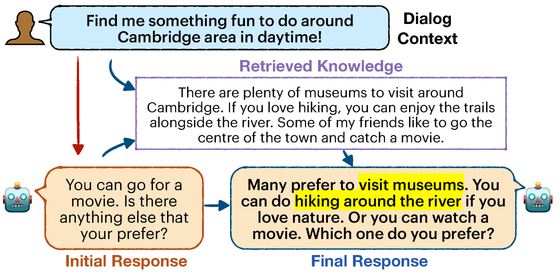
Achieving Conversational Goals with Unsupervised Post-hoc Knowledge Injection

Bodhisattwa Prasad Majumder[♣] **Harsh Jhamtani**[◇]
Taylor Berg-Kirkpatrick[♣] **Julian McAuley**[♣]
[♣]Department of Computer Science and Engineering, UC San Diego
`{bmajumde, tberg, jmcauley}@eng.ucsd.edu`
[◇]School of Computer Science, Carnegie Mellon University
`jharsh@cs.cmu.edu`

Abstract

A limitation of current neural dialog models is that they tend to suffer from a lack of specificity and informativeness in generated responses, primarily due to dependence on training data that covers a limited variety of scenarios and conveys limited knowledge. One way to alleviate this issue is to extract relevant knowledge from external sources at decoding time and incorporate it into the dialog response. In this paper, we propose a post-hoc knowledge-injection technique where we first retrieve a diverse set of relevant knowledge snippets conditioned on both the dialog history and an initial response from an existing dialog model. We construct multiple candidate responses, individually injecting each retrieved snippet into the initial response using a gradient-based decoding method, and then select the final response with an unsupervised ranking step. Our experiments in goal-oriented and knowledge-grounded dialog settings demonstrate that human annotators judge the outputs from the proposed method to be more engaging and informative compared to responses from prior dialog systems. We further show that knowledge-augmentation promotes success in achieving conversational goals in both experimental settings.

1 Introduction



Find me something fun to do around Cambridge area in daytime!

There are plenty of museums to visit around Cambridge. If you love hiking, you can enjoy the trails alongside the river. Some of my friends like to go the centre of the town and catch a movie.

You can go for a movie. Is there anything else that your prefer?

Many prefer to visit museums. You can do hiking around the river if you love nature. Or you can watch a movie. Which one do you prefer?

Figure 1: Augmenting initial response from an existing dialog model with relevant external knowledge leads to more *engaging* and *informative* responses improving the success in achieving the conversational goal (here, finding a fun activity).

Generic responses which lack specificity have been a major issue in existing dialog models (Hosseini-Asl et al., 2020; Dinan et al., 2019a). The issue in part stems from bottlenecks in dialog models due to a limited scope of scenarios and access to limited knowledge available during training. On the other hand, encoding all possible world knowledge at training time is not feasible, and even undesirable in cases where knowledge sources are dynamically varying (Ghazvininejad et al., 2018; Majumder et al., 2020b; Zhao et al., 2020; Bruyn et al., 2020; Kim et al., 2020; Prabhumoye et al., 2021). One possible approach is to incorporate relevant knowledge at decoding-time. For example, in Figure 1, the user is seeking options for a fun activity around Cambridge. While the initial dialog response suggests watching a movie as an option, it does not provide any information behind that choice.

We propose and evaluate an approach for unsupervised knowledge injection into a dialog model's response at decoding time[1]—not addressed in any previous work. We first sample a response from the model (trained on dialog data) conditioned on the dialog context. Next, we utilize the dialog context and the sampled response to query external knowledge sources. Finally, the retrieved knowledge is used to construct a more informative and engaging response (Figure 1). A major advantage of such post-hoc knowledge injection is its flexibility in adding newer knowledge sources especially where the success of achieving conversational goals relies upon the availability of relevant knowledge. Post-hoc injection also promotes efficiency in NLP applications (Schwartz et al., 2020; Strubell et al., 2019): it mitigates the need to retrain dialog models to accommodate dynamically evolving knowledge.

We experiment with two types of knowledge sources: language models, which we treat as parametric knowledge bases (Petroni et al., 2019;

---

[1] Code: `https://github.com/majumderb/poki`

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3140 - 3153
May 22-27, 2022 ©2022 Association for Computational Linguistics

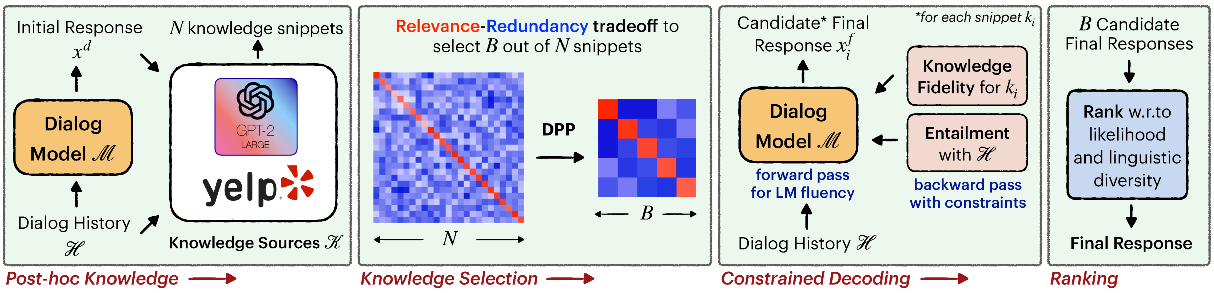

Initial Response x^d
N knowledge snippets
Dialog Model $\mathscr{M}$
GPT-2 LARGE
yelp
Dialog History $\mathscr{H}$
Knowledge Sources $\mathscr{K}$
Relevance-Redundancy tradeoff to select B out of N snippets
DPP
N
B
Candidate* Final Response x_i^f
*for each snippet k_i
Knowledge Fidelity for k_i
Entailment with $\mathscr{H}$
Dialog Model $\mathscr{M}$
forward pass for LM fluency
backward pass with constraints
Dialog History $\mathscr{H}$
B Candidate Final Responses
Rank w.r.to likelihood and linguistic diversity
Final Response
Post-hoc Knowledge
Knowledge Selection
Constrained Decoding
Ranking

Figure 2: Pipeline of POKI: It first retrieves **post-hoc knowledge** from external sources based on dialog history and an initial response from a dialog model. Then the most relevant and diverse knowledge snippets are **selected** from the retrieved set. Each selected snippet is individually combined with the initial response through **constrained decoding** to generate a candidate final response. At last, the final response is selected via an unsupervised **ranking** step. Note that POKI requires no additional training.

