

Generative Knowledge Selection for Knowledge-Grounded Dialogues

Weiwei Sun, Pengjie Ren, Zhaochun Ren*

Shandong University, Qingdao, China

sunweiwei@gmail.com, {renpengjie, zhaochun.ren}@sdu.edu.cn

Abstract

Knowledge selection is the key in knowledge-grounded dialogues (KGD), which aims to select an appropriate knowledge snippet to be used in the utterance based on dialogue history. Previous studies mainly employ the classification approach to classify each candidate snippet as “relevant” or “irrelevant” independently. However, such approaches neglect the interactions between snippets, leading to difficulties in inferring the meaning of snippets. Moreover, they lack modeling of the discourse structure of dialogue-knowledge interactions. We propose a simple yet effective generative approach for knowledge selection, called GENKS. GENKS learns to select snippets by generating their identifiers with a sequence-to-sequence model. GENKS therefore captures intra-knowledge interaction inherently through attention mechanisms. Meanwhile, we devise a *hyperlink* mechanism to model the dialogue-knowledge interactions explicitly. We conduct experiments on three benchmark datasets, and verify GENKS achieves the best results on both knowledge selection and response generation.

1 Introduction

To improve the informativeness in open-domain dialogue agents (Freitas et al., 2020), knowledge-grounded dialogues (KGD) are proposed to leverage external structured (Liu et al., 2019) and unstructured (Dinan et al., 2019) knowledge to dialogue responses. In KGD, it is pivotal to embed factual and conversationally appropriate knowledge in responses. Two classes of approaches are considered to embed knowledge: *end-to-end* and *pipeline*. End-to-end models, such as FiD (Izacard and Grave, 2021), process the document and generate the response in one shot. However, they tend to misuse knowledge (Adolphs et al., 2021). Pipeline models address this problem by explicitly identifying a specific knowledge snippet to

*Corresponding author.

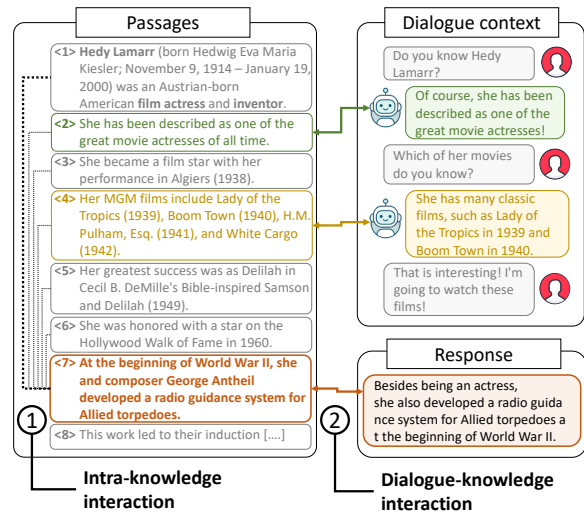


Figure 1: An example of knowledge-grounded dialogues. The dialogue agent selects a knowledge snippet (i.e., <7>) from passages and generates a response based on it. Intra-knowledge interactions and dialogue-knowledge interactions are denoted by ① and ②, respectively.

be used in the response (Adolphs et al., 2021). Typically, pipeline KGD approaches have two sub-steps, i.e., knowledge selection and response generation (Dinan et al., 2019; Kim et al., 2020): The former aims to select knowledge snippets from passages, and the latter generates responses based on them. Knowledge selection plays a vital role in KGD as it directly determines the content of the response (Lian et al., 2019; Meng et al., 2020). In this paper, we focus on selecting knowledge snippets for dialogue to enhance pipeline KGD models.

The *Classification* paradigm dominates knowledge selection studies. In this paradigm, each snippet is independently classified as “relevant” or “irrelevant” (Dinan et al., 2019; Zhao et al., 2020b). However, these approaches ignore *knowledge interactions*, which refer to flows of information within the knowledge or between knowledge and dialogues. As shown in Figure 1, we identify two

types of knowledge interactions in KGD:

Intra-knowledge interaction Intra-knowledge interaction refers to the interactions between snippets. It is worth noting that the meaning of a knowledge snippet is context-dependent and can be ambiguous when taken individually. For example, the <8> snippet in Figure 1 “*This work led to their*” has a referential element *their*, and is difficult to identify its meaning without knowing the remaining context of the sentence. However, with the existence of the remaining context, we can quickly infer that it refers to *Lamarr and George Antheil*. *This problem challenges existing methods when selecting knowledge on new topics*.

