AARGH! End-to-end Retrieval-Generation for Task-Oriented Dialog

Tomáš Nekvinda and **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{nekvinda,odusek}@ufal.mff.cuni.cz

Abstract

We introduce AARGH, an end-to-end task-oriented dialog system combining retrieval and generative approaches in a single model, aiming at improving dialog management and lexical diversity of outputs. The model features a new response selection method based on an action-aware training objective and a simplified single-encoder retrieval architecture which allow us to build an end-to-end retrieval-enhanced generation model where retrieval and generation share most of the parameters.

On the MultiWOZ dataset, we show that our approach produces more diverse outputs while maintaining or improving state tracking and context-to-response generation performance, compared to state-of-the-art baselines.

1 Introduction

Most research task-oriented dialog models nowadays focus on end-to-end modeling, i.e., the whole dialog system is integrated into a single neural network (Wen et al., 2017; Ham et al., 2020). Although recent end-to-end *generative* approaches based on pre-trained language models produce fluent and natural responses, they suffer from two major problems: (1) hallucinations and lack of grounding (Dziri et al., 2021), which result in faulty dialog management or responses inconsistent with the dialog state or database results, and (2) blandness and low lexical diversity of outputs (Zhang et al., 2020b). On the other hand, *retrieval-based* dialog systems (Chaudhuri et al., 2018) select the most appropriate response candidate from a human-generated training set, thus producing varied outputs. However, their responses might not fit the context and can lead to disfluent conversations, especially when the set of candidates is sparse. This limits their usage to very large datasets which do not support dialog state tracking or database access (Lowe et al., 2015; Al-Rfou et al., 2016).

Several recent works focus on combining the retrieval and generative dialog systems via response selection and subsequent refinement, i.e., *retrieval-augmented generation* (Pandey et al., 2018; Weston et al., 2018; Cai et al., 2019b; Thulke et al., 2021). These models are used for open-domain conversations or to incorporate external knowledge into task-oriented systems and do not consider an explicit dialog state.

Our work follows the retrieve-and-refine approach, but we adapt it for database-aware task-oriented dialog. We aim at improving diversity of produced responses while preserving their appropriateness. In other words, we do not retrieve any new information from an external knowledge base, instead, we retrieve relevant training data responses to support the decoder in producing varied outputs. To the best of our knowledge, we are the first to use retrieval-augmented models in this context. Unlike previous works, we merge the retrieval and generative components into a single neural network and train both tasks jointly, instead of using two separately trained models. Our contributions are summarized as follows:[1]

- We propose a single-encoder retrieval model utilizing dialog action annotation during training, and we show its superior retrieval capabilities in the task-oriented setting compared to two-encoder baseline models (Humeau et al., 2020).

- We propose an end-to-end *task-oriented* generative system with an *integrated* minimalistic retrieval module. We compare it to strong baselines that model response selection and generation separately.

- On the MultiWOZ benchmark (Budzianowski et al., 2018), our approaches outperform previous methods in terms of lexical diversity and achieve competitive or better results in automatic metrics and human evaluation.

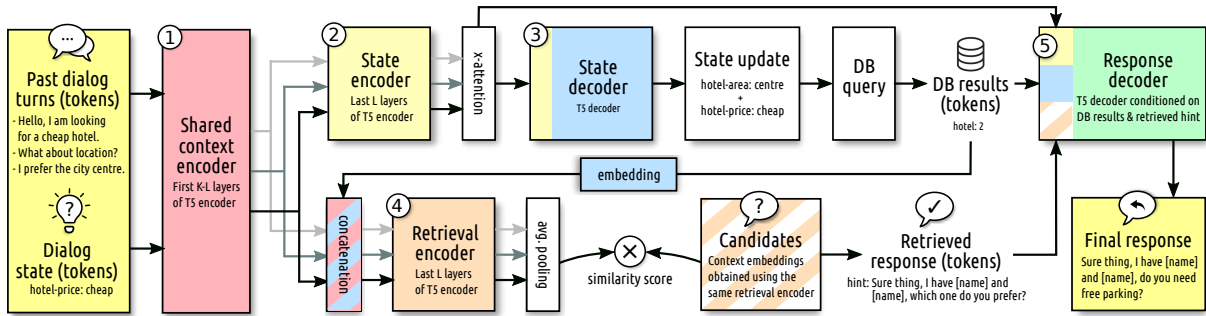[1]Code: https://github.com/Tomiinek/Aargh

Figure 1: Our retrieval-based generative task-oriented system (AARGH, see Section 3.5). Numbers in module boxes mark the order of processing during inference: (1) inputs are pushed through the shared context encoder and (2) state encoder; (3) the state decoder produces the update to the current dialog state. The new state is used to query the database whose outputs are discretized, embedded, and (4) used in the retrieval encoder whose output is reduced to a single vector via average pooling. The context embedding is used to get the best response candidate (hint). Finally, (5) the response decoder, which can attend to the state encoder outputs via cross-attention and is conditioned on the database results and the hint, generates the final system response to be shown to the user.

2 Related Work

Task-Oriented Response Generation Most current works focus on building multi-domain database-grounded systems. The breeding ground for this research is the large-scale conversational dataset MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020; Zang et al., 2020).

Recent models often benefit from *action annotation*. Zhang et al. (2020a) use action-based data augmentation and a three-stage architecture, decoding the dialog state, action, and response. Chen et al. (2019) generate responses without state tracking, exploiting a hierarchical structure of the action annotation. On the other hand, reinforcement learning models (Wang et al., 2021) learn latent actions from data without using annotation.

Recent works focus on *end-to-end* systems based on pre-trained language models. Budzianowski and Vulic (2019) fine-tune GPT-2 (Radford et al., 2019) to model task-oriented dialogs, Hosseini-Asl et al. (2020) enhance this approach with explicitly decoded system actions. Peng et al. (2021b) use auxiliary training objectives and machine teaching for GPT-2 fine-tuning. Lin et al. (2020) introduced the encoder-decoder-based framework MinTL with BART (Devlin et al., 2019a) or T5 (Kale and Rastogi, 2020) backbones (see Section 3.1).

Response Selection can be viewed as scoring response candidates given a dialog context. A popular approach is the dual encoder architecture (Lowe et al., 2015; Henderson et al., 2019b) where the response and context encoders model a joint embedding space. The encoders can take various forms:

Henderson et al. (2019a) compare encoders based on BERT (Devlin et al., 2019b) and custom encoders pre-trained on Reddit; Wu et al. (2020) pre-train encoders specifically for task-oriented conversations. Humeau et al. (2020) introduce poly-encoders, which produce multiple context encodings and add an attention layer to allow rich interaction with the candidate encoding (cf. Section 3.3).