Brown et al., 2020); and user review datasets such as Yelp reviews (Hajas et al., 2014) as non-parametric knowledge sources (§ 2). Since it is possible to gather a large amount of related knowledge given a query, we select a relevant and diverse (estimated via information-theoretic measures) subset of knowledge snippets using an unsupervised method (§ 3.1). Then, a gradient-based inference approach is used to construct an updated response that incorporates the selected knowledge (§ 3.2). Note that our framework does not require retraining the existing dialog model—it only relies upon updating the model's output hidden states at decoding time for unsupervised knowledge injection.

We experiment with two scenarios: goal-oriented and knowledge-grounded dialog where the training data covers only a fraction of the needed knowledge. Automatic evaluation reveals that our method is capable of generating highly diverse responses in both settings. In some cases, the generated response shows high overlap with the original target response showing that our unsupervised method bridges the knowledge gap between available knowledge and human-written responses present in the existing dialog corpus. An extensive human evaluation confirms that generated responses are indeed engaging, interesting, and human-like without any loss in fluency.

To pinpoint the usefulness of knowledge injection in the above settings, we design a real-time study (§5.3) where users interact with our system to reach a conversational goal (e.g. planning a holiday or knowing more about the solar system). We find that external knowledge enables users to achieve their goals more efficiently. Additionally, we observe that the our approach of sub-selecting relevant but diverse knowledge leads to responses that promote success in achieving conversational goals.

2 Post-hoc Knowledge for Dialog

Our goal is to construct a dialog response by injecting knowledge (from external textual sources) at decoding time, without having to retrain the models. Consider a dialog model $\mathcal{M}$ from which we can sample a dialog response x^d given a dialog history $\mathcal{H}$. We shall refer to the response x^d sampled from such a model without any decoding time knowledge injection as the *initial* response.

However, as motivated earlier, samples from such a dialog model often lack detail. To improve such responses, we retrieve and incorporate relevant external knowledge k into the initial response. To achieve our goal, we construct a *query* using both dialog history $\mathcal{H}$ and the initial response x^d, and gather a relevant knowledge candidate k from a knowledge source $\mathcal{K}$. The retrieved snippet can provide useful information to the end-user to achieve the conversational goal (see §5.3). We explore both parametric (e.g querying a language model) and non-parametric (e.g. deterministic retrieval using word-overlap) ways to obtain post-hoc knowledge.

2.1 Parametric knowledge sources

Pretrained language models (PTLM) are typically trained with a vast amount of text that spans a diverse range of domains. Petroni et al. (2019); Brown et al. (2020) showed that such PTLMs can be used as a source of knowledge when queried with suitable textual prompts (e.g. *Seattle is famous for ___*). To use PTLMs in our use-case, we construct useful prompts from dialog history and the *initial* response. We assemble simple prompts inspired from various knowledge-seeking situations in dialog (Shwartz et al., 2020) such as `[KP]` *is famous for ___*, *Here is what I know about* `[KP]`: ___, where `[KP]` is a key-phrase[2] extracted from dialog context. We use `gpt2-large` as the PTLM. For example, a query "Here is what I know about fun things around Cambridge:" results in "*There are plenty of museums to visit around Cambridge. If you love hiking, you can enjoy the trails alongside the river...*" as shown in Figure 1. A complete list of prompts is provided in Appendix B. We finally rank each knowledge snippet k using the likelihood obtained from the PTLM for a concatenated input of k and dialog history and choose the most likely.

2.2 Non-parametric knowledge sources

External knowledge in the form of a text corpus can be used as a non-parametric knowledge source available at decoding time. Compared to parametric knowledge sources, such sources do not generate text as knowledge snippets, but offer the advantage of high quality and reliability of human written text. We consider the dialog history and the *initial* response as a query to retrieve relevant knowledge instances from the corpus. Next, we identify the top relevant instances in the given corpus with respect to the constructed query using cosine similarity on TF-IDF based representations (Robertson et al., 1995).

3 Unsupervised Knowledge Injection in Generated Dialog

Effectively utilizing the retrieved knowledge snippets to construct an enriched dialog response encompasses two major challenges. Firstly, it is not practical to use potentially hundreds of knowledge snippets obtained from the retrieval step for a single response generation. Thus, we need to find a relevant but diverse subset of the snippets. Secondly, the dialog model $\mathcal{M}$ is trained to condition only on the dialog context, and not on the external knowledge. Hence, to leverage the knowledge snippets, we need a decoding strategy to rewrite the initial response x^d such that the resulting final response x^f should closely follow the knowledge snippet to be injected without a loss in the fluency and consistency. Thus, our method requires no additional training and only assumes a language model trained on dialog context (i.e. $\mathcal{M}$). We refer to our proposed framework (Figure 2) as **POKI** (**Po**st-hoc **K**nowledge **I**njection in Generated Dialog).

3.1 Relevance-Redundancy Tradeoff for Knowledge Selection

At each turn, we obtain N knowledge snippets from both the parametric and non-parametric sources. We wish to select a subset of B (out of N) relevant but diverse knowledge snippets.

We define relevance score of a snippet k_i with respect to the dialog history H using pointwise mutual information (PMI) as follows:

$$\mathbb{REL}_i = \text{PMI}(k_i, \mathcal{H}) = \log\left(\frac{p(\mathcal{H}|k_i)}{p(\mathcal{H})}\right),$$

Thus, a high PMI score would imply a larger semantic similarity between the snippet k_i and H. To account for redundancy between the snippet pair k_i, k_j we again use the PMI score as follows:

$$\mathbb{RED}_{ij,j>i} = \text{PMI}(k_i, k_j) = \log\left(\frac{p(k_j|k_i)}{p(k_j)}\right).$$

The redundancy score is symmetric i.e. $\mathbb{RED}_{ij} = \mathbb{RED}_{ji}$ as PMI is a symmetric measure.

We estimate probabilities (both conditional and marginal) $p(.)$ in the above equations using GPT2 language model, following past work (Padmakumar and He, 2021). The PMI measure is often considered better than other n-gram-based overlap metrics to measure the degree of association between two sentences (Kedzie et al., 2018; Padmakumar and He, 2021). Semantically similar phrases occur in both sentences that can easily be ignored by overlap based metrics.

Selection via Determinantal Point Processes. To select B knowledge snippets out of N with a relevance-redundancy trade-off, we use a subset selection process named Determinantal Point Process (DPP) (Kulesza and Taskar, 2011). DPP employs a non-uniform selection that assigns low probability to subsets (here, of knowledge snippets) that are less diverse by modeling the repulsive correlation between independently occurring datapoints (see Figure 2).