Dialogue-Knowledge interaction Previous works also neglect interactions between dialogue and knowledge. There is a discourse structure and smooth transition of involved knowledge in multi-turn dialogue. For example, *Lamarr’s* profession mentioned in the dialogue in Figure 1 is demonstrated in a parallel and multi-perspective manner, while some other cases follow a shallow-to-deep structure in dialogue.

Some recent efforts attempt to fix these problems within the classification paradigm; for example, [Li et al. \(2022\)](#) build a semantic graph for passages to capture intra-knowledge interaction, [Kim et al. \(2020\)](#) propose sequential knowledge selection to model the dialogue-knowledge interaction as latent variables. However, they are complicated, lack deep semantic interactions, and are challenging to model the two types of knowledge interaction simultaneously.

In this work, we propose **GENKS (Generative Knowledge Selection)**, a simple yet effective *generative* model that addresses these challenges. GENKS first assigns an identifier to each snippet, feeds all the snippets into the model simultaneously, and then selects snippets by generating their identifiers with a sequence-to-sequence Transformer model (e.g., BART ([Lewis et al., 2020a](#))). Compared with KGD methods with the *classification* paradigm, GENKS captures interactions between knowledge snippets through the *self-attention* mechanism in Transformer ([Vaswani et al., 2017](#)). Therefore, GENKS can obviate the ambiguity in snippets with the existence of the rest context and improve the understanding of knowledge. Moreover, we propose a *hyperlink* method to

capture the dialogue-knowledge interactions explicitly and effectively. Finally, we propose to joint knowledge selection and response generation within one generative model.

We evaluate our proposed method on three public KGD datasets: Wizard of Wikipedia ([Dinan et al., 2019](#)), Holl-E ([Moghe et al., 2018](#)), and CMU_DoG ([Zhou et al., 2018](#)). The experimental results show that GENKS significantly improves the accuracy of knowledge selection as well as the quality of response generation, by establishing new state-of-the-art on KGD benchmarks. Improvements are particularly significant on unseen topics, outperforming the BART classification model by up to 8.1% absolute. GENKS also achieves the best results as the number of dialogue turns increased, with an average of 10% improvements over the BART classification model in the last three turns. We also compare our model with recent SOTA end-to-end methods ([Shuster et al., 2021](#)), and find our model can generate responses with fewer hallucinations while having better controllability and interpretability. The effectiveness of the proposed method is also validated through human evaluation and ablative experiments.

Our contributions are summarized as follows: (1) We propose GENKS, which is the first attempt at generative knowledge selection in KGD. (2) GENKS captures intra-knowledge and dialogue-knowledge interactions simultaneously. (3) We propose a hyperlink method to enhance the interactions between dialogue and knowledge. (4) Experiments verify that GENKS establishes a new state-of-the-art on KGD¹.

2 Related work

Knowledge-grounded dialogues With the advances in large-scale language models, dialogue agents can now generate high-quality responses using parametric knowledge ([Thoppilan et al., 2022](#); [Freitas et al., 2020](#); [Bao et al., 2021](#)). However, hallucination remains a challenge, which means that the language model tends to generate plausible-looking statements that are factually incorrect ([Shuster et al., 2021](#)). To address this problem, knowledge-augmented approaches are applied in dialogue generation ([Lewis et al., 2020b](#)). In knowledge-grounded dialogues (KGD), the dialogue models first select a knowledge snippet from

¹The code is available at: <https://github.com/sunweiwei/GenKS>

passages and then generate the responses (Liu et al., 2018; Dinan et al., 2019).

