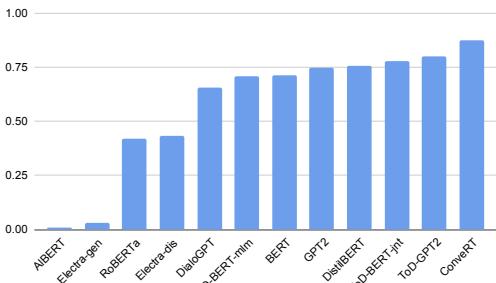
Classifier results are shown in Figure 1. We can observe that ConveRT, TOD-BERT-jnt, and TOD-GPT2 achieve the best performance, implying that pre-training with dialogue-related data captures better representations, at least in these sub-tasks. Moreover, the performance of ConveRT and TOD-BERT-jnt suggests that it is helpful to pre-train with a response selection contrastive objective, especially when comparing TOD-BERT-jnt to TOD-BERT-mlm. Moreover, most of the pre-trained models have a similar and high micro-F1 score in (d) system dialogue act prediction, as most of them are above 75% over 13 classes. Dialogue slot (c) information, meanwhile, is not well captured by

¹<https://github.com/PolyAI-LDN/polyai-models>

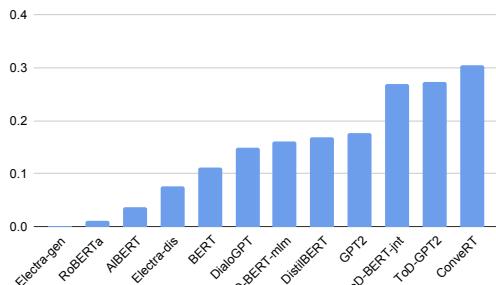
²<https://github.com/jasonwu0731/TOD-BERT>



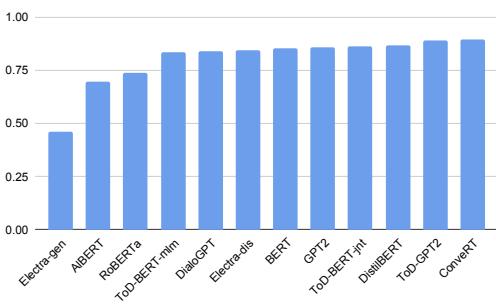
(a) MWOZ Domain (System)



(b) OOS Intent (User)



(c) MWOZ Slot (System)



(d) MWOZ Act (System)

Figure 1: The results of supervised classifier probe. The y-axis in (a) and (b) represents the accuracy. The y-axis in (c) and (d) represents the micro-F1 score.

these representations, resulting in a micro-F1 lower than 30%. On the other hand, ELECTRA-GEN, RoBERTa, and ALBERT show the worst classification results. Especially in (b) intent classification

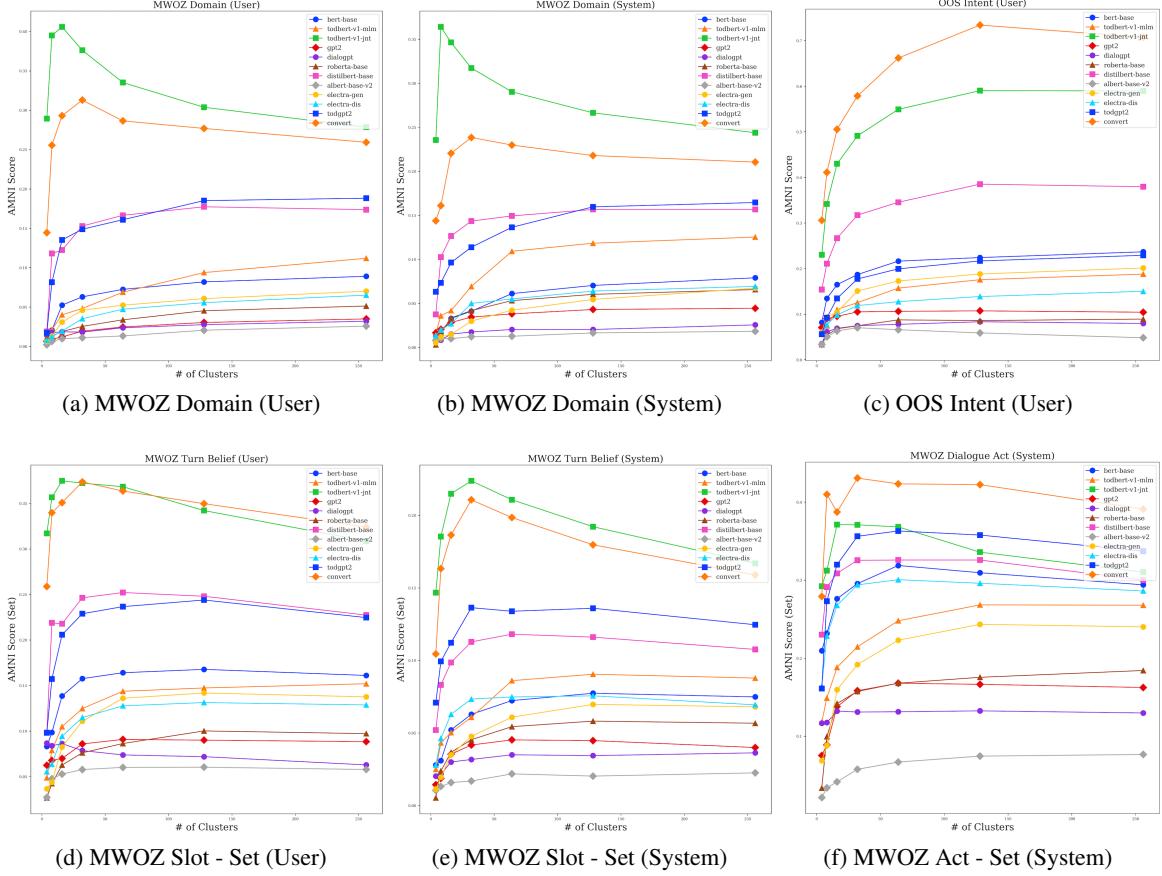


Figure 2: The ANMI evaluation of pre-trained models with the domain, intent, slot, and action labels. The X-axis is the number of clusters and the y-axis is the ANMI score (Best view in color)

