GRTr: Generative-Retrieval Transformers for Data-Efficient Dialogue Domain Adaptation

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, Hannes Schulz

Abstract—Domain adaptation has recently become a key problem in dialogue systems research. Deep learning, while being the preferred technique for modeling such systems, works best given massive training data. However, in real-world scenarios, such resources are rarely available for new domains, and the ability to train with a few dialogue examples can be considered essential. Pre-training on large data sources and adapting to the target data has become the standard method for few-shot problems within the deep learning framework. In this paper, we present GRTr, a hybrid generative-retrieval model based on the large-scale general-purpose language model GPT-2 fine-tuned to the multi-domain MetaLWoz dataset. In addition to robust and diverse response generation provided by the GPT-2, our model is able to estimate generation confidence, and is equipped with retrieval logic as a fallback for the cases when the estimate is low. GRTr is the winning entry at the fast domain adaptation task of DSTC-8 in human evaluation (4% improvement over the 2nd place system). It also attains superior performance to a series of baselines on automated metrics on MetaLWoz and MultiWoz, a multi-domain dataset of goal-oriented dialogues. In this paper, we also conduct a study of GRTr’s performance in the setup of limited adaptation data, evaluating the model’s overall response prediction performance on MetaLWoz and goal-oriented performance on MultiWoz.

Index Terms—Deep learning, neural networks, natural language processing, dialogue systems, domain adaptation

1 INTRODUCTION

Goal-oriented dialogue is an area of increasingly high interest, both from academic and industrial perspectives. Data-driven approaches to developing such systems [1] proved to be more flexible and scalable to various scenarios and domains than previous techniques widely employed in industry, mostly based on expert knowledge. The benefits of methods based on machine learning (especially deep learning) can only be experienced when there are excess amounts of training data available; however, in real-world scenarios, there’s only a small amount of initial data available for a new domain. Training techniques must make the most of this small data, i.e. work in a data-efficient way, in order to enable rapid development of dialogue models for an ever-increasing number of domains and tasks. The most promising method to achieve this under the deep learning framework has become transfer learning where a large, generic model is first trained from a highly represented source of data, after which it gets adapted to the target task.

In this paper, we explore this problem through the Eighth Dialog System Technology Challenge (DSTC), Fast Domain Adaptation task. Specifically, we propose a hybrid generative/retrieval dialogue model leveraging knowledge transfer from a large-scale pre-trained general-purpose language model. Our model is able to maintain goal-oriented dialogue in a closed domain having only been exposed to a small set of in-domain dialogues as the domain description. Our hybrid model achieves state-of-the-art performance on the MetaLWoz dataset when evaluated with human judges, and attains competitive generalization level in adapting to goal-oriented MultiWoz dataset unseen at the main training stage. Automated word overlap-based metrics demonstrate that it outperforms a series of competitive baselines — both generation-only and retrieval-only models.

2 RELATED WORK

Dialogue response generation is an actively researched area, with the sequence-to-sequence (seq2seq) model [1] gaining wide adoption in both chat-oriented [5] and goal-oriented dialogue [6]. Initially these architectures were based on Recurrent Neural Networks such as LSTM [7] or GRU [8] which were quite challenging to train on large amounts of conversational data, causing researchers to focus on improving response diversity [9] and the overall dialogue consistency [10]. Quite recently, self-attention mechanisms, like those used in the Transformer [11], have been adopted for conversation models — together with large-scale pre-training, it resulted in a new generation of seq2seq architectures.

The data efficiency of dialogue systems has also been extensively researched in the past. Initially, when modular dialogue system architecture was the prevalent approach, dialogue managers and state trackers were the components that data-efficient methods were applied to the most. As such, the dialogue state tracker domain adaptation task was initially proposed in DSTC-3 [12] — that challenge featured approaches like Bayesian Processes [13] and Recurrent Neural Networks [14]. Later research was focused on data-efficiency of dialogue managers, for instance Williams et al. [15] introduced a model designed for bootstrapping from limited training data and further fine-tuning in the reinforcement learning fashion. Furthermore, a recent paper by Vlasov et al. [16] proposed a dialogue management model...
which used a unified embedding space for user and system turns allowing for efficient cross-domain knowledge transfer.