Knowledge selection As the critical step in KGD, knowledge selection has received many studies. The existing methods mainly employ *classification* model with dual-encoder (Dinan et al., 2019; Kim et al., 2020) or cross-encoder (Zhao et al., 2020b) architecture. However, the classification paradigm is unable to capture the knowledge interaction in KGD (Kim et al., 2020; Li et al., 2022). To address this problem, Li et al. (2022) propose a graph-based method to capture the relationship between candidate snippets, Zhan et al. (2021a) and Wu et al. (2021) employ machine reading comprehension model to extract span from long document. Sequential knowledge selection has also been proposed to capture the topic transition in conversations (Kim et al., 2020; Zhan et al., 2021b; Zheng et al., 2020; Meng et al., 2020; Yang et al., 2022). Despite their effectiveness, the existing methods have two drawbacks: (1) they use compact vectors to represent dialogue and knowledge and thus lack deep semantic interactions; (2) they are complicated and challenging to capture intra-knowledge and dialogue-knowledge interactions simultaneously. We address these drawbacks by shifting the modeling paradigm of knowledge selection to identifier generation (Sun et al., 2022), and propose GENKS to capture the two types of interaction simultaneously using Transformer (Vaswani et al., 2017).

Generative knowledge selection A generative paradigm for knowledge selection is not foreign to the NLP community; for example, sequence-to-sequence models have been applied on entity retrieval (Cao et al., 2021), document ranking (Nogueira et al., 2020; Tay et al., 2022), multi-evidence retrieval (Min et al., 2021; Yavuz et al., 2022), and etc. Our proposed model GENKS differs from existing methods in the following ways: (1) we are the first to explore generative knowledge selection in KGD; (2) we consider the effectiveness of intra-knowledge interaction; (3) we design hyperlinks to capture the interaction between knowledge and dialogue.

3 GENKS

We provide an overview of GENKS in Figure 2. As shown in Figure 2, the dialogue data is first serialized into a sequence. Then a sequence-to-sequence model (i.e., BART) is employed to select knowl-

edge and get the response by generating the target sequence autoregressively. In this section, we first formulate the task in Section 3.1. Then, we detail the serialization (Section 3.2) and optimization (Section 3.3) methods.

3.1 Problem formulation

Suppose that we have a case of knowledge-grounded dialogues (C, \mathcal{K}, r) , where $C = (c_1, \dots, c_{|C|})$ is a dialogue context that contains $|C|$ utterances, r is the response to C , $\mathcal{K} = (K_1, \dots, K_{|\mathcal{K}|})$ denotes $|\mathcal{K}|$ passages that are relevant to C ; for each i , $K_i = (k_{i,1}, \dots, k_{i,|K_i|})$ denotes a passage that contains $|K_i|$ snippets. We define $m = \sum_{i=1}^{|\mathcal{K}|} |K_i|$ as the total number of snippets in \mathcal{K} . A knowledge-grounded dialogue agent is decoupled into two modules: a knowledge selection module $P(k|C, \mathcal{K})$ that selects a snippet from \mathcal{K} ; a response generation module $P(r|C, \mathcal{K}, k_s)$ where k_s is the selected snippet from knowledge selection module.

3.2 Serialization

We formulate the knowledge selection task as a procedure of sequence generation. As shown in Figure 2, the dialogue context C and knowledge candidates \mathcal{K} are mapped into a sequence and then fed into a sequence-to-sequence model. The model’s output is converted back to the selected knowledge k or the response r .

Specifically, we first assign an identifier to each snippet in \mathcal{K} , sequentially starting from $\langle k1 \rangle$ to $\langle km \rangle$. Then we convert passages \mathcal{K} into a sequence using a template that packages snippets with the corresponding identifiers and concatenates them in order; see the green block in Figure 2. Similarly, the dialogue context C is serialized by adding task prompts, i.e., task description and speaker name, as shown in the blue block in Figure 2.

In multi-turn dialogues, the knowledge appearing in the dialogue history prompts the discourse structure of knowledge transition and knowledge expression. Hence we propose a *hyperlink* method to capture the dialogue-knowledge interaction explicitly. We provide an example of the hyperlink method in Figure 2. We see that the first utterance of User1 refers to a snippet (whose identifier is $\langle k2 \rangle$) in the passage “Skateboarding”. We thus add a hyperlink to the utterance. The hyperlink includes the identifier and the title of the snippet, i.e., annotating [Skateboarding] $\langle k2 \rangle$ at the beginning of