Retrieval-Augmented Generation To benefit from both retrieval and generative models, Weston et al. (2018) proposed an open-domain dialog system utilizing a retrieval network and a decoder to refine retrieved responses. Roller et al. (2021) further developed this approach, using poly-encoders with a large pre-trained decoder. They found that their decoder tends to ignore the retrieved response hints. To combat this, they propose the α-blending method (replacing retrieval output with ground truth, see Section 3.2). Similarly, Gupta et al. (2021) and Cai et al. (2019a,b) focus on retrieval-augmented open-domain dialog, but to prevent the inflow of erroneous information into the generative part of their models, they use semantic frames or reduced forms of retrieved responses instead of raw response texts.

Thulke et al. (2021) aim at knowledge retrieval from external documents for resolution of out-of-domain questions on MultiWOZ (Kim et al., 2020). Shalyminov et al. (2020) present the only work using generation and retrieval in a single model. They finetune GPT-2 (Radford et al., 2019) for response generation in a low-resource task-oriented setup, retrieve alternative responses based on the model's embedding similarity, and choose between gener-

ated and retrieved responses on-the-fly. However, their model is not trained for retrieval, cannot alter retrieved responses, and does not take a dialog state or database into account.

3 Method

We aim at end-to-end modeling of database-aware task-oriented systems, i.e., systems supporting both dialog state tracking and response generation tasks (Young et al., 2013). We combine retrieval and generative models to reduce hallucinations and boost output diversity. We first describe our purely generative baseline (Section 3.1), then explain baseline generation based on retrieved hints (Section 3.2). We then introduce baseline retrieval models (Section 3.3) and our action-aware retrieval (Section 3.4). Finally, we describe AARGH, our single-model retrieval generation hybrid, in Section 3.5. AARGH is shown in Figure 1; other setups are depicted in Appendix A.

3.1 Generative Baseline

Our purely-generative baseline model (**Gen**) follows MinTL (Lin et al., 2020). It is based on an encoder-decoder backbone with a context encoder, shared among two decoders: one for modeling the dialog state updates, the other for producing the final system response. Both decoders attend to the encoded input tokens via an attention mechanism.

The encoder input sequence consists of a concatenation of two parts: (1) past dialog utterances prepended with <|system|> or <|user|> tokens, and (2) the initial dialog state converted to a string, e.g., *hotel [area: center] restaurant [food: African, pricerange: expensive]*. The first decoder is conditioned only on the start-of-sequence token and predicts the dialog state update as a difference between the current state and the initial state. The second decoder is conditioned on the number of database results for each queried domain, e.g. *train: 6* if there are six matching results for a train search, and generates the final response.

During inference, the input is passed through the encoder, then the state update is predicted, merged with the initial dialog state, and this new state is used to query the database (see Section 4 for details). The final system response is predicted based on the context, state, and database results.

3.2 Retrieval-Augmented Response Generation

To combine the retrieval and generative approaches, we follow Weston et al. (2018) and incorporate response *hints*, i.e., the outputs of a retrieval module (Sections 3.3, 3.4), into the generative module in their original form as raw sub-word tokens. Specifically, we add the retrieved response prepended with <|hint|> to the input of Gen's response decoder (Section 3.1), alongside the database results.

Gupta et al. (2021) state that this straightforward token-based retrieve & refine setup might lead to generating incoherent responses due to over-copying of contextually irrelevant tokens. However, using more abstract outputs of the retrieval module, e.g. semantic frames or salient words would go against our goal of reducing blandness and increasing responses lexical diversity. To smoothly control the amount of token copying, we follow Roller et al. (2021) and use the so-called α-blending. During training, we replace the retrieved utterance with the ground-truth final response with probability α. This method also ensures that the decoder learns to attend to the retrieval part of its input successfully.

3.3 Baseline Response Selection

We consider two baseline retrieval model variants:

Dual-encoder (DE) follows the very popular retrieval architecture (Lowe et al., 2015; Humeau et al., 2020) which makes use of context and response encoders. Both produce a single vector in a joint embedding space. During training, the context embedding and the corresponding response embedding are pushed towards each other, while other responses in the training batch are used as negative examples, i.e., cross-entropy loss is used:

$$\mathcal{L}(S) = \frac{1}{N} \sum_j \left(-S_{j,j} + \log \sum_i e^{S_{j,i}} \right)$$

where $S \in \mathbb{R}^{N \times N}$ is the similarity matrix between *normalized* encoded responses $\mathbf{e}^r$ and contexts $\mathbf{e}^c$ in the batch, specifically $S_{i,j} = w \cdot (\mathbf{e}_i^c \cdot \mathbf{e}_j^r)$, where $w > 0$ is a trainable scaling factor.

Inference-time retrieval is as simple as finding the nearest candidate embedding given a context embedding. The context input is similar to Gen's (see Section 3.1): a concatenation of the current updated dialog state, the number of matching database results and past user and system utterances. En-

coders are followed by average pooling and a fully-connected layer for dimensionality reduction.

Poly-encoder (PE)   an extension of DE, aiming at richer interaction between the candidate and the context. The candidate encoder is unchanged. In the context encoder, the average pooling is replaced with two levels of dot-product attention (Vaswani et al., 2017; Humeau et al., 2020). The first level summarizes the encoded context tokens into m vectors. The context tokens act as attention keys and values; queries to this attention are m learned embeddings (query codes). The second attention level provides the candidate-context interaction: it takes the m context summary vectors as keys and values, and the candidate encoder output acts as the query. The parameter m provides trade-off between inference complexity and richness of the context encoding. The loss term remains the same.

3.4   Action-aware Response Selection

We argue that the dual- or poly-encoder models are not practical for the task-oriented settings as their performance depends on the way negative examples are sampled during training (Nugmanova et al., 2019). Choosing appropriate negative examples is difficult in task-oriented datasets as system responses are often very similar to each other (with the conversations being in a narrow domain and following similar patterns). Therefore, we propose a method for candidate selection based on system action annotation, which is usually available in task-oriented datasets. We designed the method to be usable with a single encoder only, but we also include a dual-encoder version for comparison.