End-to-end dialogue response generation, the technique that followed modular architectures with the arrival of large conversational datasets, was also eventually approached in a data-efficient way. One such method used prior linguistic knowledge to improve zero-shot performance: Eshghi et al. [17] proposed a linguistically informed model based on an incremental semantic parser [18] combined with a reinforcement learning-based agent. The parser was used for both maintaining the agent’s state and pruning the agent’s incremental, word-level generation actions to those leading to syntactically correct word sequences. While outperforming end-to-end dialogue models on bAbI Dialog Tasks [19] in the extreme zero-shot case [20], this method inherited the limitations of the dialogue grammar — specifically, it is limited to a single closed domain until a wide-coverage grammar is available.

Zhao and Eskenazi [21] introduced zero-shot dialogue generation (ZSDG) framework under which a dialogue system was trained on dialogues from several source domains and a small amount of annotated utterances from the target domain. The key feature in their framework was the unified latent space which was used to encode user’s queries, dialogue contexts, and annotations.

Later, Shalyminov et al. [22], [23] proposed Dialogue Knowledge Transfer Networks which approached the problem in a few-shot setup with a separate out-of-domain pre-training stage on a large goal-oriented corpus (MetaLWOZ, [24]). In those approaches, MetaLWOZ was used as source dataset for transfer, whereas we treat it as the target dataset. While the authors used full target-domain dialogues, they ended up using only a fraction of ZSDG’s data in terms of the number of utterances.

More generally, transfer learning has been widely adopted for natural language problems with the emergence of large-scale pre-trained text representation models like ELMO [25], BERT [26], and GPT-2 [2]. When applied to dialogue response generation, the most successful approaches made use of a Transformer for chat-oriented dialogue [27] and GPT/GPT-2 for goal-oriented dialogue [28]. Our approach is based on a similar technique, though in addition to fine-tuning a pre-trained model to our task, we augment the generative model with a retrieval component in a hybrid approach.

Finally, another recent approach applied to the problem of few-shot dialogue generation is meta-learning [29], under which the task is split into multiple subtasks corresponding to dialogue domains. For each of them, a specialized dialogue model was trained, with their training progress then merged into the main model. In general, the intuition behind meta-learning is training a base model which would be best suited for data-efficient fine-tuning — otherwise known as rapid adaptation — making the most efficient gradient updates from the few data points available in the target domain.

### Fast Dialogue System Domain Adaptation

Goal-oriented dialogue systems can be challenging to bootstrap: for a new domain, little data is available to train, e.g., a natural language understanding (NLU) module or other parts of the pipeline. Often, a Wizard-of-Oz (WOZ, [30], [31]) schema can be used to obtain some initial test data, however, this requires training human agents for the task and setting up a complex pipeline. The value of WOZ data is limited, since “users” are mostly hired and might not conform to...
real users. Additionally, any change in the chatbot interface requires collecting more data.

### 3.1 DSTC-8, Fast Domain Adaptation Task

In the Eighth Edition of DSTC, its Domain Adaptation task focuses on building a model that predicts user responses for a goal-oriented dialogue system for which only limited in-domain data is available. Such data could be collected from (e.g., customer service transcripts) or written by the developers themselves. From this in-domain data, the support set, one would like to extrapolate responses to novel dialogue contexts (the target). However, the support set is typically too small to train a dialogue response generation model. Instead, the approach assumed in the challenge is to adapt (or fine-tune) a generic dialogue model trained on a large corpus of conversations over multiple source domains.

MetaLWOZ, the main dataset for the task (to be described in detail later in Section 5) represents natural cooperative information-seeking dialogues between two humans in a variety of domains and tasks within those, and is not annotated with any conventional goal-oriented markup. And the domain adaptation task is focused on modeling the user side in the dialogue. This can be considered a more challenging task than predicting the system’s responses since normally, user’s utterances are less predictable and more varied, and a prospective successful dialogue model should have the underlying user’s goal in it to generate relevant queries to the system. The desired properties of the hypothetical ideal model are reflected in the Task’s evaluation procedure combining human judgments and a series of automatic metrics (see Section 7 for more detail). Possible applications of an adaptive user-side dialogue model include Reinforcement Learning-based setups which are highly dependent on the quality of the user simulator, as well as data augmentation-based training setups for improved robustness and coverage.

Technically, the problem setup is as follows: having trained the base model on the source domains, the model is then fed with one target dialogue and a support set at a time. The model’s task is to predict the next user turn of the target dialogue, taking into account the support set before producing a prediction. At prediction time, each target dialogue is processed in isolation from other target dialogues, such that the model cannot use knowledge or state obtained from other target/support data.