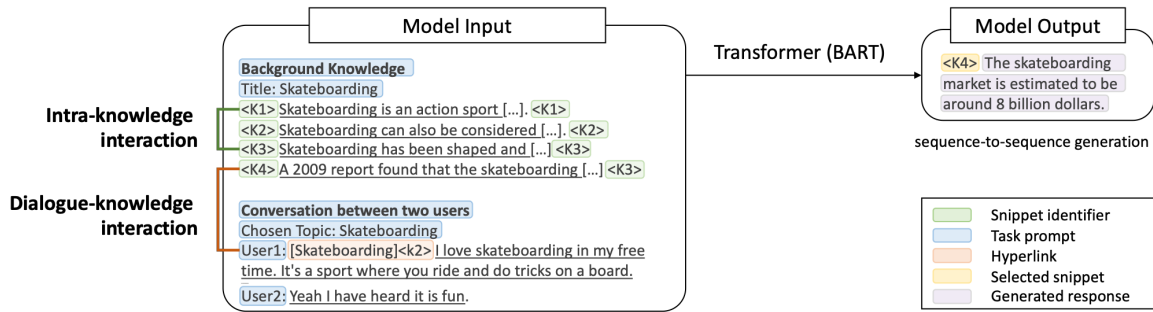


Figure 2: Overview of GENKS. The dialogue context and the knowledge are serialized and fed into a seq-to-seq model, BART. The outputs are the identifier of the selected snippet (i.e., <k5>) and the response.

this utterance (as shown in the red block in Figure 2). Finally, we splice the passages and dialogue context sequences as input for a Transformer model (i.e., BART). Therefore, the model can capture the intra-knowledge and dialogue-knowledge interactions through a *self-attention* mechanism (Vaswani et al., 2017).

3.3 Optimization

The knowledge selection model is optimized by the cross-entropy loss: $\mathcal{L} = -\log P(k_{true}|C, \mathcal{K})$, where k_{true} denotes the label knowledge. Since k_{true} needs to be labeled manually and is not available in some scenarios (Zhou et al., 2018), we construct pseudo-labels for model training following Zhao et al. (2020b) in cases the knowledge label is absent. In particular, we calculate the F1 score (Dinan et al., 2019) between each knowledge snippet and the response. We use the snippet with the highest score as the pseudo label. Such a method is based on the intuition that human responses provide hints regarding the relevance of the snippets (Zhao et al., 2020b; Li et al., 2020).

Since both knowledge selection and response generation are modeled with the *generative* paradigm, we unify the two modules with one joint generative model. In this joint model, the knowledge selection and the response generation are optimized jointly, with shared parameters. To this end, we splice the knowledge identifier k_{true} and response r into one sequence (as shown in Figure 2). Then, we optimize the sequence-to-sequence model using cross-entropy loss on all the tokens of the target sequence. In inference, the model generates knowledge identifier k_s and responses r in an autoregressive fashion. We note that the end-to-end model allows the two tasks to be mutually enhanced and improves the model’s efficiency.

4 Experimental setup

4.1 Datasets

We conduct experiments on Wizard of Wikipedia (WoW) (Dinan et al., 2019), Holl-E (Moghe et al., 2018), and CMU_DoG (Zhou et al., 2018). The statistical details on these three datasets are shown in Table 7 in the appendix.

- **WoW** is an open-domain KGD dataset using Wikipedia passage as background knowledge. The test set of Wizard is split into seen and unseen versions, where the unseen test set contains 58 new topics not discussed in the training data.
- **Holl-E** focuses on the movie domain. The background knowledge consists of plots, comments, and movie reviews collected from different websites. Holl-E has two versions of the test set: single test and multi-reference test. In the multi-reference test, there are multiple human-annotated ground-truth knowledge and corresponding responses for each instance.
- **CMU_DoG** focuses on the domain of movies. The workers discuss a movie in depth given the background knowledge (e.g., introduction, plots, and key scenes).

4.2 Baselines

We compare GENKS with baselines of two categories: (i) *End-to-end methods* that generate response directly without explicit knowledge selection, and (ii) *Pipeline methods* that explicitly select knowledge snippet to be used in response.

The end-to-end methods we consider are:

- **BART** (Lewis et al., 2020a) that generates responses without access to the external passage and uses knowledge inside model parameters instead.
- **BART FID** (Izacard and Grave, 2021) concatenates and encodes each candidate knowledge

with dialogue separately and fuses all the encoded representation in the decoder to generate the response.