Action-aware-encoder (AAE)   Using two separate encoders to encode the response and the context might be impractical due to large model size. Some recent works (e.g., Wu et al. (2020); Roller et al. (2021)) use a single shared encoder instead, and Henderson et al. (2020) discuss parameter sharing between the two encoders. In view of that, we propose a single-encoder action-aware retrieval model. We train it to produce embeddings of dialog contexts which are close to each other if the corresponding responses in the training data have similar action annotation. More precisely, we adapt Wan et al. (2018)'s generalized end-to-end loss, originally developed for batch-wise training of speaker classification from audio: To form training mini-batches, we first sample M random dialog actions,

and for each of those actions, we sample N examples that include the particular action in their system action annotation. We then encode dialog contexts corresponding to the sampled examples into normalized embeddings $\mathbf{e}_{m,n}$, and compute the similarity matrix as follows:

$$
S_{ji,k} = \begin{cases} w \cdot (\mathbf{e}_{j,i} \cdot \mathbf{c}_j^{\{i\}}) & \text{if k = j} \\ w \cdot (\mathbf{e}_{j,i} \cdot \mathbf{c}_k^{\varnothing}) & \text{otherwise} \end{cases}
$$

$$
\mathbf{c}_j^A = \frac{1}{N - |A|} \sum_{i \in [N]-A} \mathbf{e}_{j,i}
$$

where $S \in \mathbb{R}^{N \cdot M \times M}$, $i \in [N]$, $j, k \in [M]$, and $A \subseteq [N]$ is a set of indices. Same as for DE, $w > 0$ is a trainable scaling factor of the similarity matrix. In other words, the similarity matrix describes the similarity between embeddings of each example and centroids, i.e., the means of N embeddings that correspond to the same particular action. For stability reasons and to avoid trivial solutions, we follow Wan et al. (2018) and exclude $\mathbf{e}_{j,i}$ from the centroid calculation when computing $S_{ji,j}$.

We then maximize the similarity between the examples and their corresponding centroids while using other centroids as negative examples:

$$
\mathcal{L}(S) = \frac{1}{N \cdot M} \sum_{j,i} \left(-S_{ji,j} + \log \sum_k e^{S_{ji,k}} \right)
$$

During inference, we rank the responses from the training set according to the cosine similarity of their corresponding contexts and the query context. Again, the contexts consist of the current updated dialog state, the number of matching database results and past utterances.

Action-aware-dual-encoder   (AADE)   This setup follows the DE architecture (see Section 3.3), but it is trained in a similar way as AAE, i.e., we form training mini-batches identically and for each of M distinct actions in the batch, we treat all N examples as positive examples.

3.5   Hybrid End-to-end Model

To further simplify the retrieval-augmented setup, reduce the number of trainable parameters and gain back computational efficiency, we introduce an end-to-end Action-Aware Retrieval-Generative Hybrid model (**AARGH**), which jointly models both response selection and context-to-response generation (see Figure 1). It is a natural extension of the

Gen generative model (Section 3.1), enabled by our new single-encoder action-aware response retrieval (AAE, Section 3.4).

A new retrieval encoder, which produces normalized context embeddings, shares most parameters with the original encoder, which is followed by the two decoders and is partially responsible for state tracking and response generation. To build the retrieval encoder, we fork the last L layers of the original encoder and condition them on the outputs of the shared preceding layers, concatenated with an embedding of the number of current database results. To obtain this embedding, we convert the number of database results into a small set of bins, which are then embedded via a learnt embedding layer of size E.[2] The new retrieval encoder is followed by average pooling and trained using the same objective as AAE (see Section 3.4).

During inference, we pass the input through the partially shared context encoder and decode and update the dialog state. The new state is used to query the database. Database results are embedded and added to the output of the last encoder shared layer to form the input to the retrieval encoder, which produces the context embedding and a retrieved response. Based on state, database results, and retrieved response, the response decoder produces the final (delexicalized) response.

4 Experimental Setup

Models   Our models are based on pre-trained models from HuggingFace (Wolf et al., 2020): We implement Gen and the generative parts in our retrieval-based models using T5-base (Kale and Rastogi, 2020). Retrieval encoders in DE, AADE, PE and AAE are implemented as fine-tuned BERT-base (Devlin et al., 2019a). AARGH is built upon T5-base, same as Gen; we fork the last $L = 2$ out of $K = 12$ encoder layers. The choice of L is a trade-off between model performance and size.[3] The database embedding has size $E = 4$. For simplicity, we do not use specialized backbones pre-trained on dialogs such as ToD-BERT (Wu et al., 2020). PE uses $m = 16$ query codes (see Section 3.3) and single-headed attention mechanisms.

---

[2] This conversion is dataset-specific and not used in other compared models such as Gen. We use the label 0 if there are no results, 1 for 0 matching results, 2, 3, 4 if there are 1, 2 or 3 results, respectively, 5, 6 if there are less than 6 or 11 results, and 7 if there are 11 or more results.

[3] We noticed a performance drop when using $L = 1$, and $L = 3$ did not bring any large gain.
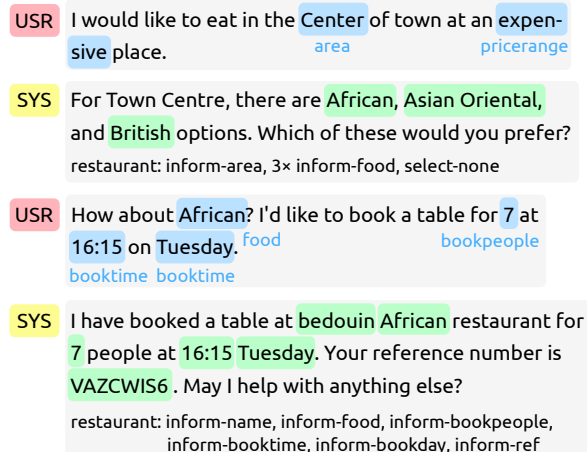

Figure 2: Part of a short conversation from MultiWOZ. It has ● user and ● system turns, and annotated ● slot spans. Both, user and system affect the ● dialog state. Actions are shown below system texts.

Data and database   We experiment on the MultiWOZ 2.2 dataset (Budzianowski et al., 2018; Zang et al., 2020) which is a popular dataset with around 10k task-oriented conversations in 7 different domains such as trains, restaurants, or hotels (see Figure 2). A single conversation can touch multiple domains. The dataset has an associated database, dialog state annotation, dialog action annotation of system turns, and slot value span annotation for easy delexicalization (Wen et al., 2015), thus enabling development of realistic end-to-end dialog systems.[4] To query the database using the belief state, we use the fuzzy matching implementation by Nekvinda and Dušek (2021). To filter out inactive domains from database results during inference, we follow previous work and estimate the currently active domain from dialog state updates.