We build an $N \times N$ kernel matrix $\mathcal{D}$, which is real, symmetric and positive semi-definite. The diagonal entries $\mathcal{D}_{ii}$ are populated by the squared relevance score of the i-th knowledge $\mathbb{REL}_i$ and the off-diagonal entries $\mathcal{D}_{ij}$ are $\beta \times$ squared redundancy scores $\mathbb{RED}_{ij}$. We adjust β in such a way that $\mathcal{D}$ always remains positive semi-definite (more details in (Wilhelm et al., 2018)). To select a subset of B, a DPP assigns a probability of sampling such a subset proportional to the determinant

---

[2] It possible that a lack of key-phrases results in no knowledge. Key-phrase extraction details are in Appendix B.

of the submatrix $\mathcal{D}_B$ of $\mathcal{D}$, constructed using the indices of the subsetted items. The DPP probability is geometrically related to the volume of the parallelepiped spanned by the selected knowledge snippets. Diverse knowledge snippets tend to be orthogonal in their space hence span larger volume (Kulesza and Taskar, 2012).

Choosing B-size submatrix from N-size $\mathcal{D}$ is a combinatorial problem and can become prohibitively costly when N is very high. Hence, we use a greedy method (Wilhelm et al., 2018) where we initialize the selection with the most relevant k_i and subsequently select the next k_j that maximizes the determinant of the resultant submatrix.

3.2 Gradient-based Constrained Decoding for Knowledge Injection

Upon selecting B knowledge snippets, we want to individually inject each knowledge snippet into x^d to construct a candidate final response x^f at inference time.

Previous works have addressed the problem of unsupervised modification of already-generated text using gradient-based decoding (Dathathri et al., 2020; Qin et al., 2020) that employs an iterative procedure consisting of a forward and a backward pass. The forward pass on the generative model (here, $\mathcal{M}$) encourages fluency of the generated text while the backward pass performs gradient ascent on certain desired constraints. Note that due to the discrete nature of x_d, it is not possible to directly update it via back-propagation. Therefore, we maintain the sequence of hidden representations of each output token as z from the dialog model. Each output token $x^d_{(t)}$ is realized via $p(x^d_{(t)}) \sim \mathrm{softmax}(Wz_{(t)}/\tau)$, where τ is the temperature hyperparameter, W is the output embedding matrix (shared with the input), and $Wz_{(t)} \in \mathcal{R}^V$ (V is the size of the vocabulary).

Constraints. Following Majumder et al. (2021a), we define a **knowledge fidelity** objective that encourages x^f to be minimally different from the knowledge snippet k. We achieve this by minimizing the cross entropy loss (CE) between knowledge tokens $k_{(1)}, \ldots, k_{(T)}$ as labels and $Wz_{(1)}, \ldots, Wz_{(T)}$ as the logits.

We further notice that injected knowledge can influence the generation in such a way that it contradicts with responses uttered during previous turns. Hence, we also want x^f to be entailed with the dialog history $\mathcal{H}$. We build an **entailment** classifier $\theta(z, \mathcal{H})$ that predicts the probability of x^f (ideally, the hidden representation z of x^f) entailing $\mathcal{H}$. The classifier $\theta(z, \mathcal{H})$ is a bag-of-words classification layer with hidden states z from $\mathcal{M}$ and fine-tuned using the DNLI dataset (Welleck et al., 2019) to predict whether the current response is entailed with previous responses or not.

Decoding. In the subsequent forward and backward passes, the hidden representation z is gradually perturbed via gradient ascent on the respective objectives. During backward pass, the objective with constraints is

$$\mathcal{L}(\mathcal{H}, k; z) = \alpha \log \theta(z, \mathcal{H}) - \lambda\, \mathrm{CE}(k, Wz)$$

with hyperparameters α and λ. We use back-propagation to update z with the gradient $\nabla_z \mathcal{L}(\mathcal{H}, k; z)$ while the parameters of $\mathcal{M}$ remain fixed. The updated latent representations of z after the backward pass are denoted as z^{bw}.

A forward pass with $\mathcal{M}$ is required to regularize the hidden states z toward the original dialog model objective to obtain z^{fw}. Corresponding to the tth token, the hidden states for the $t+1^{\text{th}}$ time step are computed via a weighted addition of backward and forward hidden states, i.e., $z_{(t+1)} = \gamma \times z^{bw}_{(t)} + (1-\gamma) \times z^{fw}_{(t)}$ where $\gamma \in (0, 1)$ is a hyperparameter.

During generation, we start by sampling the initial response x^d with greedy decoding from $\mathcal{M}$. The hidden states z (of x^d) are iteratively updated by alternate backward and forward passes. The final response is sampled as $x^f \sim \mathrm{softmax}(Wz/\tau)$. The number of iterations ($= 5$) and the γ ($= 0.45$) were chosen by maximizing the Z-normalized sum of dialog model perplexity and linguistic diversity (% of distinct bigrams) in a greedy hyperparameter search. More details are in Appendix B.

3.3 Unsupervised Ranking of Candidate Final Responses

Several previous works often over-generate and use an additional ranking step in order to select the final candidate in unsupervised text generation (Qin et al., 2020; Shwartz et al., 2020; Paranjape and Manning, 2021). Similarly, here we want to rank the generated candidate final responses according to the diversity of the generated text as well as the conditional likelihood of generation given the dialog history. For diversity, we measure the percentage of distinct bigrams present in the response. For conditional likelihood, we use

System	Acc	BLEU	BRTSc	D-2	ENTR
KCopy	70.1	4.1	62.3	3.16	2.41
SimpleTOD (2020)	70.1	**15.0**	**79.2**	0.56	0.90
SimpleTOD+ (2021)	69.8	12.1	68.1	0.81	1.11
Arranger (2021)	70.2	12.3	68.5	0.93	1.15
Rewriter (2021)	70.2	12.1	69.4	1.03	1.45
POKI	**71.1**	13.7	74.5	**3.78**	**2.67**
w/o Entailment	69.9	10.9	67.8	**3.67**	**2.56**
w/o Kw Fidelity	70.0	12.3	71.2	0.95	1.19
Gold	100	100	100	0.78	0.86

Table 1: Automatic metrics on the test set of MultiWoZ. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$).

System	BLEU	BRTSc	D-2	ENTR
KCopy	13.4	74.3	**3.64**	3.12
KGuide (2017)	16.7	71.5	2.54	2.12
KGround (2019)	18.3	72.5	2.87	2.35
BART (2020a)	**19.8**	73.4	2.97	2.55
RAG (2020b)	**19.9**	73.1	1.03	1.45
POKI	**19.4**	**76.8**	**3.65**	**3.44**
w/o Entailment	18.1	74.2	3.17	**3.39**
w/o Kw Fidelity	18.8	73.3	2.75	2.54
Gold	100	100	2.98	2.59

Table 2: Automatic metrics on the test set of Wizard-of-Wikipedia. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$).

the pre-trained GPT2 model to obtain the log probability when the dialog history, followed by the generated response, passed as a concatenated input. Since these two scores can have varied scale, we perform Z-normalization on the individual scores and add them to obtain a single score for ranking. The highest ranked candidate response is finally rendered to the user.