- **BART RAG-DPR** is a baseline adopted by Adolphs et al. (2021), which uses DPR-retrieved passages and produces response using RAG.
- **BART FiD-RAG DPR-Poly** (Shuster et al., 2021) uses DPR-Poly to retrieve passage and uses FiD-RAG to generate the response.

Regarding the pipeline baselines, according to their knowledge selection modeling paradigm, we sub-categorize pipeline baselines into four groups:

(1) The *Classification methods*, includes:

- **SKT** (Kim et al., 2020) proposes sequential knowledge selection.
- **DiffKS** (Zheng et al., 2020) captures the knowledge differences between adjacent turns.
- **DukeNet** (Meng et al., 2020) models the knowledge shift and tracking processes with a dual learning scheme.
- **KnowledGPT** (Zhao et al., 2020b) exploits pre-trained language models in KGD.
- **MIKe** (Meng et al., 2021) distinguish user-initiative and system-initiative.
- **K-Mine** (Lotfi et al., 2021) proposes a score-and-aggregate module.
- **TAKE** (Yang et al., 2022) propose a topic-shift aware network.

(2) The *MRC methods*, includes:

- **CoLV** (Zhan et al., 2021a) proposes a collaborative latent variable model.
- **DIALKI** (Wu et al., 2021) proposes a MRC-based model to extract span from passage.

(3) The *Graph-based methods*, includes:

- **Graph** (Li et al., 2022) builds a semantic graph upon candidate documents and employs a GNN model.

(4) And the *Knowledge generation methods*, includes:

- **K2R** (Adolphs et al., 2021) uses the RAG-based model to generate knowledge text and then generates dialogue response based on it.

4.3 Evaluation metrics

In WoW, we choose perplexity (**PPL**) of the ground-truth responses, unigram **F1**² (Dinan et al., 2019), **Knowledge-F1** (Shuster et al., 2021), and **BLEU-4** (Papineni et al., 2002) score as metrics. In Holl-E, we additionally use **ROUGE-1**, and **ROUGE-2** following Meng et al. (2020). In

²<https://github.com/facebookresearch/ParLAI>

CMU_DoG, we additionally use embedding-based metrics includes **Average**, **Extreme**, and **Greedy** following Zhao et al. (2020b).

In addition, we randomly sample 100 examples from the WoW test seen and WoW test unseen, respectively, and recruit three experts for human evaluation. The annotators are asked to judge the model-generated response in four ways:

- **Fluency**, which measures whether the response is fluency in expression;
- **Coherence**, which measures whether the response is coherence to the dialogue context;
- **Relevance**, which measures whether the knowledge used in the response is relevant to the dialogue; and
- **Factuality** measures whether the response’s content is factual. In Factuality evaluation, the experts check the content using Google.

The annotators are asked to assign a score in {0, 1} (representing “nonfactual” and “factual”) for factuality, and a score in {0, 1, 2} (representing “bad”, “fair”, and “good”) for the others.

4.4 Implementation details

We implement the GENKS using BART large (with 400M parameters) (Lewis et al., 2020a) in HuggingFace’s Transformers library. We truncate the dialogue context to 256 tokens, then truncate the knowledge so that the total length is less than 1024 tokens. During inference, the responses are decoded using a greedy search. See Appendix A for more details.

Typically, the number of passages in \mathcal{K} is large, so that the input sequence exceeds the maximum input length of BART (i.e., 1024 tokens). To address this problem, we take advantage of a lightweight passage selector based on DistilBERT (with 66M parameters) (Sanh et al., 2019), which aims to rank the passages in \mathcal{K} . Specifically, we concatenate each passage with dialogue context and encode the sequence using DistilBERT. Finally, the representation of [CLS] token is used to estimate the relevance score of the passage through a learnable MLP classifier. The passage selector is optimized via contrastive learning objective (Nogueira and Cho, 2019), in which the model learns to assign a higher score to positive passages than negative passages. During inference, we keep only the top-1 passage ranked by the passage selector. The passage selector gets Recall@1 of 75.5%, 76.5%, and 68.0% for the WoW test seen, WoW test unseen,