and (c) dialogue slot tagging, some of them seem to have zero useful information to make a prediction.

Mutual information results using K-means clustering are shown in Figure 2. Due to the space limit, we report the results using GMM in the Appendix, as the two of them have similar trends. The x-axis is the number of clusters in each subplot, ranging from 4 to 256, and the y-axis is the ANMI score between a predicted clustering and its corresponding true clustering. In general, the mutual information probe results are similar to what we observe in the classifier probe. We can find that Tod-BERT-jnt and ConveRT are those with the highest mutual information, and they are usually followed by TOD-GPT2 and DistilBERT.

Another observation is that representations from those pre-trained language models, especially the top ones, seem to have more connection with user intent and system dialogue act labels than domain and slot labels. The average ANMI scores across 12 models and 7 different number of clusters for intent and dialogue act are 0.193 ± 0.169

and 0.226 ± 0.107 , respectively. But domain and dialogue slot only have 0.086 ± 0.087 and 0.077 ± 0.057 ANMI scores in average. We discuss each subplot in detail in the following:

Figure 2 (a) and (b) show the mutual information between predicted clustering and the true domain labels on the MWOZ dataset. A user utterance seems to have higher domain mutual information than a system response. TOD-BERT-jnt, in this case, outperforms others by a large margin, achieving around 0.4 ANMI with 8 clusters. Figure 2 (c) is about user intent using user utterances. ConveRT surpasses others by far in the mutual information of intent, achieving over 0.7 ANMI at 128 clusters when the true number of classes equals to 151. Other than the top three models (ConveRT, TOD-BERT-jnt, and DistilBERT), the remaining pre-trained models have ANMI scores lower than 0.2.

Figure 2 (d) and (e) show the mutual information evaluation using the slot labels. When comparing (d) to (e), we can find that user utterances contain

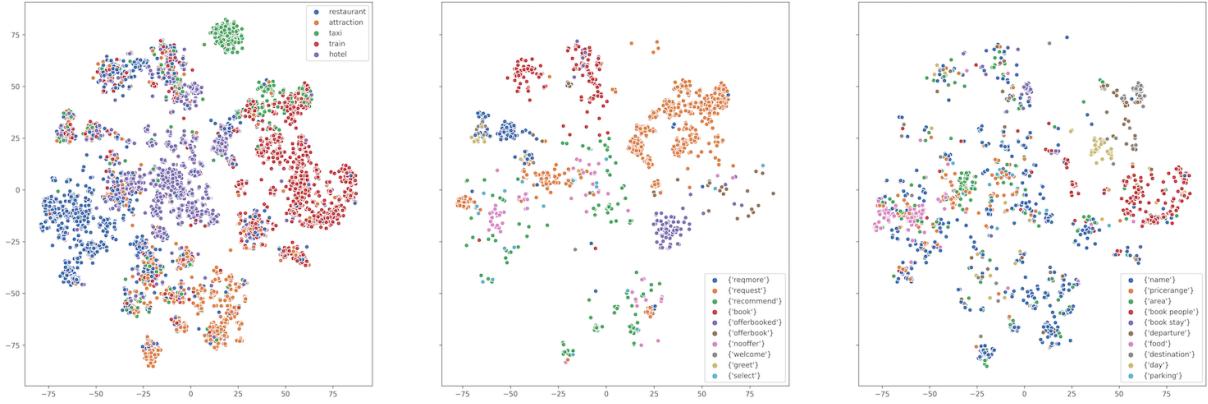


Figure 3: The tSNE visualization of dialogue representations from ToD-BERT-jnt. (Best view in color)

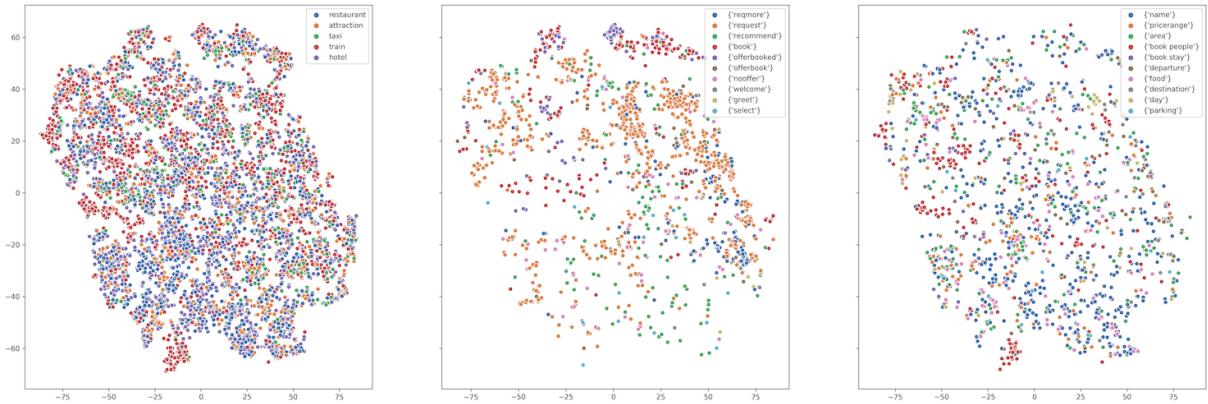


Figure 4: The tSNE visualization of dialogue representations from GPT2. (Best view in color)

more slot information than system responses (Max around 0.35 and 0.25). It is not surprising because a user in task-oriented dialogue is usually the slot information provider, informing what location or which cuisine s/he prefers. ToD-BERT and ConveRT perform similar in this case, still outperform others by a big margin.