### 4 Proposed model

We use a language model pre-trained on a very large and diverse collection of textual data providing a strong language and then adapt the model for our tasks in the form of fine-tuning. Our base model is GPT-2 [27], a transformer-based language model. In order to adapt GPT-2 for dialogue generation, we first augment the input embedding for each token in the dialogue with (1) a speaker tag embedding identifying the speaker and (2) a turn embedding, identifying the turn number in the current dialogue. These additional embedding matrices are learned solely using the dialogue data. The input token embeddings are then obtained by summing up these representations. We also add two task-specific output layers (or “heads”) for our purposes: a language modeling (LM) head and a next-sentence prediction (NSP) classification head, both trained from randomly initialized parameters.

We fine-tune GPT-2 for response generation by minimizing the negative log-likelihood of response tokens given the concatenation of dialogue context and the previous tokens in the response,

\[
\mathcal{L}_{LM} = - \log P_{LM}(x | C) = - \sum_{i=1}^{[X]} \log P_{LM}(x_i | x_{i-1}, ..., x_1, C),
\]

where \(X\) is the response and \(C\) is the dialogue context, i.e. the concatenation of the tokens in the previous utterances.

To predict the next sentence, we proceed as follows: given a context/response pair \((C, X)\), the classification head is trained to produce a binary label \(y\), which is 1 if \(X\) is the correct response given the context \(C\), and 0 if \(X\) is a distractor.
As every test dialogue in the target domain/task is accompanied with a small support set of dialogues from the same domain/task, we make use of this data by further fine-tuning the dialogue model on the support dialogues. Crucially, we make sure not to accumulate any information between test dialogues: after each fine-tuning on the support set, we reset the weights of the model to the dialogue prior obtained by the fine-tuning stage described in the previous section.

In order to add diversity to the responses, GPT-2 uses nucleus (top-\(p\)) sampling \cite{28} during generation, i.e. the model’s vocabulary \(V\) is pruned to \(V^p\), the smallest set such that
\[
\sum_{x \in V^p} p(x | x_{1:t-1}, C) \geq p, \tag{4}
\]
and the final distribution from which the words are sampled is rescaled as follows:
\[
P'(x | x_{1:t-1}) = \begin{cases} \frac{P(x | x_{1:t-1}, C)}{\sum_{x \in V^p} P(x | x_{1:t-1}, C)} & \text{if } x \in V^p \\ 0 & \text{otherwise} \end{cases} \tag{5}
\]

### 4.2 Hybrid generative-retrieval prediction

In our experiments, we found that retrieval baselines are quite effective in the automatic metrics considered. Combining retrieval techniques with our generative model in a hybrid approach produced a stronger model.

The retrieval component is set up as follows (see Algorithm 1): when predicting the \(t\)-th turn of the test dialogue, the model embeds its context of length \(t-1\) as well as all the support dialogue contexts of the same length \(t-1\) using the fine-tuned dialogue encoder. The encoding for the dialogue context is the hidden state of the last layer of the Transformer model at the position corresponding to the last token in the context. Then, it selects the nearest support context to the target context and picks its next turn as the retrieved candidate response. The responses considered as retrieved candidates lie within the retrieval window of a fixed length centered at the target response position, \(t\).

Finally, the model’s own generated response and the best retrieved candidate response are ranked using the NSP classification head, i.e. both responses are concatenated with the ground-truth context and the one with the higher \(P_{\text{NSP}}\) (Eq. 3) is selected. The model is visualized in Figure 1.
We use MetaWOZ, the dataset for DSTC-8 Track 2 “Fast Domain Adaptation” [24]. It contains more than 37,000 human-human dialogues spanning the total of 227 tasks in 47 domains. The dialogues are collected in a Wizard-of-Oz style: human participants were assigned the role of bot or user, then given a problem domain and related specific task, and instructed to reach the user’s goal over at least 10 dialogue turns. At the evaluation time, the data is organized in tuples as follows [35]:

```
{Train-Inform: [('Leave', '10:00')], 'Train-Request': [('Id', '?')]
```

### 5 Datasets

We use MetaWOZ, the dataset for DSTC-8 Track 2 “Fast Domain Adaptation” [24]. It contains more than 37,000 human-human dialogues spanning the total of 227 tasks in 47 domains. The dialogues are collected in a Wizard-of-Oz style: human participants were assigned the role of bot or user, then given a problem domain and related specific task, and instructed to reach the user’s goal over at least 10 dialogue turns. At the evaluation time, the data is organized in tuples as follows [35]:

```
{Train-Inform: [('Leave', '10:00')], 'Train-Request': [('Id', '?')]
```

### 6 Baselines and Competing Models

We compare our hybrid model to the retrieval baselines provided by the DSTC-8 organizers. The baselines ignore the training data and rely solely on the support sets: they embed each support dialogue’s context and find the one nearest to the target context using cosine distance as the metric. They then return the turn following the identified context as the predicted response. There are two retrieval-only baselines, which differ in their encoder: (1) BERT-based, taken off-the-shelf, and (2) SentencePiece/FastText-based, modeled after [34] with embeddings pre-trained on the Reddit Conversations corpus.