| Methods | WoW | | Holl-E | |
|-----------------------------------|-------------|-------------|-------------|-------------|
| | Seen | Unseen | Single | Multi |
| <i>Classification methods</i> | | | | |
| SKT (Kim et al., 2020) | 26.8 | 18.3 | 29.2 | 39.2 |
| DukeNet (Meng et al., 2020) | 26.4 | 19.6 | 30.0 | 40.3 |
| DiffKS (Zheng et al., 2020) | 25.5 | 19.7 | 33.0 | - |
| KnowledGPT (Zhao et al., 2020b) | 28.0 | 25.4 | - | - |
| MIKe (Meng et al., 2021) | 28.4 | 21.5 | 31.9 | 41.8 |
| K-Mine (Lotfi et al., 2021) | 29.7 | 28.3 | 31.7 | - |
| TAKE (Yang et al., 2022) | 28.8 | 25.8 | - | - |
| <i>Other methods</i> | | | | |
| CoLV (Zhan et al., 2021a) | 30.1 | 18.9 | 32.7 | - |
| DIALKI (Wu et al., 2021) | <u>32.9</u> | <u>35.5</u> | - | - |
| Graph (Li et al., 2022) | 29.4 | 30.8 | <u>37.7</u> | <u>46.1</u> |
| GenKS | 34.2 | 36.6 | 37.9 | 46.8 |
| <i>Variants for comparison</i> | | | | |
| - BART classification | 29.8 | 29.7 | 34.0 | 44.0 |
| - BART classification w/ position | 30.1 | 31.2 | 34.0 | 44.0 |
| - Hierarchical classification | 30.0 | 31.4 | 33.8 | 43.7 |
| - Without passage selector | 31.4 | 32.0 | 34.5 | 44.4 |
| - Unorder knowledge snippets | 31.8 | 33.3 | 36.5 | 45.8 |
| - Without hyperlink | 33.4 | 35.4 | 36.9 | 45.4 |

Table 1: Knowledge selection accuracy on WoW (seen and unseen test set) and Holl-E (single reference and multi-reference test set). **Bold** denote the best results with significant improvements over the previous SOTA (t-test, $p < 0.05$). Underline denote second best results.

and Holl-E, respectively.

5 Experimental results

5.1 Performance on knowledge selection

We evaluate the knowledge selection effectiveness of GENKS on WoW and Holl-E, respectively³. In Table 1, we compare the knowledge selection accuracy of GENKS with previous pipeline methods. Results show that GENKS achieves the best knowledge selection accuracy on both datasets and consistently outperforms baselines.

We find that GENKS particularly excels at topics that do not appear during training (see WoW unseen test split). For example, the classification models both have noticeable accuracy drops on the unseen topic. In contrast, models that model the intra-knowledge interaction (e.g., GENKS, Graph, DIALKI) can better understand the knowledge of unseen topics⁴.

To evaluate the Performance of GENKS as dialogue goes deeper, we compare GENKS with

³We cannot evaluate the knowledge selection accuracy on CMU_DoG because the knowledge snippets used in responses are not manually labeled.

⁴The higher results on unseen than seen might be due to the smaller number of topics in the unseen test set.

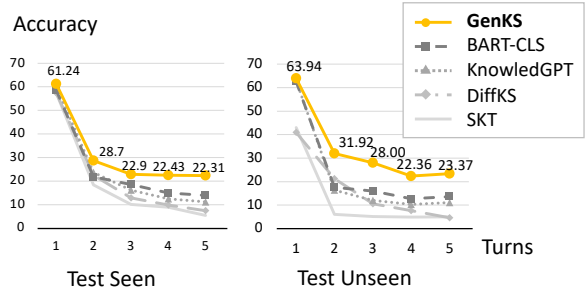


Figure 3: Knowledge selection accuracy over different dialogue turns. BART-CLS represents a text-matching model with cross-encoder architecture.

four classification baselines (SKT, DiffKS, KnowledGPT, and BART-CLS) overturns. Figure 3 shows the results. Both methods achieve good accuracy in the first few turns. However, as the conversation dives deeply into a topic, a significant performance decline can be seen in baseline methods. In contrast, GENKS that explicitly captures the multi-turn dialogue-knowledge interaction, achieves a relatively high accuracy (around 22%-23%).