Figure 2 (f) shows the mutual information for the predicted clustering of system dialogue acts. We can find that most of the pre-trained language models have shown a relatively high ANMI score (average 0.226) and closed the gap between their performance and the top model. ConveRT works the best, in this case, followed by TOD-BERT-jnt and TOD-GPT2, in which two of them seem to have similar ANMI scores.

5 More Analysis

Difference Between Probes Ideally, both probes should distinguish the goodness of different pre-trained language models, i.e., features that can be easily classified or features with high correlation with true distributions are preferred. However, we

found that although, in general, the trends we observe from two probing methods are similar, they are not the same in terms of the ranking. When comparing the ranking of GPT2 and DialoGPT models in Figure 1 and Figure 2, we found that they obtain almost the worse ANMI scores but work quite good in classification accuracy. This observation means that their representations of different classes are “close” to each other as a low ANMI score suggesting a more noisy clustering. Still, at the same time, it is not hard to find a hyperplane that can well discriminate those features.

We discuss some possible reasons for this interesting observation in the following. The first guess is that these features may not follow a Gaussian distribution, as we assume during clustering, suggesting that more advanced clustering techniques can be investigated in future work. The second guess is that these features have an unavoidable clustering noise that can be denoised or debiased easily by a strong supervision signal. The third guess, which may be a possible reason, is that these features are clustered by some other factors that are

ConveRT	
Cluster 1 (Failed Booking)	i am sorry but dojo noodle bar is solidly booked at that time . i can try a different time or day for you .
	1 moment while i try to make the reservation of table for 8 , friday at 16:30 . booking was unsuccessful . can you try another time slot ?
	i am very sorry i was unable to book at acorn guest house for 5 nights , would you like to try for a shorter stay ?
	i am afraid that booking is unsuccessful . would you like a different day or amount of days ?
Cluster 2 (Train Time)	there are 5 trains available , may i book 1 for you that leaves at 7:40 and arrives at 10:23 ?
	tr0330 departs at 14:09 and arrives by 15:54 . would you like a ticket ?
	the tr2141 arrives by 15:27 . would you like me to reserve some seats for you ?
	i have train tr4283 that leaves cambridge at 5:29 and arrives in bishops stortford at 6:07 . would you like to make reservations ?
Cluster 3 (Restaurant Request)	i have a train that leaves cambridge 14:01 arriving in birmingham new street at 16:44 . would that work ?
	there are 21 restaurant -s available in the centre of town . how about a specific type of cuisine ?
	there are 9 indian restaurant -s in centre what price range do you want ?
	i am sorry , there are no catalan dining establishments in the city centre . would you like to look for a different cuisine or area ?
Cluster 4 (Confirm Booking)	i found 4 restaurant -s with the name tandoori that serve indian food on the south , west , and east . do you have a location preference ?
	there are no singaporean restaurant -s , but there are cheap ones offering several different cuisines .
	all set . your reference number is k2bo09vq .
	i have got you booked for 16:30 . the reference number is eq0yaq1g .
Cluster 5 (Hotel Request)	your reservation was a success and the reference number is jtwxfm7m .
	i have got your booking set , the reference number is 9rmfgjma .
	i booked tr3932 , reference number is fiw5abo2 .
	what part of town there are none in the west .
	i can help you with that . do you have any special area you would like to stay ? or possibly a star request ?
	there are no colleges close to the area you are requesting , would you like to chose another destination ?
	sure , what area are you thinking of staying ?
	i would be happy to help . may i ask what price range and area of town you are looking for ?

Table 3: Clustering results of the ConveRT model. The samples are picked from each randomly selected five clusters with K=32. We can roughly label a topic for each cluster.

not tested, and at the same time, the factors we are interested in are scattered in groups for different classes in a similar way. Intuitively, there are four clustering results shown in Figure 5, where GPT2 and DialoGPT may fall into the (d) clustering type, which has a lower mutual information score but higher classification accuracy.

As a result, we suggest a simple rule of thumb regarding which probing results. In short, the results of the classifier probe could be useful if a supervised approach for a downstream task is designed, e.g., user dialogue act prediction and dialogue state tracking. On the other hand, the mutual information probe is more effective for an unsupervised problem, e.g., utterance clustering and dialogue parsing tasks.

Visualization In Figure 3 and Figure 4, we visualize the embeddings of TOD-BERT-jnt and GPT2 given the same system responses from the MWOZ test set. Each point is reduced from its high-dimension features to a two-dimension point using the t-distributed stochastic neighbor embedding (tSNE). We use different colors to represent different domains (left), dialogue acts (middle), and turn slots (right). As one can observe, TOD-BERT-jnt has more clear group boundaries and better clustering results than GPT2. Visualization plots for other pre-trained models are shown in the Appendix.