Another baseline provided is a generation-only model, a bidirectional LSTM-based model [5] trained on MetaWOZ.

All the submissions at the final stage of the challenge are as follows [35]:

- **Team A** trained a BiLSTM on the provided Reddit corpus, then fine-tuned the model at test-time using a mixture of MetaWOZ and MultiWOZ support dialogues, augmented to the context of the target dialogue, and dynamically-sampled Reddit threads,
- **Team B** — the work described in this chapter,
- **Team C** first fine-tuned GPT-2 on the MetaWOZ training corpus, then fine-tuned it further on the support sets of the MetaWOZ and MultiWOZ test sets,
- **Team D** trained a BiLSTM encoder and attentional LSTM decoder on both Reddit and MetaWOZ training corpora, without any fine-tuning to the test sets.

### 7 Experimental setup and evaluation

We perform training in two stages: training of the base model and fine-tuning it to the target dialogue’s support set. At the first stage, we train the model for the maximum of 5 epochs with early stopping. The second stage goes on for 1 epoch in the interest of time. GPT-2 models use the context of 3 exchanges, or 5 turns: bot-user-bot-user-bot, predicting the next user’s utterance. We mainly used the ‘small’ GPT-2 checkpoint by HuggingFace — we also tried the ‘medium’ one, but found no improvement with it in our task.
The main systems’ goal is to generate appropriate responses towards maintaining a natural cooperative dialogue on the user’s side, so the main evaluation is performed involving human judges. Specifically, Amazon Mechanical Turk workers were tasked to compare the candidate responses given the dialogue context. Each comparison was pairwise between the results of two systems presented in random order. Judges ranked the responses against the following criteria:

- **Usefulness** — whether the response is useful given the dialogue context and the user’s overall final goal,
- **Informativeness** — whether the response specifically contains information relevant to the conversation,
- ** Appropriateness** — whether the response is appropriate (on-topic, of a reasonable length, not repetitive) to the conversation,
- **Ease of answering** — given a hypothetical conversational bot on the system side, whether the response will be a valid input for it and presumably straightforward to process.

For each pairing, three independent comparisons were performed against each metric. The number of comparisons required was reduced by letting the Multisort algorithm [36] determine which responses to compare, causing more similar systems with similar performance to be compared more often with each other. Bootstrapping over the 100 randomly chosen dialogue contexts was used to determine average ranks and assess the ranking robustness [37].

### 7.2 Automatic evaluation

In addition to human evaluation, we also assess model performance using automatic metrics. The models were evaluated on Meta1WOZ against word-overlap metrics such as BLEU-1–3, CIDER, METEOR, ROUGE-L using the NLGEval package [38]. Although not ideal for the specifics of dialogue and spoken language in general [39], [40], such metrics approximate the overall quality of a response generation model and are especially useful for intermediate evaluation.

As was mentioned in Section [5], at this stage, we have two Meta1WOZ setups: pure task and cross-task — we evaluate them independently.

We also perform additional evaluation of Entity/Intent F1 of the Multi1WOZ dataset in pure task mode with pre-trained NLU taggers from the ConvLab package [41]. There is no Multi1WOZ data available at the first stage (base model training), so all the exposure our model has to this dataset is via support dialogues. Complementary to Meta1WOZ evaluation, this stage is designed for assessing the models’ goal-oriented performance.

### 7.3 Few-shot evaluation

We are also interested in evaluating our model’s performance in the extremely limited-data setup. For that, we run adaptation & prediction with GRT on 8–64 support dialogues. We sample those dialogues from the original DSTC-8 support sets several times and perform independent runs of adaptation & prediction with the corresponding few-shot support sets. Specifically, in this paper we report the average performance over 3 such runs against all our automated metrics: word overlap on Meta1WOZ pure/cross task and Intent/Intent+Slot F1 score on Meta1WOZ.

### 8 Results and Discussion

#### 8.1 Human evaluation

Results of pairwise comparisons are shown in Table 1. Our GRT system’s responses (Team B) were preferred by the judges in 56% of direct comparisons. This surpasses the next best system (Team C) performance by more than 4%, with only the gold human responses being chosen more frequently.