Input and output format   We use the same formats for all models. Target responses are delexicalized using MultiWOZ 2.2 span annotation, and we limit the context to 5 utterances. MultiWOZ action labels include domain, action, and slot name, e.g., *train-inform-price*. We remove domains from the labels to limit data sparsity.

Training procedure   DE, AADE, PE and AAE are trained in two stages. The retrieval part is trained first and provides response hints to the generative model during the second phase. Modules in AARGH are trained jointly, but we alternate param-

---

[4] Unlike the similar-sized Taskmaster (Byrne et al., 2019) and SGD (Rastogi et al., 2020) datasets, which lack databases and annotation detail.
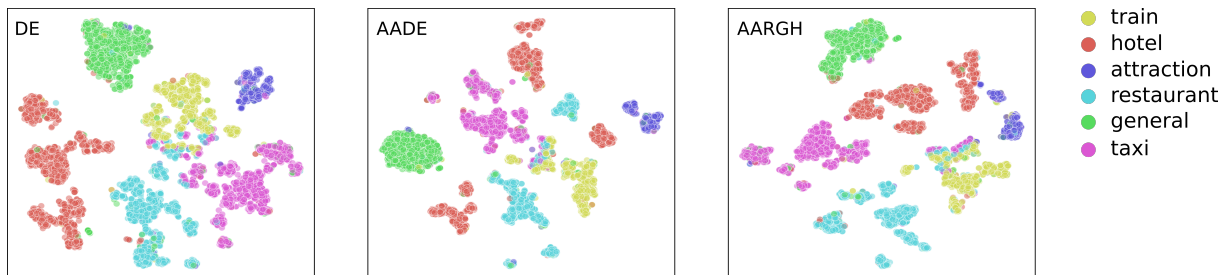
Figure 3: t-SNE projection of test set context embeddings (colored by domains) of retrieval modules of our models. The colors indicate the different MultiWOZ domains that are associated with the corresponding dialog turns.

eter updates for the retrieval encoder and the rest of the network. To do so, we use two separate optimizers. AARGH's hints used in the response decoder during training are refreshed after every epoch. All models are optimized using Adam (Kingma and Ba, 2015) and cosine learning rate decay with warmup. With respect to memory limits of our hardware, we set $N = 6$, $M = 8$ for batch sampling during training of retrieval parts of AAE and AARGH.

α-blending   We experiment with two α-blending values: a conservative one ($\alpha = 0.05$, marked "↓") and a greedy one ($\alpha = 0.4$, marked "↑"), targeting a mostly generation-focused and a mostly retrieval-focused setting.[5]

Decoding   We use greedy decoding for dialog state update generation. For response generation, we report results with greedy decoding in Section 5 and with beam search in Appendix B.

5   Evaluation and Results

We focus on end-to-end modeling, which includes dialog state tracking and response generation. All reported results are on MultiWOZ test set with 1000 dialogs, averaged over 8 different random seeds. We generated responses given ground truth contexts. We follow MinTL and predict the dialog state cumulatively for each conversation turn, which means that state tracking errors may compound. See Appendix C for an example end-to-end conversation without any ground-truth information.

5.1   Response selection

First, we assess the performance of retrieval components of DE, AADE, PE, AAE and AARGH. We cannot use the popular R@k metric (Chaudhuri et al., 2018) as AAE and AARGH use embeddings of dialog contexts (not responses) of candidates as

Setting	BLEU	Action IoU	% full match	% no match	% uniq. hints
Random	2.1	5.1 ± 0.2	1.1	85.0	93.5
DE	8.9	34.7 ± 0.5	11.0	29.7	54.1
AADE	7.9	30.9 ± 1.7	8.9	33.4	24.2
PE	8.8	35.0 ± 0.8	11.4	28.9	44.1
AAE	**12.8**	**37.1** ± 0.2	**14.5**	**28.6**	88.6
AARGH	12.6	36.6 ± 0.2	14.2	29.0	**89.6**

Table 1: Evaluation of retrieval components of our models (Section 3.3, 3.5). See Section 5.1 for details.

the search criterion and would always score 100%. Instead, we use the action annotation and measure the intersection over union (IoU), full-match and no-match rates on sets of actions associated with top-1 retrieved and ground-truth responses. We add BLEU (Papineni et al., 2002; Liu et al., 2016a) between ground-truth and retrieved responses and the proportion of distinct retrieval outputs to assess their lexical similarity to references and diversity.

Table 1 shows that AAE and AARGH significantly outperform other setups on all measures except for the no-match rate,[6] where PE has comparable results. This is expected as they use the additional action annotation during training, unlike DE and PE. AADE performs surprisingly bad. According to the unique hints rate, AAE and AARGH retrieve a much wider range of outputs, which could improve lexical diversity of final responses. The higher BLEU, Action IoU and full match rates suggest that the models retrieve responses more similar to the ground truth.

To further compare the approaches to response selection, we computed the Silhouette coefficient (Rousseeuw, 1987) based on the active domain and action annotation (see Table 3).[7] We omit PE

---

[5]The values were chosen empirically, based on preliminary experiments on development data.

[6]According to a paired t-test with 95% confidence level.

[7]In the case of action-based clustering, we treat each action as a separate cluster; each example can belong to multiple clusters. The clustering measure is calculated for each cluster and averaged over all actions which are weighted by the size