What utterances are clustered together? In Table 3, we show the clustering examples of system responses from the top performance model Con-

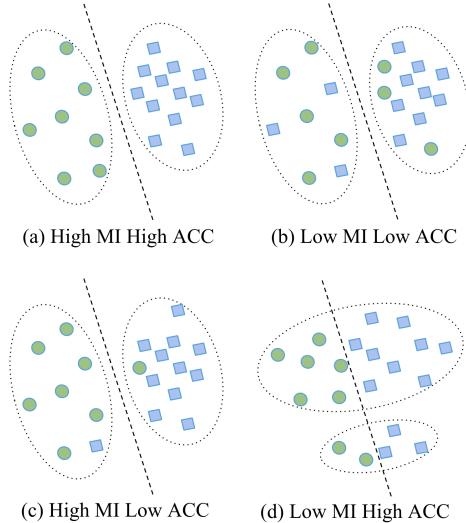


Figure 5: Illustration of four different type of clusterings related to mutual information and accuracy.

veRT. We use $K = 32$ clustering and randomly select five clusters and five samples each. We found that most of the utterances are related to an unsuccessful booking in the cluster 1, containing “I am sorry,” “solidly booked,” or “booking was unsuccessful.” We also found other clusters showing good clustering results, such as selecting departure or arrival time for a train ticket or requesting more user preference for a restaurant reservation. More clustering results are shown in the Appendix.

6 Conclusion

We investigate representations from pre-trained language models for task-oriented dialogue tasks, including domain identification, intent detection, slot tagging, and dialogue act prediction. We use a supervised classifier probe and a proposed unsupervised mutual information probe. From the ranking results of two different probings, we show a list of interesting observations to provide model selection guidelines and shed light on future research towards a more advanced language modeling learning for dialogue applications.

References

- Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.
- Siqi Bao, Huang He, Fan Wang, and Hua Wu. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.
- Paweł Budzianowski and Ivan Vulić. 2019. Hello, it’s gpt-2—how can i help you? towards the use of pre-trained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Ivan Vulić, et al. 2019. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.
- John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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)*, pages 4129–4138.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.
- Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. *arXiv preprint arXiv:2004.03061*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Douglas A Reynolds. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Hannes Schulz, Jeremie Zumer, Layla El Asri, and Shikhar Sharma. 2017. A frame tracking model for memory-enhanced dialogue systems. *arXiv preprint arXiv:1706.01690*.

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

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrette Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019a. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogues. *arXiv preprint arXiv:2004.06871*.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

OOS Intent
'translate', 'transfer', 'timer', 'definition', 'meaning_of_life', 'insurance_change', 'find_phone', 'travel_alert', 'pto_request', 'improve_credit_score', 'fun_fact', 'change_language', 'payday', 'replacement_card_duration', 'time', 'application_status', 'flight_status', 'flip_coin', 'change_user_name', 'where_are_you_from', 'shopping_list_update', 'what_can_i_ask_you', 'maybe', 'oil_change_how', 'restaurant_reservation', 'balance', 'confirm_reservation', 'freeze_account', 'rollover_401k', 'who_made_you', 'distance', 'user_name', 'timezone', 'next_song', 'transactions', 'restaurant_suggestion', 'rewards_balance', 'pay_bill', 'spending_history', 'pto_request_status', 'credit_score', 'new_card', 'lost_luggage', 'repeat', 'mpg', 'oil_change_when', 'yes', 'travel_suggestion', 'insurance', 'todo_list_update', 'reminder', 'change_speed', 'tire_pressure', 'no', 'apr', 'nutrition_info', 'calendar', 'uber', 'calculator', 'date', 'carry_on', 'pto_used', 'schedule_maintenance', 'travel_notification', 'sync_device', 'thank_you', 'roll_dice', 'food_last', 'cook_time', 'reminder_update', 'report_lost_card', 'ingredient_substitution', 'make_call', 'alarm', 'todo_list', 'change_accent', 'w2', 'bill_due', 'calories', 'damaged_card', 'restaurant_reviews', 'routing', 'do_you_have_pets', 'schedule_meeting', 'gas_type', 'plug_type', 'tire_change', 'exchange_rate', 'next_holiday', 'change_volume', 'who_do_you_work_for', 'credit_limit', 'how_busy', 'accept_reservations', 'order_status', 'pin_change', 'goodbye', 'account_blocked', 'what_song', 'international_fees', 'last_maintenance', 'meeting_schedule', 'ingredients_list', 'report_fraud', 'measurement_conversion', 'smart_home', 'book_hotel', 'current_location', 'weather', 'taxes', 'min_payment', 'whisper_mode', 'cancel', 'international_visa', 'vaccines', 'pto_balance', 'directions', 'spelling', 'greeting', 'reset_settings', 'what_is_your_name', 'direct_deposit', 'interest_rate', 'credit_limit_change', 'what_are_your_hobbies', 'book_flight', 'shopping_list', 'text', 'bill_balance', 'share_location', 'redeem_rewards', 'play_music', 'calendar_update', 'are_you_a_bot', 'gas', 'expiration_date', 'update_playlist', 'cancel_reservation', 'tell_joke', 'change_ai_name', 'how_old_are_you', 'car_rental', 'jump_start', 'meal_suggestion', 'recipe', 'income', 'order', 'traffic', 'order_checks', 'card_declined', 'oos'

Table 4: OOS intent

Name	# Dialogue	# Utterance	Avg. Turn
MetaLWOZ	37,884	432,036	11.4
Schema	22,825	463,284	20.3
Taskmaster	13,215	303,066	22.9
MWOZ	10,420	71,410	6.9
MSR-E2E	10,087	74,686	7.4
SMD	3,031	15,928	5.3
Frames	1,369	19,986	14.6
WOZ	1,200	5,012	4.2
CamRest676	676	2,744	4.1

Table 5: The data statistics is from Wu et al. (2020).