4 Experimental Setup

4.1 Scenarios and Datasets

We experiment with two dialog scenarios: goal-oriented and knowledge grounded. Both setups are knowledge intensive but the training data in such setups often contains only a fraction of the needed knowledge. For the goal-oriented setting, we use the Multi-domain Wizard-of-Oz (Budzianowski et al., 2018) dataset. For knowledge grounded dialog, we use the Wizard-of-Wikipedia (Dinan et al., 2019b) dataset. More details are in Appendix A.

Multi-domain Wizard-of-Oz (MultiWOZ) is a multi-domain dialog dataset (we use v2.0 (Hosseini-Asl et al., 2020)) consisting of goal-oriented human-human conversations. The dataset spans seven domains (restaurant, train, attraction, hotel, taxi, hospital, police) and contains 10,438 dialogs with 13.68 average turns. Since, we do not need any training data, we only use an evaluation set (of 7K utterances).

Wizard-of-Wikipedia (WoW) is a knowledge grounded dialog dataset which involves retrieving relevant knowledge from Wikipedia, reading and conditioning on it, and finally generating dialog responses (Dinan et al., 2019b). The dataset contains 201K utterances from 22K dialogues spanning 1300 diverse topics, from which we use only the test set. The associated Wikipedia knowledge base has 5.4M articles and 93M sentences.

4.2 Baselines and Ablations

Baselines for MultiWOZ. For MultiWOZ, we consider several baselines following (Sun et al., 2021) for knowledge injection. First, we use the current state-of-the-art model, SimpleTOD, for goal-oriented dialog (Hosseini-Asl et al., 2020). Sun et al. (2021) extends SimpleTOD by adding chitchat candidates to dialog histories during training. They also have other variants that either *concatenate* output from SimpleTOD and candidate chitchats (Arranger) or *rewrite* by combining both output and chitchat snippets (Rewriter). We also have a trivial baseline (KCopy) which appends the retrieved knowledge snippet k from POKI with the initial response x_d.

Baselines for WoW. For WoW, we use two current-best knowledge-grounded models, KGround (Wolf et al., 2019) and BART (Lewis et al., 2020a) that concatenate the associated knowledge snippets (present in WoW) and the dialog history as inputs to generate the response with supervision. KGuide (Zhao et al., 2017) and RAG (Lewis et al., 2020b) have an additional knowledge selection step modeled by a latent variable before response generation similar to knowledge grounded models. We also use the KCopy baseline, as described for MultiWOZ.

Variants of POKI. To investigate the impact of various decoding constraints in POKI, we consider the following two variants of POKI—w/o Entailment and w/o Knowledge (Kw) Fidelity (§ 3.2). In POKI, we use SimpleTOD as the base dialog model in goal-oriented scenarios and use BART (which is a state-of-the-art model for WoW) as the base dialog model in the knowledge-grounded scenario. For all variants of POKI, we use gradient-based inference for decoding the final response.

	POKI vs	**SimpleTOD**			**Rewriter**			**w/o Entailment**			**w/o Kw Fidelity**			**Gold**		
	Criteria	win	loss	κ	win	loss	κ	win	loss	κ	win	loss	κ	win	loss	κ
MultiWOZ	Coherent	**93.2**	4.4	0.76	**85.6**	10.2	0.75	**98.7**	0.8	0.72	**77.8**	17.8	0.78	26.2	**34.4**	0.69
	Engaging	**94.3**	4.5	0.78	**89.7**	7.9	0.79	**98.7**	0.6	0.80	**71.5**	20.5	0.80	42.4	37.4	0.78
	Interesting	**92.7**	5.4	0.72	**91.2**	8.3	0.73	**88.6**	8.9	0.68	**98.7**	0.8	0.75	49.7	45.6	0.67
	Humanlike	**85.4**	10.7	0.68	**87.4**	7.3	0.65	**61.9**	30.5	0.71	**81.7**	14.0	0.74	29.7	**37.8**	0.66
		RAG			**BART**			**w/o Entailment**			**w/o Kw Fidelity**			**Gold**		
WoW	Coherent	**95.4**	4.5	0.78	**88.5**	9.6	0.72	**94.3**	3.4	0.68	**83.6**	10.7	0.65	23.8	**25.3**	0.73
	Engaging	**89.3**	7.7	0.72	**87.8**	8.3	0.71	**97.7**	0.8	0.70	**71.5**	25.4	0.69	25.4	26.7	0.73
	Interesting	**96.3**	3.5	0.74	**83.3**	9.9	0.75	**79.8**	17.2	0.70	**93.5**	4.5	0.71	35.9	**37.8**	0.76
	Humanlike	**91.4**	7.1	0.68	**92.4**	6.5	0.66	**84.5**	10.5	0.67	**81.8**	13.5	0.71	42.3	41.9	0.68

Table 3: Pairwise comparison (% win/loss cases, tie not reported) between responses from POKI and from other baselines as well as ground truth. Difference between bold and non-bold numbers is statistically significant ($p < 0.001$). κ denotes Cohen's Kappa (Cohen, 1960) between a pair of annotators. Complete details of the human evaluation are in Appendix C.

5 Results and Discussion

5.1 Automatic Evaluation

Our primary goal is to generate responses enriched with relevant external knowledge. Arguably, a system which can effectively leverage additional knowledge at decoding time should generate more diverse responses. We measure percentage of distinct bigrams as Distinct-(D-2) (Li et al., 2016) and geometric mean of entropy values of empirical frequency distributions of n-grams ($n = 1, 2, 3$) as Entropy (ENTR) (Jhamtani et al., 2018) for diversity. Additionally, we report overlap between generated responses and corresponding ground truth as per BLEU and BERTScore (BRTSc). For MultiWOZ, we also report the final goal accuracy (Acc) following (Hosseini-Asl et al., 2020).

MultiWOZ. Table 1 shows POKI outperforms all the baselines in terms of diversity of generated responses. More importantly, we see POKI promotes accuracy of reaching the final dialog state i.e. the goal. For ablated versions of POKI, we find the entailment constraint has little effect on diversity while dropping the knowledge adherence constraint negatively influences accuracy and diversity. All variants of SimpleTOD and all versions of POKI show departure from the results obtained by SimpleTOD on BLEU and BERTScore since all of these versions add external knowledge that were not explicitly present in the data. However, we observe that the departure is not significant and POKI achieves a much closer BERTScore to SimpleTOD compared to baselines.

WoW. Despite all systems for WoW use knowledge explicitly in the knowledge-grounded dialog generation task, Table 2 shows POKI generates the most diverse responses. Similar to MultiWOZ, the knowledge adherence constraint still remains a significant factor for increasing diversity, one of the main goals of knowledge injection. For WoW, we instead see POKI outperform even BART (previous SOTA) in terms of BERTScore when injected with external knowledge indicating the need of the external knowledge for modeling WoW dialogs.