Setting	BLEU	Inform	Success	Unique trigrams	BCE	Hint-BLEU	Hint-copy	Joint acc.
Corpus	-	93.7	90.9	25,212	3.37	-	-	-
SOLOIST (Peng et al., 2021a)	13.6	82.3	72.4	7,923	2.41	-	-	-
PPTOD (Su et al., 2022)	18.2	83.1	72.7	2,538	1.88	-	-	-
MTTOD (Lee, 2021)	19.0	85.9	76.5	4,066	1.93	-	-	-
MinTL (Lin et al., 2020)	19.4	73.7	65.4	2,525	1.81	-	-	-
Gen (equiv. to MinTL)	18.6 ± 0.3	77.0 ± 1.2	66.4 ± 1.0	3,209	1.94	-	-	54.1 ± 0.2
AAE (retrieval only)	12.8 ± 0.1	79.9 ± 0.6	58.3 ± 0.7	22,457	3.34	100.0	100.0 %	-
DE +Gen ↓	**17.6** ± 0.3	80.9 ± 0.5	68.8 ± 0.6	8,190	**2.36**	32.5	15.2 %	54.2 ± 0.1
AADE +Gen ↓	17.3 ± 0.3	81.2 ± 0.9	69.1 ± 1.0	6,613	2.29	26.7	12.8 %	54.3 ± 0.1
PE +Gen ↓	17.4 ± 0.3	79.9 ± 0.9	66.8 ± 1.0	7,736	2.35	31.3	14.5 %	**54.4** ± 0.2
AAE +Gen ↓	17.5 ± 0.6	**82.0** ± 1.0	**70.3** ± 0.8	8,152	2.32	32.0	16.2 %	54.2 ± 0.2
AARGH ↓	17.3 ± 0.3	81.2 ± 0.6	69.5 ± 0.5	**8,200**	2.33	28.4	14.2 %	53.8 ± 0.2
DE +Gen ↑	12.3 ± 0.3	87.8 ± 0.3	69.1 ± 0.5	18,800	3.20	80.4	76.5 %	54.2 ± 0.2
AADE +Gen ↑	**14.6** ± 0.4	81.0 ± 0.8	66.7 ± 0.4	10,723	2.72	51.7	44.8 %	54.2 ± 0.1
PE +Gen ↑	12.9 ± 0.4	86.0 ± 0.8	67.1 ± 0.6	16,632	3.13	74.0	69.1%	**54.4** ± 0.1
AAE +Gen ↑	11.9 ± 0.2	**90.5** ± 0.3	**71.3** ± 0.3	19,436	**3.23**	91.1	89.3 %	54.3 ± 0.2
AARGH ↑	12.1 ± 0.2	89.6 ± 0.2	70.7 ± 0.5	**19,813**	3.21	87.6	85.0 %	53.6 ± 0.2

(The left margin labels $\alpha=0.05$ for the first block of five models and $\alpha=0.4$ for the second block of five models.)

Table 2: Response generation and state tracking evaluation on MultiWOZ using automatic metrics, including the bi-gram conditional entropy (BCE) and number of unique trigrams. We compare previous work, the baseline and retrieval-based generative models. See Section 5.2 for details about the metrics; Section 3, 4 for model descriptions.

Silhouette coefficient	DE	AADE	AAE	AARGH
per Domain	0.098	**0.179**	0.151	0.159
per Action	0.147	0.316	0.312	**0.320**

Table 3: Evaluation of domain and action separation (Section 5.1). We show averages over 8 random seeds.

because its context embeddings depend on queries, i.e., the candidate embeddings (other models output the same context regardless of candidates). DE has the worst results; other systems perform similarly, but AARGH is the best on action separation while AADE has the best scores for domains.

We see that AADE's context encoder is successful in clustering, but it lags behind in terms of correct action selection. Unlike AARGH and AAE, AADE retrieves candidates based on response embeddings. We hypothesize that lower response variability (compared to context variability) leads the model to prefer responses seen more frequently during training. AARGH and AAE are not affected by this as they use purely context-based retrieval.

Figure 3 provides a visualisation of the domain clusters projected using t-SNE (van der Maaten and Hinton, 2008). It supports the findings of our evaluation based on the Silhouette coefficient: We see that visualisations of AARGH and AADE embedding spaces look similarly whereas DE's clusters appear more noisy.

---

of the corresponding clusters.

5.2 Response generation

We evaluate the response generation abilities of our models using automatic metrics and human assessment of delexicalized texts (see Table 4 for examples).

Evaluation with automatic metrics We use the corpus-based evaluator by Nekvinda and Dušek (2021) to measure commonly used metrics on MultiWOZ (Inform & Success rates, BLEU) as well as lexical diversity measures, namely the number of distinct trigrams in the outputs and bigram conditional entropy (Li et al., 2016; Novikova et al., 2019). State tracking joint accuracy is calculated with scripts adapted from TRADE (Wu et al., 2019). To better understand the effect of using retrieved hints and to quantify the amount of copying, we calculate BLEU between retrieved hints and final generated responses (Hint-BLEU) and the proportion of generated responses exactly matching the corresponding retrieved hints (Hint-copy).

We include comparisons with recent strong end-to-end models on MultiWOZ: SOLOIST (Peng et al., 2021a), MTTOD (Lee, 2021), PPTOD (Su et al., 2022), and MinTL (Lin et al., 2020), which has the same architecture as Gen. To show the importance of the generative parts of our models, we also include AAE without the refining decoder.

Table 2 shows scores obtained with greedy decoding (see Appendix B for beam search results). All models have similar state tracking performance.

System	Response
Context	User: I need a taxi to take me to Yippee Noodle Bar by 9:45.
Corpus	Absolutely, where will the taxi be picking you up from today?
Gen	I can help with that. Where are you departing from?
DE +Gen ↓	Where will you be departing from?
AARGH ↓	Okay, I can help with that. Where are you coming from?
AARGH ↑	Of course, I just need to know your *destination* so I can book that for you.
Context	User: I am looking for an expensive hotel with free parking.
	System: We have 5 places that offer all of that. Any preference on star rating or area?
	User: No, that's not important to me. But, I would prefer to stay in a guesthouse with free parking.
Corpus	Sorry, there are no [type] [price] places with free parking. Perhaps you'd prefer a different type or price range?
Gen	I'm sorry, but there are no [type] that meet your criteria. Would you like to try a different price range?
DE +Gen ↓	I'm sorry, but there are no [type] that meet your criteria. Would you like to try a different price range or area?
AARGH ↓	I'm sorry, but I don't have any [type] that meet your criteria. Would you like to try a [type] or a [type]?
AARGH ↑	I'm sorry but I don't have any [type] matching that criteria. Could we look at [type] instead?
Context	User: Thank you so much. You have been very helpful.
Corpus	Great! I hope you enjoy your stay!
Gen	You're welcome. Have a great day!
DE +Gen ↓	You're welcome. Have a great day!
AARGH ↓	Is there anything else I can help you with today?
AARGH ↑	You are welcome. I can also help arrange transportation or suggest attractions, if you wish.

Table 4: Side-by-side system output comparison (see Sections 3 and 4 for model details). The responses are continuations of conversations SNG0016, SNG1048, and MUL2138 from MultiWOZ.