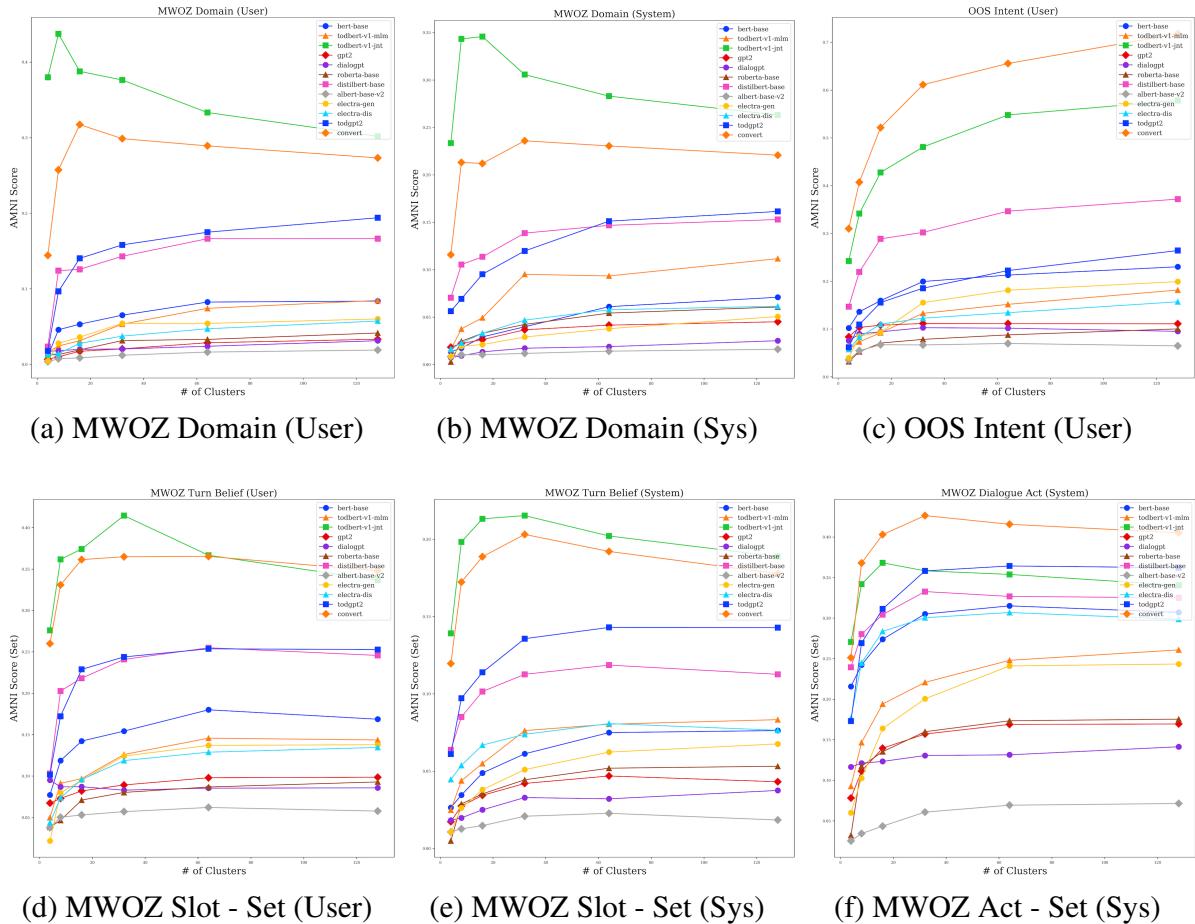


Figure 6: The ANMI evaluation of pre-pretrained models with domain, intent, slot, and action labels using GMM. (Best view in color)

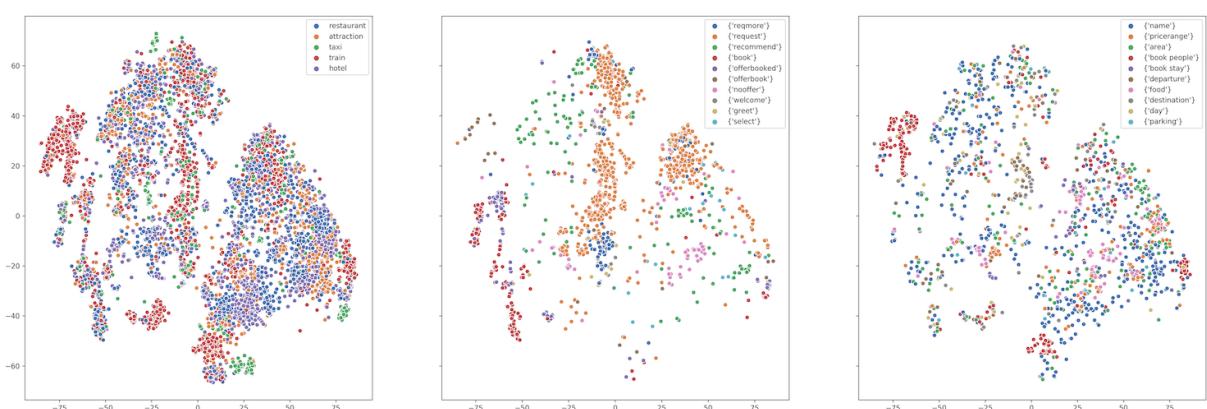


Figure 7: The tSNE visualization of dialogue representations from the ToD-BERT-jnt. (Best view in color.)