5.2 Human Evaluation

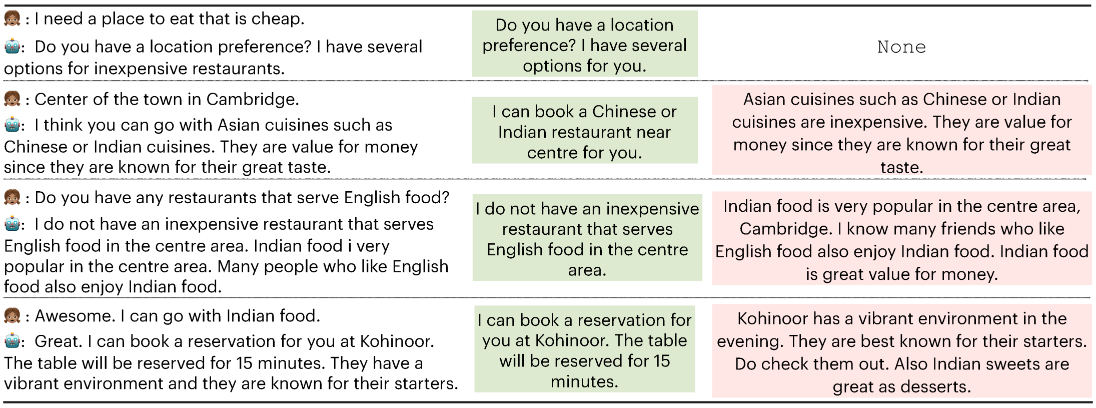
We conduct a comparative human evaluation with 300 samples to evaluate the quality of generated dialog responses following ACUTE-Eval (Li et al., 2019). We show a generated response from POKI to an annotator with its associated dialog history to annotate if knowledge injection makes the final response more *engaging*, *interesting* and *humanlike* compared to a baseline response. As sanity check, we also investigate if the response remain *coherent* after knowledge injection. Each sample is evaluated by two annotators[3].

MultiWOZ. Table 3 records the pairwise comparison showing POKI consistently outperforms baselines on all criteria. Responses from POKI are more engaging and interesting compared to SimpleTOD and Rewriter, demonstrating that gradient-based decoding is effective for knowledge injection. In POKI, entailment constraint mostly influences coherence whereas knowledge fidelity constraint is important for engagingness and interestingness.

WoW. Table 3 shows POKI outperforms baselines that use grounding knowledge during training in all criteria showing that external knowledge can be useful even in the knowledge-grounded setting to make the conversation engaging and interesting. It also indicates the limitation of the training signal or lack of access to sufficient knowledge and

---

[3] More details of the setup are in Appendix C.



👩 : I need a place to eat that is cheap. 🤖: Do you have a location preference? I have several options for inexpensive restaurants.	Do you have a location preference? I have several options for you.	None
👩 : Center of the town in Cambridge. 🤖: I think you can go with Asian cuisines such as Chinese or Indian cuisines. They are value for money since they are known for their great taste.	I can book a Chinese or Indian restaurant near centre for you.	Asian cuisines such as Chinese or Indian cuisines are inexpensive. They are value for money since they are known for their great taste.
👩 : Do you have any restaurants that serve English food? 🤖: I do not have an inexpensive restaurant that serves English food in the centre area. Indian food i very popular in the centre area. Many people who like English food also enjoy Indian food.	I do not have an inexpensive restaurant that serves English food in the centre area.	Indian food is very popular in the centre area, Cambridge. I know many friends who like English food also enjoy Indian food. Indian food is great value for money.
👩 : Awesome. I can go with Indian food. 🤖: Great. I can book a reservation for you at Kohinoor. The table will be reserved for 15 minutes. They have a vibrant environment and they are known for their starters.	I can book a reservation for you at Kohinoor. The table will be reserved for 15 minutes.	Kohinoor has a vibrant environment in the evening. They are best known for their starters. Do check them out. Also Indian sweets are great as desserts.

Figure 3: POKI converses with a user who is looking for some restaurant options (left column). In each turn, a knowledge snippet (right column) is injected into an initial response (middle column). More examples are in Appendix D.

room for improvement in terms of how knowledge is utilized. A large gap in win percentages in favor of POKI for evaluating how 'humanlike' is a response when compared to state-of-the-art methods suggests knowledge injection leads to more natural conversation. Here too, both decoding constraints show similar trends to MultiWOZ.

Qualitative Analysis. Figure 3 shows a conversation by POKI with a user who seeks to find restaurant options around Cambridge. We observe that in most of the turns the injected knowledge appeared as an additional justification over the initial responses making the dialog engaging and effective to reach the user's goal (also noted by human judges in §5.3). For example, in turn 3, we observe that adding the extra information about Indian cuisine helped user to reach a conclusion when their original choice of English cuisine was absent.

Effect of Response Length. Qualitatively, as seen in Figure 3, responses generated by POKI are longer than those from the initial response due to the post-hoc knowledge injection. In the human evaluation sample, we found that 37% of responses from POKI are similar or smaller in length compared to responses from the best baseline. We investigate if response length acted as a confounding factor during human evaluation. Among all the cases where POKI was *lost* over a baseline, 45% (± 2% when bootstrapped with 1000 subsets of size 50) of responses from POKI were longer than those from the comparing baseline. Among *win* cases for POKI, we observe 49% (± 3% when bootstrapped with 1000 subsets of size 50) POKI responses were longer than those from the comparing method. This indicates that human users did not only choose longer responses as better.

5.3 User Study for Effectiveness of Knowledge Injection

Relevant knowledge injection has the benefit of adding more justification to terse dialog outputs and hence influencing the task outcome positively. Mirroring observations from (Ghandeharioun et al., 2019), a real-time full conversation evaluation is needed to investigate if POKI could achieve the conversational goal any better than baselines.

We recruited 60 users for this study[4]. One half of the users interacted with POKI, while the other half interacted with the best baseline model that does not augment dialog responses with external knowledge. We construct a *speculative goal* for each user to accomplish via the conversation. We allow users to end the conversation any time they would like and ask them whether the system helped them to reach their conversation goal along with additional comments to justify their annotation. Users who interacted with a knowledge-augmented system also asked if the system provided any knowledge that user has not explicitly asked for but indeed the extra information helped them to reach the conversational goal (Majumder et al., 2021b). Finally, we also ask if they would like to engage with the system they interacted with in future.

For goal-oriented dialog, we construct speculative goals (e.g. looking for entertainment options) manually from the ground truth for 300 dialog samples. Since we are not using the underlying databases, we made sure speculative goals do not require specific information (e.g. booking availability, flight information, etc.). For knowledge-grounded dialog, we provide the intended topic of

---

[4] More details of the participants and the study setup are in Appendix C.