AARGH has slightly lower numbers, which is not surprising as it shares a substantial part of the encoder with its retrieval component. As expected, we notice a huge difference in Hint-BLEU and Hint-copy of versions with different α-blending probabilities (↓ vs. ↑).[8] The performance boost over Gen and retrieval-only AAE is, for ↓ variants, mainly in terms of Success. In ↑, more frequent hint copying reduces BLEU and improves lexical diversity; we also see higher Inform. AAE +Gen and AARGH (both ↓ and ↑) perform better than corresponding DE +Gen or PE +Gen on Inform and Success rates.[9] Differences between AAE +Gen and AARGH are not statistically significant and their Success scores are better than MinTL, competitive with PPTOD and SOLOIST but lower than MTTOD. In terms of lexical diversity, all models are better than most generative baselines.[10]

Human evaluation We arranged an in-house human evaluation on the delexicalized outputs of Gen (i.e., MinTL's architecture), DE +Gen ↓, AARGH ↓ and AARGH ↑. We used side-by-side relative ranking evaluation, which has been repeatedly found to increase consistency compared to rating isolated examples (Callison-Burch et al., 2007; Belz and

	Gen	DE +Gen ↓	AARGH ↓	AARGH ↑
Mean Ranking	2.03	1.99	**1.91**	2.3
Ranked #1	36.1%	35.5%	**40.8%**	37.9%
Ranked #2	34.7%	36.7%	33.5%	18.8%
Ranked #3	18.8%	20.5%	19.7%	18.8%
Ranked #4	10.4%	7.2%	**6.1%**	24.6%

Table 5: Human evaluation results – mean ranks (1-4) established from 50 evaluated conversations.

Kow, 2010; Kiritchenko and Mohammad, 2017). Participants were given full dialog context and current database results, and we asked them to rank responses of the compared models from the best-fitting to the worst, where multiple responses could be ranked the same (see Appendix D for details). We collected rankings for 346 turns of 50 conversations from 5 linguists with experience in natural language generation. All of them were given a different set of dialogs and they were instructed to focus on consistency with the context and database results, naturalness, and attractiveness of the responses. See Table 5 for results.

Although AARGH ↑ scored the best on automatic metrics, it has worse mean ranks than other models, which all have similar mean ranks.[11] This confirms previous findings of low correla-

---

[8]Hint-copy of 15% roughly means one turn per dialog.

[9]According to a paired t-test with 95% confidence level.

[10]The ↓ variants are similar to SOLOIST, which, however, reaches diversity by employng sampling (Holtzman et al., 2020) instead of greedy decoding.

[11]According to Friedman test with 95% confidence level and Nemenyi post-hoc test; only the difference between AARGH ↑ and other models is statistically significant.

tion between automatic metrics and human assessments (Liu et al., 2016b; Novikova et al., 2017). Upon detailed manual error analysis, we found that AARGH ↑ often copies whole hints including words that do not fit the context, i.e., contradictions to earlier statements or noisy non-delexicalized values from the training set. AARGH ↓ performs slightly better than the baselines and is more often ranked best and least often ranked worst.

6 Conclusion

We present AARGH, an end-to-end task-oriented dialog system, combining retrieval and generative approaches. It uses an embedded single-encoder retrieval component which extends a purely generative model without the need for a large number of new parameters. AARGH features an action-aware response selection training objective. Our experiments on the MultiWOZ dataset show that AARGH outperforms baselines in terms of automatic metrics and human evaluation and it is competitive with state-of-the-art models such as SOLOIST or MT-TOD. We showed that our proposed action-aware retrieval training objective supports retrieval of a larger variety of unique and relevant responses in the task-oriented setting and makes efficient use of the available system action annotation. Further, using the retrieval module improves dialog management in terms of the Success rate. A limitation of our approach is the need for careful hyper-parameter setting, coupled with the risk of overuse of retrieved responses that match the dialogue state but are not appropriate for the context.

In future work, we would like to confirm our results on more datasets and explore more complex ways of usage of the retrieved responses to encourage the model to copy interesting language structures while ignoring inappropriate tokens or relics of faulty delexicalization.

Acknowledgements

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational Contextual Cues: The Case of Personalization and History for Response Ranking. *arXiv:1606.00372*.

Anja Belz and Eric Kow. 2010. Comparing Rating Scales and Preference Judgements in Language Evaluation. In *INLG 2010 - Proceedings of the Sixth International Natural Language Generation Conference*, Trim, Co. Meath, Ireland.

Pawel Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019*, pages 15–22, Hong Kong.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 4515–4524, Hong Kong, China.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019a. Skeleton-to-Response: Dialogue Generation Guided by Retrieval Memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 1219–1228, Minneapolis, MN, USA.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019b. Retrieval-guided Dialogue Response Generation via a Matching-to-Generation Framework. 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 2019*, pages 1866–1875, Hong Kong, China.

Chris Callison-Burch, Cameron S. Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007.

(meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007*, pages 136–158, Prague, Czech Republic.

Debanjan Chaudhuri, Agustinus Kristiadi, Jens Lehmann, and Asja Fischer. 2018. Improving Response Selection in Multi-Turn Dialogue Systems by Incorporating Domain Knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018*, pages 497–507, Brussels, Belgium.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-, Volume 1: Long Papers*, pages 3696–3709.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 2197–2214, Virtual Event / Punta Cana, Dominican Republic.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset With State Corrections and State Tracking Baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France.

Prakhar Gupta, Jeffrey P. Bigham, Yulia Tsvetkov, and Amy Pavel. 2021. Controlling Dialogue Generation with Semantic Exemplars. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 3018–3029, Online.

DongHoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 583–592, Online.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019a. A Repository of Conversational Datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and Accurate Conversational Representations From Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online.

Matthew Henderson, Ivan Vulic, Daniela Gerz, Iñigo Casanueva, Pawel Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrksic, and Pei-Hao Su. 2019b. Training Neural Response Selection for Task-Oriented Dialogue Systems. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-, Volume 1: Long Papers*, pages 5392–5404.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-Text Pre-Training for Data-to-Text Tasks. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, pages 97–102, Dublin, Ireland.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tür. 2020. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020*, pages 278–289, Online.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015Proceedings*, San Diego, CA, USA.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017 4, Volume 2: Short Papers*, pages 465–470, Vancouver, Canada.

Yohan Lee. 2021. Improving End-to-End Task-Oriented Dialog System with A Simple Auxiliary Task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1296–1303, Virtual Event / Punta Cana, Dominican Republic.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego California, USA.