5.2 Quality of generated responses

We report response generation evaluation results on WoW in Table 2. The results on Holl-E and CMU_DoG are available in Table 8 and Table 9 in the appendix. The results of baselines are cited from original papers or re-evaluated using officially released checkpoints.

Compared with previous pipeline models, GENKS achieve the best Performance on almost all metrics. For example, GENKS surpasses KnowledGPT by 0.7% and 2.4% in terms of F1 on WoW seen and WoW unseen, respectively. Note that the improvements on the unseen test set are more notable than on the seen test set, which agrees with the experimental results regarding knowledge selection. GENKS also achieve competitive results compared to SOTA end-to-end models. For example, GENKS performs comparably to BART FiD-RAG DPR-Poly on WoW seen and outperformed on WoW unseen.

5.3 Ablation study about knowledge selection

To analyze the effect of each component in GENKS, we designed several variants and conducted an ablation study about knowledge selection. Results are listed in Table 1, “Variants for comparison”. The details of compared variants and the findings are as follows:

BART classification We use BART to classify each

| Methods | WoW Seen | | | | WoW Unseen | | | |
|--|-------------|--------------|-------------|-------------|-------------|--------------|--------------|-------------|
| | PPL | F1 | KF1 | B4 | PPL | F1 | KF1 | B4 |
| <i>End-to-end models</i> | | | | | | | | |
| BART (Lewis et al., 2020a) | 14.7 | 20.9 | 17.4 | 1.7 | 18.9 | 18.8 | 15.1 | 0.9 |
| BART FiD (Izacard and Grave, 2021) | 17.0 | 21.5 | 20.0 | 3.6 | 18.4 | 20.6 | 19.2 | 3.2 |
| BART RAG-DPR (Adolphs et al., 2021) | 11.5 | 22.6 | 26.1 | 3.7 | 13.1 | 21.5 | 22.7 | 3.0 |
| BART FiD-RAG DPR-Poly (Shuster et al., 2021) | 11.4 | 22.1 | 29.7 | 4.1 | 13.1 | 21.1 | 27.1 | 3.8 |
| <i>Pipeline models</i> | | | | | | | | |
| DukeNet (Meng et al., 2020) | 48.3 | 19.3 | 18.5 | 2.4 | 69.4 | 17.1 | 16.5 | 1.7 |
| CoLV (Zhan et al., 2021a) | 39.5 | 20.3 | 18.2 | 2.8 | 54.3 | 18.5 | 17.5 | 2.1 |
| KnowledGPT (Zhao et al., 2020b) | 19.2 | 22.0 | 23.8 | 3.7 | 22.3 | 20.5 | 22.1 | 3.0 |
| K-Mine (Lotfi et al., 2021) | 13.2 | 21.8 | - | - | 16.4 | 21.1 | - | - |
| K2R RAG-DPR (Adolphs et al., 2021) | 18.3 | 22.0 | 27.3 | 3.7 | 22.3 | 19.9 | 23.2 | 2.8 |
| K2R BART RAG-DPR (Adolphs et al., 2021) | 17.9 | 21.3 | 29.2 | 3.5 | 21.1 | 19.9 | 24.3 | 2.5 |
| GenKS | 13.1 | 22.9* | 29.5 | 4.5* | 13.2 | 22.7* | 28.1* | 4.6* |
| <i>Ablative variants</i> | | | | | | | | |
| - With BART classification knowledge | 14.7 | 22.0 | 25.9 | 3.5 | 16.2 | 21.1 | 24.4 | 3.1 |
| - Without identifiers generation | 13.8 | 21.7 | 23.2 | 3.7 | 14.1 | 21.8 | 23.3 | 3.9 |
| - Without hyperlink | 14.2 | 22.1 | 27.2 | 3.9 | 15.5 | 22.3 | 26.9 | 4.2 |
| - With oracle knowledge | 8.9 | 38.8 | 74.2 | 13.1 | 10.5 | 38.9 | 74.5 | 12.8 |

Table 2: Evaluation results on WoW seen and unseen test set in terms of response quality. We compare against the ground-truth dialogue response in terms of perplexity (PPL), F1, Knowledge F1 (KF1), and BLEU-4 (B4). The four groups lists previous end-to-end models, previous pipeline models, GenKS, and ablative variants. The best results are highlighted with **bold**, and the second-best results are highlighted with underline. * indicates significant improvements over all baselines with p-value < 0.05.

candidate snippet into two classes: “relevant” or “irrelevant”. The results show that BART in the classification paradigm performs worse than GENKS by a large margin.