MultiWOZ	**# turns ↓**	**Goal**	**Know**	**Would use**
Rewriter	8 ± 2	69%	35%	56%
POKI	$\mathbf{4 \pm 3}$	**86%**	**84%**	**76%**
WoW	**# turns ↑**	**Goal**	**Know**	**Would use**
BART	10 ± 2	56%	70%	48%
POKI	$\mathbf{16 \pm 3}$	**76%**	**89%**	**71%**

Table 4: Real-time user study with average # of turns for successful goal completion, % of time the goal was achieved, % of success cases users were helped by an *additional* knowledge (Know) that was not explicitly asked to reach their goal, and if users would like to use the system in future.

	Relevant		**Factual**		**BRTSc for WoW**	
Source	**Random**	**DPP**	**Random**	**DPP**	**Random**	**DPP**
Parametric	82%	**89%**	65%	**83%**	74.2	**81.3**
Non-parametric	81%	**83%**	97%	**98%**	65.2	**76.8**

Table 5: Evaluation for the quality of the knowledge snippets for random and DPP-based selection.

System	**MultiWOZ**	**WoW**
Supervised	17.6 ± 5.2 ms	23.6 ± 4.6 ms
PPCM (2020)	30.9 ± 7.5 ms	32.6 ± 4.2 ms
POKI	34.2 ± 8.4 ms	35.7 ± 5.7 ms
POKI, only decoding	31.6 ± 2.7 ms	32.3 ± 3.4 ms

Table 6: Mean and std. error of clock-time taken per token

discussion (e.g. science fiction) present in the data; the speculative goal here is to know more about, or to have an engaging conversation about the topic.

Results. First of all, we find that POKI is unanimously preferred by users compared to the baseline during the user study. More importantly, we see that when the user successfully accomplished their goal, 84% of those times they found the additional knowledge helpful in the goal-oriented setting (MultiWOZ) as compared to a baseline (Rewriter) that did not use any external knowledge. Most importantly, POKI takes significantly fewer turns for users to accomplish the goal as compared to Rewriter implicitly indicating injected knowledge (we observe high correlation, 0.67) contributes toward more efficient conversations.

For the knowledge-grounded setting (WoW), both BART and POKI have access to external knowledge sources. However, 89% (compared to 70%) of success scenarios were directly influenced by the additional post-hoc knowledge. For knowledge-grounded dialog, a longer conversation is indicative of engagingness on a particular topic (Gopalakrishnan et al., 2019), hence users preferred to converse with POKI for more turns as compared to a BART baseline. We quote a comment from a user who found a conversation about the Korean culture with POKI was particularly engaging—"*Before this conversation, I had less knowledge about Korean movies and art-forms. This gave me a new perspective and a handful of popular opinions to look at it.*".

5.4 Discussion

Performance of Knowledge Selection. The knowledge selection step in POKI acts an information bottleneck where the quality of the generated response directly depends on the quality of the selected knowledge[5]. We perform a human evaluation on 200 snippets to measure the relevance and the factual correctness in two scenarios: when we randomly select a retrieved snippet or select via DPP. In Table 5, we see that the parametric knowledge source (`gpt2-large`) generates more relevant knowledge snippets than a non-parametric one. We attribute this to 1) a large and diverse dataset (webtext) used during pretraining of `gpt2` as compared to yelp reviews (restricted domains) we used for retrieval, and 2) the limited recall of relevant knowledge when using word-overlap based retrieval. However, large language models are still prone to generate non-factual knowledge. We observe that DPP-based selection in POKI is able to sub-select more factual knowledge which then positively influences the final response quality. For WoW, we also compare the selected snippets with the gold knowledge available in the dataset that in turn show high fidelity in terms of BERTScore.

Time Complexity. Madotto et al. (2020) shows that iterative gradient-based decoding could be slower than generating response using single forward pass from an existing model. When we benchmark POKI in an Nvidia 2080Ti GPU, in Table 6, we see that knowledge generation (or retrieval) could be a computational bottleneck for POKI. However the greedy selection and the constrained decoding step do not add significant computational load. Furthermore, POKI's performance is comparable with PPCM (Madotto et al., 2020)—a more efficient version of gradient-based decoding. The efficiency of the knowledge retrieval step can be improved with better indexing (Johnson et al., 2021) which we leave as a future work.

---

[5] A statistical analysis on number of knowledge snippets retrieved/generated and selected is provided in Appendix B.

6 Related Work

Knowledge grounded dialog datasets such as Wizard-of-Wikipedia (Dinan et al., 2019a) and Topical chat (Gopalakrishnan et al., 2019) typically consist of dialog responses paired with relevant knowledge available as collected annotations. Hence, models trained on such datasets are restricted to the knowledge sources they were exposed to at training time. Past work (Sun et al., 2021; Majumder et al., 2020a; Su et al., 2020; Komeili et al., 2021; Adolphs et al., 2021; Ghazvininejad et al., 2018; Tuan et al., 2020; Lewis et al., 2020c; Guu et al., 2020) has looked into injecting extra knowledge sources at training time in a bid to add knowledge not available originally as paired to dialog responses. However, such approaches require re-training the model if some new knowledge source were to be used. Moreover, while previous work focuses on just improving specificity of dialog response using external knowledge, we also study the effect of additional knowledge in achieving conversational goals.

Improving the diversity of dialog responses by using diversity-promoting sampling has been explored in past work (Fan et al., 2018; Holtzman et al., 2020). We use a gradient-based decoding method, building on past work in this direction (Dathathri et al., 2020; Qin et al., 2020; Madotto et al., 2020; Majumder et al., 2021a). However, we propose new objectives to inject post-hoc knowledge obtained based on already generated dialog—an unsupervised knowledge injection method that has not been explored so far.

7 Conclusion

We propose a framework for unsupervised knowledge injection into dialog responses. We show that knowledge can be obtained post-hoc from *any* knowledge sources that can improve users' ability to reach their conversational goal more effectively. In future, our idea can be generalized to setups where external knowledge can justify model's predictions such as conversational recommendation.

Acknowledgements

We thank anonymous reviewers for providing valuable feedback. BPM is partly supported by a Qualcomm Innovation Fellowship, a Friends of the International Center Fellowship–UC San Diego, NSF Award #1750063, and MeetElise.

References

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021. Reason first, then respond: Modular generation for knowledge-infused dialogue. *CoRR*, abs/2111.05204.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2020. BART for knowledge grounded conversations. In *Converse@KDD*, volume 2666. CEUR-WS.org.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *EMNLP*.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019a. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *ACL*.

Guillaume Gautier, Guillermo Polito, Rémi Bardenet, and Michal Valko. 2019. DPPy: DPP Sampling with Python. *Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS)*.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Àgata Lapedriza, and Rosalind W. Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *NeurIPS*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Peter Hajas, Louis Gutierrez, and Mukkai S. Krishnamoorthy. 2014. Analysis of yelp reviews. *CoRR*, abs/1407.1443.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *NeurIPS*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *ACL 2018*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*.