Furthermore, from the bootstrap ranking distribution (Figure 2), we see that, apart from the gold human responses, our model’s outputs are consistently preferred over other submissions by the judges. Of all metrics used, the most notable are ‘appropriateness’ and ‘usefulness’. On the former, GRT responses have the second visible peak at rank 1.
competing with gold responses. On usefulness however, rank 1 is held by the gold responses with no variation, and our model has the second visible peak at rank 3, thus almost tying with Team C.

8.2 Automatic evaluation

Results on MetaWOZ and MultiWOZ against automatic evaluation metrics are shown in Figures 3 and 4 respectively. We observe that the retrieval baselines attain very competitive performance on both datasets, with FastText embeddings from Reddit leading to overall better results than off-the-shelf BERT, especially in the pure task setting.

With GRT, we performed an ablation study to have a closer look into its performance. We evaluated three versions:

- GPT-2 base — generation-only model trained on MetaWOZ which ignores all the support data,
- GPT-2 + sup — generation-only model fine-tuned to the support dialogues
- GRT base — the hybrid approach not fine-tuned to support data
- GRT@N — our full hybrid approach with different retrieval window lengths (we experimented with N=1, 3, and 5).

As seen in Figure 3, there is strong dependence on support dialogues (‘base’ vs. ‘+sup’) as the base model mostly struggles to compete with the baselines. Adding retrieval logic (GRT vs. ‘+sup’) results in further performance gains. HRED and GPT-2 base, the two models that did not use support dialogues, had comparable performance on MetaWOZ. We note that expanding the retrieval window (GRT@1 vs GRT@3 vs GRT@5) has most benefit on MetaWOZ pure task as the support responses are sufficiently similar to the target ones, while on the cross-task dataset this emphasis on the retrieval harms the model’s accuracy.

In goal-oriented metrics on MultiWOZ (see Figure 4), the same performance pattern is observed with retrieval models, but GPT-2 in the generation-only version performs surprisingly better when not fine-tuned to support set (‘base’). On the other hand, the hybrid model experiences even more performance gain than on MetaWOZ. Presumably, generating responses for this dataset is harder due to the fact that it is not represented at the main training stage, and there is not much utterance overlap with MetaWOZ, so little knowledge transfer takes place in this experiment. Correspondingly, expanding the retrieval window sensibly improves the performance. Compared to other submissions, we observe that GRT still outperforms most of the competitors and only gives way to Team A’s system. We hypothesize here the best MultiWOZ model (Team A) was fitted to the automatic evaluation metrics too tightly, with the negative side effect observable in human evaluation results of Table 1 and Figure 2 where this system was prevalently ranked 4th and 5th.

8.3 Few-shot evaluation

The results with limited amounts of support data are shown in Figure 5. As we can see, in the setups when the target dialogues are similar to the support ones — MetaWOZ pure task and MultiWOZ — more support data positively affects GRT’s performance both in word overlap and in goal-oriented Intent/Intent+Slot F1 metrics. On the other hand, on MetaWOZ cross-task, where support dialogues share less similarity with the target ones, there is no more clear performance trend. It can be argued that MetaWOZ cross-task resembles more of a zero-shot setup, where the model doesn’t get exposed to the ‘true’ target-domain dialogues no matter the size of the support set.

8.4 Retrieval and Generation Frequency

In Table 4 we show per-domain ratios of retrieved/generated responses from the hybrid model. We find that the majority of the responses are generated, and the retrieval logic works as the fallback option. On MetaWOZ, which the model had more exposure during the training, generated responses ratio is generally slightly higher than that on MultiWOZ which was only seen by the model via support dialogues. Consequently, the model’s overall confidence on this dataset is lower, which results in more frequent fallbacks.

In Figure 6 we plot how the retrieval ratio of GRT@1 behaves in the few-shot setup with different sizes of the support set. We see that (1) in both MetaWOZ setups, the model falls back to retrieval significantly less often than on MultiWOZ when the support data is strictly limited (8—16 dialogues), and (2) on MultiWOZ, we don’t see a clear trend for the increase in the retrieval ratio with more support data as we see on MultiWOZ pure and cross-task — instead, the ratio fluctuates around an overall high level. The first observation corresponds to what we saw in Table 4, and the second observation can potentially explain the result for MetaWOZ cross-task in Figure 5: the model keeps retrieving from support data proportionally to its amount, although retrieved responses from other tasks might not always be applicable for target task.