Andrea Lin, Zhaojiang andMadotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016a. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 2122–2132, Austin, Texas, USA.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016b. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 2122–2132, Austin, Texas, USA.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using T-Sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, Flavours of Success: The Case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online.

Jekaterina Novikova, Aparna Balagopalan, Ksenia Shkaruta, and Frank Rudzicz. 2019. Lexical Features Are More Vulnerable, Syntactic Features Have More Predictive Power. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019*, pages 431–443, Hong Kong, China.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 2241–2252, Copenhagen, Denmark.

Aigul Nugmanova, Andrei Smirnov, Galina Lavrentyeva, and Irina Chernykh. 2019. Strategy of the Negative Sampling for Training Retrieval-Based Dialogue Systems. In *IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019*, pages 844–848, Kyoto, Japan.

Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. 2018. Exemplar Encoder-Decoder for Neural Conversation Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 1329–1338, Melbourne, Australia.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002*, pages 311–318, Philadelphia, PA.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021a. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Trans. Assoc. Comput. Linguistics*, 9:907–824.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021b. RADDLE: An Evaluation Benchmark and Analysis Platform for Robust Task-oriented Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 4418–4429, Virtual Event.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. Technical report, Open AI.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 8689–8696, New York, NY, USA.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021*, pages 300–325, Online.

Peter J. Rousseeuw. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. 2020. Hybrid Generative-Retrieval Transformers for Dialogue Domain Adaptation. In *DSTC8@AAAI*, New York, NY, USA.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 4661–4676, Dublin, Ireland.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient Retrieval Augmented Generation From Unstructured Knowledge for Task-Oriented Dialog. In *DSTC9@AAAI*, Online.

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*, volume 30.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez-Moreno. 2018. Generalized End-to-End Loss for Speaker Verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, pages 4879–4883, Calgary, AB, Canada.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2021. Modelling Hierarchical Structure between Dialogue Policy and Natural Language Generator with Option Framework for Task-oriented Dialogue System. In *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1711–1721, Lisbon, Portugal.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J. Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Volume 1: Long Papers*, pages 438–449, Valencia, Spain.

Jason Weston, Emily Dinan, and Alexander H. Miller. 2018. Retrieve and Refine: Improved Sequence Generation Models For Dialogue. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018*, pages 87–92, Brussels, Belgium.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos*, pages 38–45, Online.

Chien-Sheng Wu, Steven C. H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 917–929, Online.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-, Volume 1: Long Papers*, pages 808–819.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proc. IEEE*, 101(5):1160–1179.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A Dialogue Dataset With Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020a. Task-Oriented Dialog Systems That Consider Multiple Appropriate Responses under the Same Context. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 9604–9611, New York, NY, USA.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020*, pages 270–278, Online.

A   Model Architectures

Figure 4 shows architectures of the baseline (Gen), dual-encoder-based model (DE), and single-encoder action-aware model (AAE). See Figure 1 for details about AARGH and Section 3 for description of the models.

B   Beam Search Results

See Table 6 for the results of beam search-based response generation evaluation, and compare the results with greedy decoding evaluation (see Section 5.1 and Table 2). For all models, we used beams of size 8 during the decoding

In the case of conservative α-blending, beam search decoding results in higher lexical diversity for all retrieval-augmented systems. However, the gains with respect to Inform and Success rates are mostly very small or not present at all in the case of AADE and AARGH. All BLEU scores are slightly lower which corresponds with the higher output diversity. We notice that the numbers for the baseline without a retrieval component have an opposite trend. Beam search decoding causes lower lexical diversity and higher BLEU. We attribute this to the fact that beam search decoding prefers safer responses with a higher overall probability.

When using higher α-blending, the differences become small even in the case of lexical diversity. We hypothesize that all the retrieval-based models are not substantially influenced by the particular response decoding strategy because they strongly rely on the retrieved hints and their copying.

C   End-to-end Conversation

Figure 5 shows a multi-domain (restaurant and taxi) end-to-end conversation between our end-to-end retrieval-based model AARGH (See Section 3.5).

D   Human Evaluation Interface

We used the graphical user interface depicted in Figure 6 for human evaluation. A full dialog context, i.e., all past utterances corresponding to the particular turn, and the number of database results were shown to participants. We asked participants to rank provided responses from the best to the worst. They evaluated only two conversations in a single run and we sampled the conversations from the test set so that all participants receive roughly the same number of turns to assess. Evaluated responses were shown side-by-side; each of them had a dedicated discrete scale from 1 to 4 where 1 was labeled as the best and 4 as the worst. More responses could receive the same ranking. Participants could move forward and backward in the conversations and they could switch to another conversation anytime.
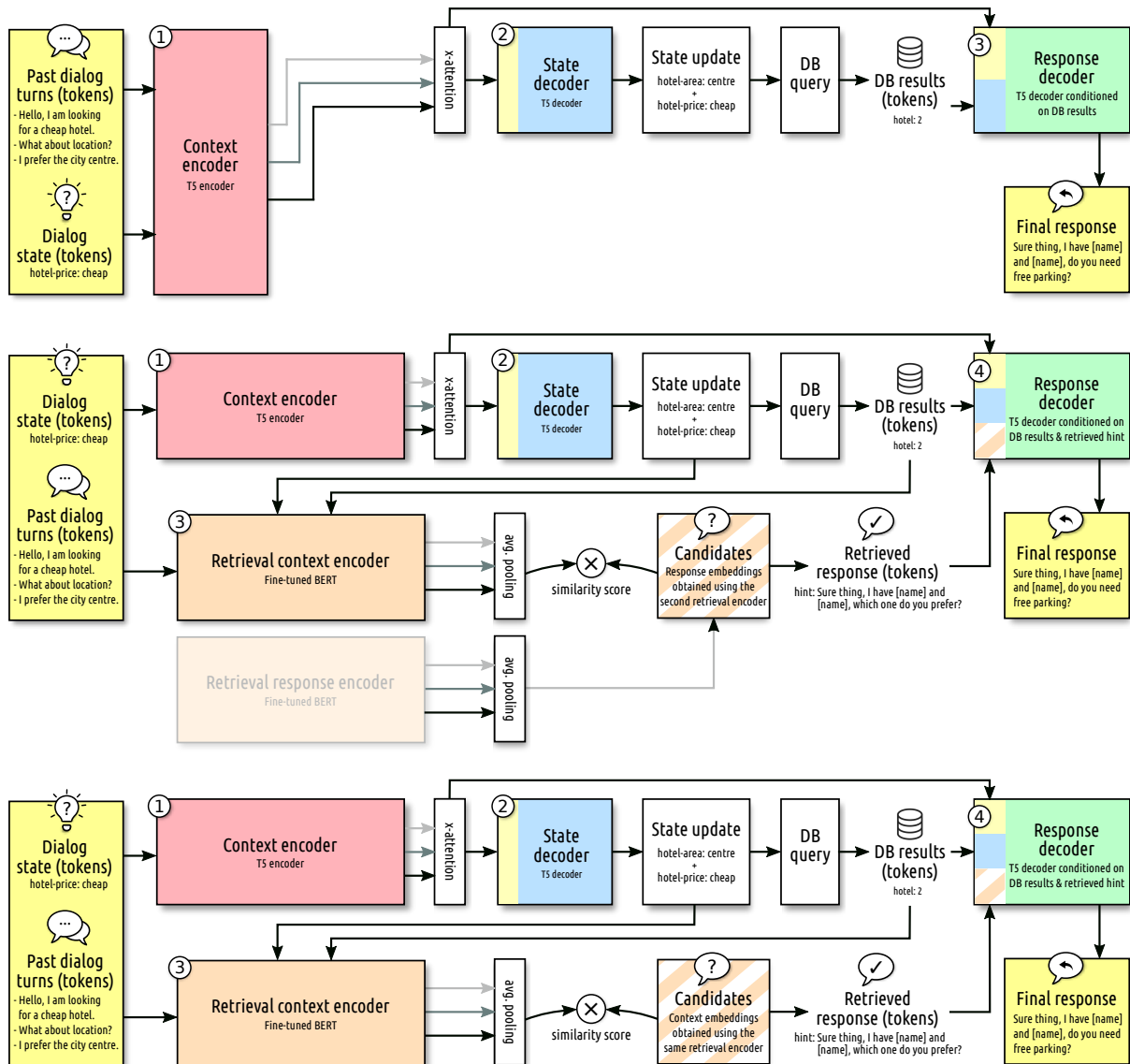
295