Chris Kedzie, Kathleen R. McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *EMNLP*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *ICLR*. OpenReview.net.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *CoRR*, abs/2107.07566.

Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In *ICML*. Omnipress.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5(2-3):123–286.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020c. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Findings of EMNLP*.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian J. McAuley, and Harsh Jhamtani. 2021a. Unsupervised enrichment of persona-grounded dialog with background stories. In *ACL*.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian J. McAuley. 2020a. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *EMNLP*.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian J. McAuley. 2020b. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *EMNLP*.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian J. McAuley. 2021b. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. *NAACL*.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *EACL*.

Ashwin Paranjape and Christopher D. Manning. 2021. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *NAACL-HLT*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP*.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W. Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *NAACL-HLT*.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *EMNLP*.

Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1995. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Inf. Process. Manag.*, 31(3):345–360.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM*, 63(12):54–63.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *EMNLP*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *ACL*.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *ACL*.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chats to enhance task-oriented dialogues. *NAACL*.

Yi-Lin Tuan, Wei Wei, and William Yang Wang. 2020. Unsupervised injection of knowledge into dialogue generation via language models. *CoRR*, abs/2004.14614.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *ACL*.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical diversified recommendations on youtube with determinantal point processes. In *CIKM*. ACM.

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

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *EMNLP*.

A Datasets

MultiWOZ. To compare with previous works, we use MultiWoz 2.0 following (Hosseini-Asl et al., 2020). Note that we do not need any training data for our models since we perform post-hoc knowledge injection.

WoW For Wizard-of-Wikipedia, all baselines and the original dialog model for POKI use available paired knowledge present in the training data (not a part of our pipeline). However, POKI additionally uses the external knowledge snippets selected via DPP.

B Implementation Details

We open-source our code at: `https://github.com/majumderb/poki`. We use the publicly available implementation[6] for DPP (Gautier et al., 2019).

We obtain the MultiWOZ 2.0 from the official release [7]. Similarly, we obtain the Wizard-of-Wikipedia from ParlAI repository [8]. We adapted codes from original PPLM (Dathathri et al., 2020) repository[9] and modified them for our own objective function. We obtained the Yelp review dataset from the official website[10]. Yelp dataset contains 8,635,403 reviews. For diversity calculation (in automatic evaluation), we use NLTK[11] to extract n-grams.

Network architecture For MultiWOZ, we use the SimpleTOD[12] as the base model. Whereas for WoW, we use BART[13] as the base model. For the parametric knowledge source, we use `gpt2-large`[14].

Hyperparameters POKI does not require any training since we perform gradient-based decoding at the inference time. For hyperparameters involved in the decoding stage, we maximize the

---

[6] `https://github.com/guilgautier/DPPy`

[7] `https://github.com/budzianowski/multiwoz`

[8] `https://parl.ai/projects/wizard_of_wikipedia/`

[9] `https://github.com/uber-research/PPLM`

[10] `https://www.yelp.com/dataset`

[11] `https://www.nltk.org/_modules/nltk/util.html`

[12] `https://github.com/salesforce/simpletod`

[13] `https://huggingface.co/transformers/model_doc/bart.html`

[14] `https://huggingface.co/transformers/model_doc/gpt2.html`

Z-normalized sum of dialog model perplexity and linguistic diversity (% of distinct bigrams) of the generated response in a greedy fashion to select the best values. For our best method, in objective function $\mathcal{L}$, we use α as 1 and λ as 1. We keep generation length to be 100 to encourage longer generations. We train the entailment classifier using code from PPLM repository[15]. The weight γ for mixing forward and backward passes was set to 0.45. We run 5 backward-forward passes to obtain a candidate final response.

Filtering knowledge candidates from PTLMs Our initial experiments suggests that that knowledge generated from PTLMs can be inappropriate (contains bias or toxic content) and misleading/nonfactual. Sun et al. (2021) collected annotations of dialog responses with labels `positive` (useful, social), `negative` (inappropriate and misleading). We learn a binary classifier to classify a knowledge snippet as positive or negative and use it as a filtering criteria.

Key-phrase extraction Given a sentence from the context, we first extract n-gram (n ∈ 1,2,3,4) key-phrases using YAKE (Yet-Another-Keyword-Extractor) (Campos et al., 2020) and retain only those that contain at least a noun.

Prompts We curated prompts inspired by various knowledge-seeking situations (such as for: more information, opinion, review) (Shwartz et al., 2020) and are listed in Table 7.

`[KP]` is famous for ____
The popular opinion about `[KP]` is ____
Here is what I know about `[KP]`: ____
My friend says that `[KP]` is: ____
Here is some information about `[KP]`: ____
Here are some reviews about `[KP]`: ____
I think `[KP]` is: ____
I read on the internet about `[KP]` and found that ____
Today I learned about `[KP]` that ____

Table 7: Manually curated prompts to query the PTLM

Statistics on generated and selected knowledge snippets For both datasets, we retrieve 100 most relevant knowledge snippets from non-parametric source (here, yelp reviews), and generate 5 candidate knowledge snippets (using nucleus sampling (Holtzman et al., 2020), $p = 0.95$) for each key-phrase extracted from an input instance (dialog history + initial response). After knowledge selection by DPP, on an average (over validation set), 5 snippets were selected for MultiWoz and 8 snippets were selected for WoW.

C Human Evaluation and User Study Setup

Human Evaluation We hired two Anglophone (Lifetime HIT acceptance % > 85) annotators for every test sample. Figure 4 shows a sample question for the pairwise comparison between response generated by POKI and a baseline for informativeness. The exact formulations for all criteria are provided as below:

- **Coherent**: Which version is more consistent with the dialog history?
- **Engaging**: *Which version is more likely to hold your attention and make you want to hear more?*
- **Interesting**: *Which version arouses your curiosity or tells you something new or useful?*
- **Humanlike**: *Which version is more natural and personable?*

All differences in values from human evaluations are significant with $p < 0.05$ from bootstrap tests on 1000 subsets of size 50. A snapshot of our human evaluation interface is shown in Figure 4. The order of two candidate responses (R1 and R2) is made random for each question.