Overall, we observe in Table 5 that there are many cases in the data where the gold response cannot possibly be inferred...
We present GRT, a hybrid generative-retrieval model based on GPT-2 with fast domain adaptation via transfer learning. It attains robust and diverse language generation performance across domains, and employs retrieval logic as a fallback for the cases of low generation confidence. The latter is estimated using a trainable component, part of the unified multi-task learning setup.

Our method is the winning entry at the fast domain adaptation task of DSTC-8 as evaluated by human judges. In additional automatic evaluation, it also outperforms a series of baselines on MetaWOZ and MultiWOZ.

The additional few-shot evaluation we conducted in this paper demonstrates how GRT is able to adapt to strictly limited data, with 8—64 support dialogues — which is closer to a real-world case of dialogue system bootstrapping. We observed clear trends on the model’s performance improvement with more support dialogues on MetaWOZ pure task and MultiWOZ. The third setup we evaluated — MetaWOZ cross-task — did not demonstrate as much contribution of the support data, partly because it represents a different, essentially zero-shot setup, with the target dialogue being less similar to the support ones.

Overall, it is evident that the problem of data-efficient dialogue response prediction needs further research, and one promising direction that we are going to explore in our own future work is the meta-learning framework [29], or ‘learning to fine-tune’. Based on splitting the task into multiple subtasks and solving them with separate versions of the model and continuously merging models of individual learners, meta-learning approach naturally fits our multi-domain setup and may lead to improved fine-tuning performance.

### 9 Conclusion and future work

We present GRT, a hybrid generative-retrieval model based on GPT-2 with fast domain adaptation via transfer learning.

<table>
<thead>
<tr>
<th>Dataset / domain</th>
<th>Generated (%)</th>
<th>Retrieved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MetaWOZ pure task</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>booking flight</td>
<td>64.10</td>
<td>35.90</td>
</tr>
<tr>
<td>hotel reserve</td>
<td>63.80</td>
<td>36.20</td>
</tr>
<tr>
<td>tourism</td>
<td>57.40</td>
<td>42.60</td>
</tr>
<tr>
<td>vacation ideas</td>
<td>61.70</td>
<td>38.30</td>
</tr>
<tr>
<td><strong>MetaWOZ cross task</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>booking flight</td>
<td>68.20</td>
<td>31.80</td>
</tr>
<tr>
<td>hotel reserve</td>
<td>74.80</td>
<td>25.20</td>
</tr>
<tr>
<td>tourism</td>
<td>73.90</td>
<td>26.10</td>
</tr>
<tr>
<td>vacation ideas</td>
<td>74.70</td>
<td>25.30</td>
</tr>
<tr>
<td><strong>MultiWOZ</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>attraction</td>
<td>55.60</td>
<td>44.40</td>
</tr>
<tr>
<td>hospital</td>
<td>60.00</td>
<td>40.00</td>
</tr>
<tr>
<td>hotel</td>
<td>63.00</td>
<td>37.00</td>
</tr>
<tr>
<td>police</td>
<td>52.10</td>
<td>47.90</td>
</tr>
<tr>
<td>restaurant</td>
<td>61.30</td>
<td>38.70</td>
</tr>
<tr>
<td>taxi</td>
<td>64.30</td>
<td>35.70</td>
</tr>
<tr>
<td>train</td>
<td>61.00</td>
<td>39.00</td>
</tr>
</tbody>
</table>

This is reflected in the way human judges are asked about response quality.
APPENDIX A

PAIRWISE DISTANCES BETWEEN GENERATED/RETRIEVED RESPONSES

Fig. 7: A histogram of pairwise distances between generated and retrieved GPT candidates — Meta1WOZ pure task (other datasets’ distribution is similar)

REFERENCES


Igor Shalyminov is a PhD student at Heriot-Watt University, UK. His research focuses on data-efficient methods for dialogue systems, including few-shot & zero-shot training, representation learning and transfer, and robustness to anomalous input.

Alessandro Sordoni is a Researcher at Microsoft Research. His research interest lies in the field of natural language processing, with a focus on unsupervised learning of natural language representations, meta- and data-efficient learning, grammar induction and robust out-of-distribution generalization.

Adam Atkinson is an Applied Scientist at Microsoft Turing. His research interests include dialogue systems, meta-learning, and large-scale deep learning for language applications.

Hannes Schulz is a Researcher at Microsoft Research. His research interests include low-resource dialogue systems, unsupervised representation learning and conversational systems in combination with computer vision.