Figure 4: Architecture of the baseline (Gen, *top*), dual-encoder-based model (DE, *middle*) and single-encoder action-aware model (AAE, *bottom*). Numbers in module boxes mark the order of processing during inference.

	Setting	BLEU	Inform	Success	Num. trigrams	Bi-gram entropy	Hint-BLEU	Hint-copy
	Gen	19.1 ± 0.3	73.1 ± 1.8	63.0 ± 1.7	2683	1.81	-	-
$\alpha = 0.05$	DE	16.1 ± 0.3	81.1 ± 0.5	68.3 ± 0.8	10098	**2.49**	41.9	25.2 %
	AADE	16.0 ± 0.4	78.0 ± 1.1	65.9 ± 1.0	7378	2.33	32.6	19.2 %
	PE	15.9 ± 0.4	80.6 ± 0.9	66.9 ± 1.0	9470	2.48	40.7	24.4 %
	AAE	**16.4 ± 0.4**	**82.5 ± 0.8**	**69.8 ± 0.6**	**10457**	2.46	44.2	29.0 %
	AARGH	16.2 ± 0.3	79.5 ± 0.5	68.0 ± 0.3	9072	2.36	36.2	22.2 %
$\alpha = 0.4$	DE	12.3 ± 0.3	87.7 ± 0.3	68.7 ± 0.6	19103	**3.28**	83.1	79.6 %
	AADE	13.7 ± 0.4	77.8 ± 0.9	63.2 ± 0.4	10997	**2.76**	55.8	50.2 %
	PE	**12.7 ± 0.3**	85.9 ± 0.9	66.3 ± 0.7	17178	3.19	78.3	74.2 %
	AAE	12.0 ± 0.2	**90.3 ± 0.5**	**71.2 ± 0.4**	19448	3.21	92.6	90.1 %
	AARGH	12.2 ± 0.2	89.3 ± 0.3	70.0 ± 0.6	**19763**	3.22	88.6	86.0 %

Table 6: Beam search-based response generation on MultiWOZ using automatic metrics. For each model setup, We use beams of size 8 during response decoding and report results averaged over 8 random seeds. We compare the baseline (Gen) and retrieval-based generative models (See Section 3 and 4). See Section 5.2 for details about the metrics. Cf. Table 2 showing results obtained using greedy decoding.

Usr Hello, I need to book a restaurant. It must be cheap and in the city centre. Can you help me?

Certainly, there are [choice] [pricerange] restaurants to choose from in the [area]. Are you in the mood for any particular type of food?   I can help you with that. There are several cheap restaurants in the centre. Are you looking for any particular type of food?   Aargh

Usr I would like to taste French cuisine. Do you have some restaurants with frogs and similar food?

There are no [food] restaurants located in [area]. Are you interested in another type of cuisine?   I'm sorry, there are no french restaurants in the centre. Would you like to try a different type of cuisine?   Aargh

Usr Ok, no problem. I actually like noodles too. What about Chinese restaurants?

There are [choice] such restaurants. I reccomend [name]. Would you like a reservation?   There are 3 such restaurants in the centre. Charlie Chan and Rice House. Would you like to book one of these?   Aargh

Usr Book Charlie Chan for three persons on Friday 8pm please.

Booking was successful for [bookday]. The table will be reserved for 15 minutes. Reference number is: [ref]. Can I help with anything else today?   Booking was successful. The table will be reserved for 15 minutes. Reference number is: 0000010. Can I help with anything else today?   Aargh

Usr I also need a taxi from El Shaddai to the restaurant.

Of course! What time would you like the taxi to leave or arrive?   What time would you like the taxi to leave or arrive by?   Aargh

Usr I need to be in the restaurant at 19:45.

I booked a [type] to pick you up. The contact number for the taxi is [phone].   I booked a yellow lexus to pick you up. The contact number for the taxi is 13582494566.   Aargh

Usr Ok thanks, that's all for today.

Thank you for allowing me to help you. Have a great day!   You're welcome. Have a great day!   Aargh

Figure 5: End-to-end conversation between ● the user and ● our retrieval-based AARGH model with conservative α-blending (see Section 3). For the system turns, we show delexicalized hints proposed by the retrieval module (left boxes in italics) and the corresponding lexicalized final responses (right boxes). We highlighted ● the parts of hints present in the final texts and ● the parts of final responses newly-introduced by the model during refining.

Figure 6: Our graphical user interface used for human evaluation.