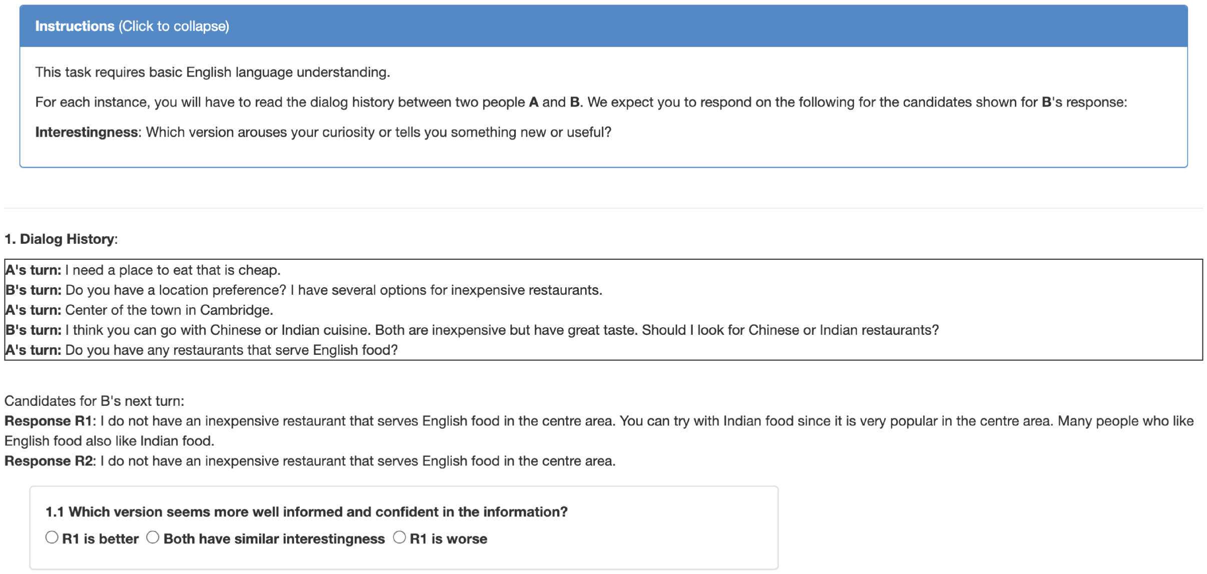
User Study For user study, we similarly recruited 60 Anglophone users who have at least high-school level of education and are comfortable with handling internet-based technologies. Each session (depending on the systems they interacted) lasted on an average 30 minutes (for MultiWOZ) and 60 minutes (for WoW) including on-boarding, performing actual task and answering post-task questions.

D Qualitative Examples

Figure 5 shows a complete dialog in the knowledge-grounded scenario where the user discusses about 'science-fiction'. Figure 6 shows more utterance level examples for both goal-oriented and knowledge-grounded scenarios.

---

[15] `https://github.com/uber-research/PPLM/blob/master/run_pplm_discrim_train.py`



Instructions (Click to collapse)

This task requires basic English language understanding.

For each instance, you will have to read the dialog history between two people **A** and **B**. We expect you to respond on the following for the candidates shown for **B**'s response:

Interestingness: Which version arouses your curiosity or tells you something new or useful?

1. Dialog History:

A's turn: I need a place to eat that is cheap.
B's turn: Do you have a location preference? I have several options for inexpensive restaurants.
A's turn: Center of the town in Cambridge.
B's turn: I think you can go with Chinese or Indian cuisine. Both are inexpensive but have great taste. Should I look for Chinese or Indian restaurants?
A's turn: Do you have any restaurants that serve English food?

Candidates for B's next turn:
Response R1: I do not have an inexpensive restaurant that serves English food in the centre area. You can try with Indian food since it is very popular in the centre area. Many people who like English food also like Indian food.
Response R2: I do not have an inexpensive restaurant that serves English food in the centre area.

1.1 Which version seems more well informed and confident in the information?

○ **R1 is better** ○ **Both have similar interestingness** ○ **R1 is worse**

Figure 4: Human evaluation setup for pairwise comparison between POKI and another baseline

Ethical considerations

We do not foresee any immediate ethical concerns for our method as we use several constraints (less divergence from the extracted knowledge, consistency with the dialog context) that allow the generation to be restricted to the context. In general, we expect our dialog system to be more engaging and accessible to the user. Since we use PTLMs as knowledge source, we inherit the general risk of generating biased or toxic language, which should be carefully filtered. In our work, we perform explicit filtering steps to make sure that the knowledge is *appropriate*. Furthermore, our selection step promotes more factually correct knowledge to be selected. However, the generations may incorporate biases that are already present in the dialog datasets due to crowd-sourced data collection. Finally, our generations are limited only to the English language. Hence we suggest that a system like ours should likely not be used as a 'black box,' but would best be used in a setting where its outputs can be 'audited'. **Carbon footprint:** Our system uses post-hoc knowledge injection which refrains from retraining newer dialog models to accommodate dynamically evolving external knowledge. This promotes green NLP applications (Schwartz et al., 2020; Strubell et al., 2019) reducing carbon footprints that stem from training (or even finetuning) large language models.

👧🏾 : I think science fiction is an amazing genre for anything. Future science, technology, time travel, FTL travel, they're all such interesting concepts. 🤖: I read a lot of science fiction books. My favorite author is Issac Asimov. What about you?	I read a lot of science fiction books. What about you?	Issac Asimov is one of the most popular science fiction writers. His writing style is quite unique.
👧🏾 : Awesome! I like Issac Asimov too. Sci-fi storytellers should on philosophical issues. 🤖: Yes! I like how science fiction discusses dystopias, alien invasion and time travel. They address the current problems and how we can save ourselves from them.	I like how they address problems that can destroy our future.	Science fiction is known for discussing dystopias, alien invasions and time travel. They talk about how we destroy our current world and how we can save ourselves.
👧🏾 : Great that you think in the same way. I gtg. 🤖: Thanks! You are welcome.	Thanks! You are welcome.	None

Figure 5: POKI converses with a user who is discussing about science fiction, in a knowledge-grounded dialog scenario (left column). In each turn, a initial response (middle column) is augmented with a knowledge snippet (right column) using constrained gradient-based decoding. Human judges unanimously noted this conversation as more engaging as compared to the initial responses.

👧🏾 : I'd like to dine for an expensive cuisine. 🤖: I'd recommend the British cuisine. It is popular because of its freshness. There are plenty to eat at your leisure.	I'd recommend British cuisine.	Expensive British cuisine is popular because of its easy accessibility, freshness. There are plenty for you to eat at your leisure.

(a)

👧🏾 : Are there any theatres in the centre of town, Cambridge? 🤖: I would recommend the Mumford theatre. It has good parking and drinking places nearby. The place is an intimate venue.	There are 4 theatres in the centre. I would recommend the Mumford theatre.	Mumford theatre has good parking and drinking places nearby. It has lovely intimate venue and the door prices are very reasonable.

(b)

👧🏾 : Tell me more about Picasso. 🤖: Picasso is one of the finest artists in the modern time. He was raised in Spain before he spent most of his adult life in France.	Picasso is one of the finest artists in the modern time.	Picasso was raised in Spain before going on to spend most of his adult life working as an artist in France.

(c)

Figure 6: Utterance level examples (left column) in (a) and (b) goal oriented scenario; and (c) knowledge-grounded scenario. POKI updates the initial response (middle column) with a knowledge snippet (right column) using constrained gradient-based decoding.