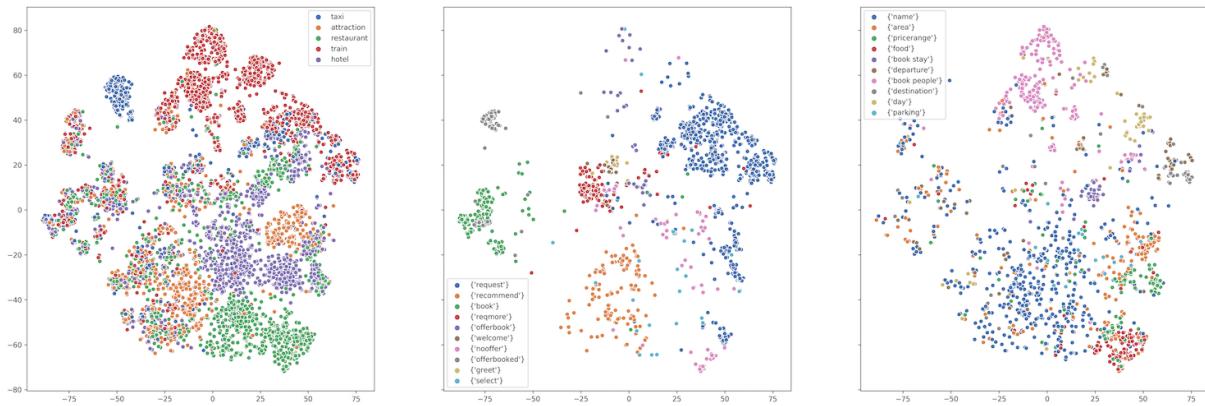


Figure 8: The tSNE visualization of dialogue representations from the ConveRT. (Best view in color)

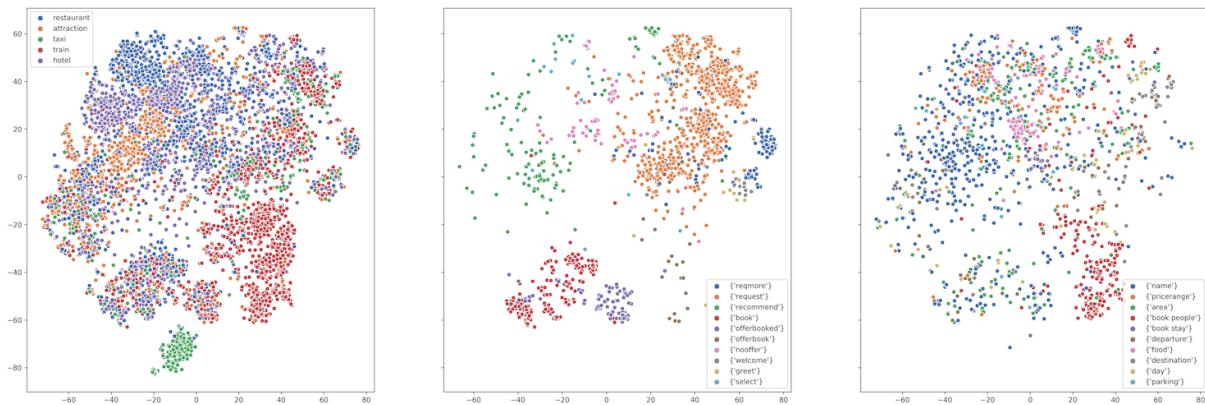


Figure 9: The tSNE visualization of dialogue representations from the DistilBERT. (Best view in color.)

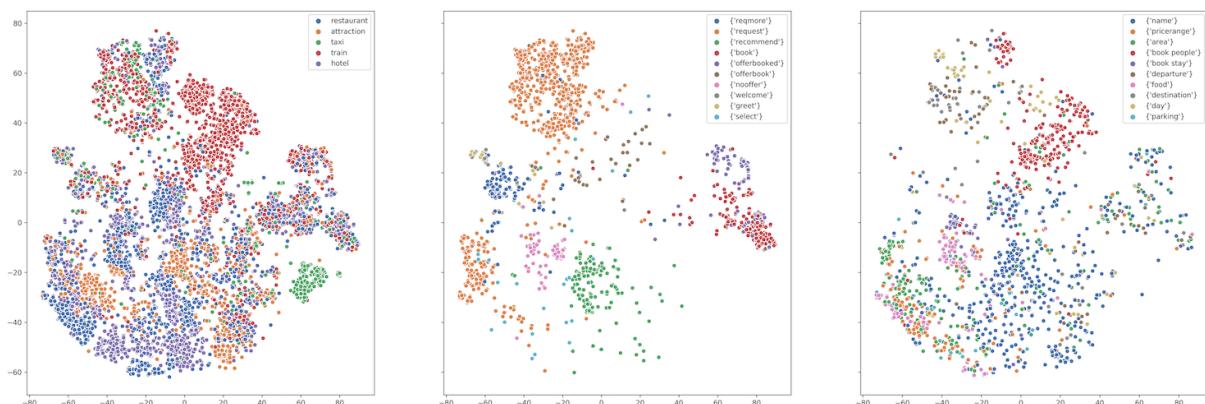


Figure 10: The tSNE visualization of dialogue representations from the ToD-GPT. (Best view in color.)



Figure 11: The tSNE visualization of dialogue representations from the ELECTRA-Dis. (Best view in color.)

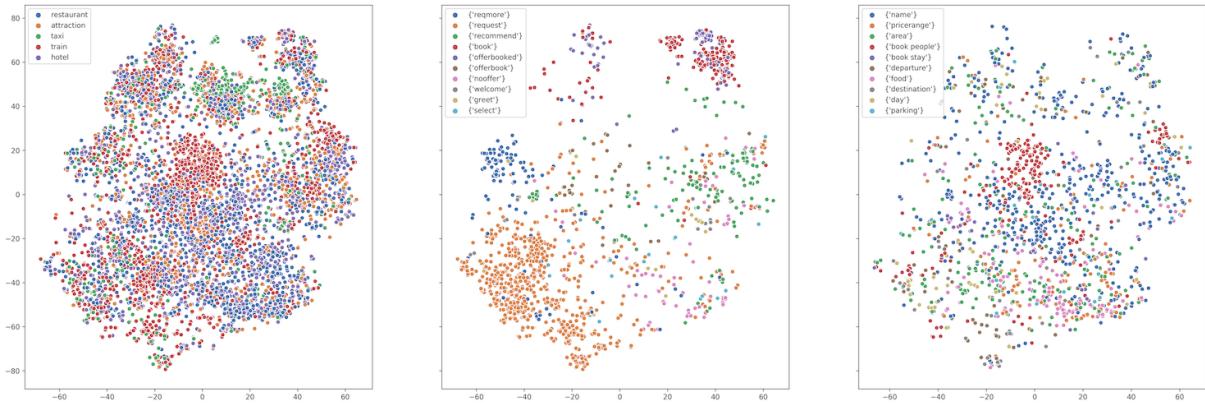


Figure 12: The tSNE visualization of dialogue representations from the ELECTRA-Dis. (Best view in color.)

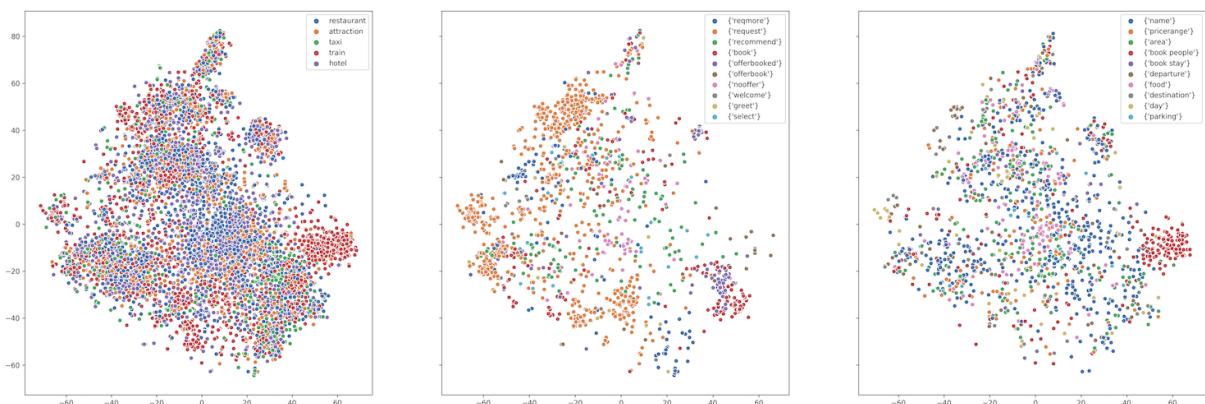


Figure 13: The tSNE visualization of dialogue representations from the RoBERTa. (Best view in color.)

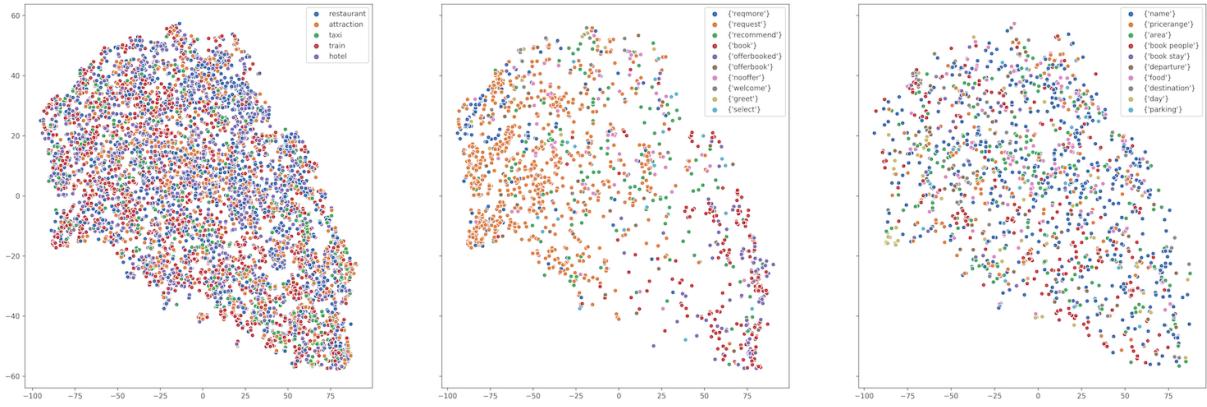


Figure 14: The tSNE visualization of dialogue representations from the DialoGPT. (Best view in color.)



Figure 15: The tSNE visualization of dialogue representations from the ALBERT. (Best view in color.)

TOD-BERT-jnt	
Cluster 1 (Restaurant Request)	i have many options available for you ! is there a certain area or cuisine that interests you ? there are 21 restaurant -s available in the centre of town . how about a specific type of cuisine ? do you have any specific type of food you would like ? there 33 place -s that fit your criteria . do you have a particular cuisine type in mind so that i can narrow the results down ? is there a particular cuisine you are looking for ?
Cluster 2 (Taxi/Train)	what time do you want to leave and what time do you want to arrive by ? do you have a time preference ? when would you like to leave and arrive ? what time would you like to leave the junction ? wonderful , i can help you . what time on sunday would you like to depart ?
Cluster 3 (Attraction Recommend)	i can recommend the allenbell . it's in the east , is cheap yet has a 4 star rating and free wifi and parking . can i help you book ? the university arms is an expensive , 4 star hotel with free wifi . comparatively , the alexander bed and breakfast is a cheap -ly priced guesthouse , also 4 stars . i have found the guesthouse you were wanting . would you like me to book this for you ? how about the express by holiday inn cambridge . it's in the east . the expensive i is actually not much more than the other 2 . i would highly recommend it . that would be at the express by holiday inn cambridge . it's in the east .
Cluster 4 (Hotel Inform)	the address is hills road city centre . their address is unit g6 , cambridge leisure park , clifton road , the postcode is cb17dy . the address is corn exchange street . is there anything else i can help you with ? yes , the phone number is 01223277977 . the address is hotel felix whitehouse lane huntingdon road , and the post code is cb30lx . want to book ? the bridge guest house is at 151 hills road and their number is 01223247942 .
Cluster 5 (Welcome/End)	you are welcome . is there anything else i can help you with today ? great , is there anything else that you need help with ? is there anything else that you would like ? no problem . can i help you with anything else ? is there something else i can help you with then ?

Table 6: Clustering results of the TOD-BERT-jnt model. The samples are randomly picked from each randomly selected five clusters (using K=32).

GPT2	
Cluster 1	there are 9 indian restaurant -s in centre what price range do you want ? do you have any specific type of food you would like ?
	105 minutes is the total travel time . can i help you with anything else ?
	there are lots to choose from under that criteria . what day would you like to travel on ?
	i have the cote in the centre . it is in the expensive range . would you like to make a booking ?
Cluster 2	your reference number is x5ny66zv . i booked tr3932 , reference number is fiw5abo2 .
	nusha is in the south , and the phone number is 01223902158 .
	they are located at 12 lensfield road city centre , postcode cb21eg , and phone number 01842753771 .
	it would cost 16.50 pounds .
Cluster 3	i hope i have been of help the entrance fee is free . anything else i can do for you today ?
	sure , lookout for a blue volvo the contact number is 07941424083 . can i help with anything else ?
	1 moment while i try to make the reservation of table for 8 , friday at 16:30 .
	i have 3 options for you 2 in the north in the moderate price range and 1 that s expensive in the east .
Cluster 4	when would you like to leave and arrive ? booking was unsuccessful . can you try another time slot ?
	on what day will you be traveling ?
	tr3823 will arrive at 16:55 , would that work for you ?
	okay , what day did you have in mind ?
Cluster 5	saffron brasserie is an expensive restaurant that serves italian food there are 21 restaurant -s available in the centre of town . how about a specific type of cuisine ?
	i have 5 different restaurant -s to choose from . there are 4 in the centre of town , and 1 in the west . do you have a preference ?
	i have about 5 different entertainment venue -s if that is what you are looking for . do you have a preference on the area its located in ?
	there are no colleges close to the area you are requesting , would you like to chose another destination ?

Table 7: Clustering results of the GPT2 model. The samples are randomly picked from each randomly selected five clusters (using K=32).

DialoGPT	
Cluster 1	it is located in jesus lane your booking was successful , the reference number is waeyaq0m . may i assist you with anything else today ?
	your booking is successful ! your reference number is iigra0mi . do you need anything else ?
	1 moment while i try to make the reservation of table for 8 , friday at 16:30 .
	this booking is successful for 1 night . your reference number is 85bgkwo4 . is there anything else i can assist you with ?
Cluster 2	sure , how many days and how many people ? i recommend castle galleries and it s free to get in !
	i have plenty of trains departing from leicester , what destination did you have in mind ?
	i have 5 colleges in the centre area . what specific college are you looking for ?
	oh yes quite a few . which part of town will you be dining in ?
Cluster 3	i have many options available for you ! is there a certain area or cuisine that interests you ? there are lots to choose from under that criteria . what day would you like to travel on ?
	actually all 5 have free wifi . what star rating would you like ?
	i have found the guesthouse you were wanting . would you like me to book this for you ?
	yes , the hamilton lodge has internet .
Cluster 4	its entrance fee is free . sure , lookout for a blue volvo the contact number is 07941424083 . can i help with anything else ?
	how many people is the reservation for ?
	how about train tr3934 ? it leaves at 12:34 and arrives at 13:24 . travel time is 50 minutes .
	sure , the phone number is 01223902112 and they are in postcode cb58sx . can i help you with anything else today ?
Cluster 5	yes i can . what restaurant are you looking for ? what time would you like to leave the junction ?
	no problem . can i help you with anything else ?
	you are welcome . is there anything else i can help you with today ?
	is there anything else i can help you with ?

Table 8: Clustering results of the DialoGPT model. The samples are randomly picked from each randomly selected five clusters (using K=32).