BART classification w/ position To understand the influence of position bias, we splice the snippet’s position into the classification model’s input. We find that the results are improved to a certain extent (about 1% improvement), but there is still a clear gap compared with GENKS.

Hierarchical classification This variant first uses the passage selector model of GENKS to rank the passages and then selects the snippets in the top-ranked passage using BART classification w/ position. The results show that the passage selector does not affect the classification model’s Performance.

Without passage selector When the passage selector model of GENKS is removed, the model has more probability of truncating the label knowledge, resulting in an evident decline in Performance.

Unorder knowledge snippets To disable the intra-knowledge interaction, we unorder the snippets so that order of the snippets is inconsistent with the original passages. This variant shows a decline in selection accuracy, especially on unseen topics, indicating that keeping the order of the snippets in

the passage is necessary.

Without hyperlinks We remove the hyperlinks in the dialogue context. About a 1% accuracy drop is seen, indicating the effectiveness of hyperlinks.

5.4 Ablation study about response generation

As shown in Table 2, we also conduct an ablation study about response generation. The details of compared variants and the findings are as follows:

With BART classification knowledge When replacing the generated identifier with the knowledge selected by BART classification, a performance decline is witnessed—the F1 value drops by 0.7% and 1.8% on Wizard seen and unseen, sustaining the effectiveness of the knowledge selection of GENKS.

Without identifier generation This variant removes the identifier generation by directly generating the response. We see notable performance drops, especially in the KF1 metric. The results indicate that explicit training and inference about knowledge selection enable to use of more appropriate knowledge in response generation.

Without hyperlinks This variant removes hyperlinks from GENKS. It performs worse than GENKS, probably due to its lower accuracy of knowledge selection than GENKS.

Use the oracle knowledge We replace the model-predicted snippet identifier with the oracle one

| Methods | WoW Seen | | | | WoW Unseen | | | |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | Flu. | Coh. | Rel. | Fact. | Flu. | Coh. | Rel. | Fact. |
| BART | 1.82 | 1.51 | 1.45 | 0.82 | 1.76 | 1.50 | 1.47 | 0.76 |
| BART FiD | 1.88 | 1.70 | 1.55 | 0.84 | 1.85 | 1.67 | 1.53 | 0.82 |
| TMN | 1.59 | 1.41 | 1.08 | 0.62 | 1.42 | 1.30 | 0.98 | 0.59 |
| DukeNet | 1.69 | 1.56 | 1.22 | 0.71 | 1.66 | 1.47 | 1.10 | 0.72 |
| KnowledGPT | 1.89 | 1.67 | 1.58 | 0.87 | 1.87 | 1.68 | 1.51 | 0.83 |
| GenKS | 1.90 | 1.72 | 1.69 | 0.89 | 1.91 | 1.71 | 1.67 | 0.91 |

Table 3: Human evaluation results. Flu, Coh, Rel, and Fact denote Fluency, Coherence, Relevance, and Factuality, respectively.

| | KS | RG | ACC | F1 |
|----------------------------------|------|------|------|------|
| BART RAG (Adolphs et al., 2021) | 0 | 11.5 | - | 21.5 |
| BART FiD* (Shuster et al., 2021) | 0 | 25.4 | - | 21.1 |
| KnowGPT (Zhao et al., 2020b) | 12.5 | 7.4 | 25.4 | 20.5 |
| K2R* (Adolphs et al., 2021) | 11.8 | 8.2 | - | 19.9 |
| DIALI* (Li et al., 2022) | 8.2 | - | 35.5 | - |
| Graph* (Li et al., 2022) | 15.4 | - | 30.8 | - |
| GenKS | 2.9 | 8.8 | 36.6 | 22.9 |

Table 4: Inference time (minutes) on one GPU on WoW unseen test set. The values of model with * are estimated based on the model size and input/output length. KG and RG denote inference time for knowledge selection and response generation stage, respectively.

(knowledge used by ground-truth response). The results (e.g., KF1=74) suggest that GENKS can effectively locate and incorporate the corresponding knowledge into the responses following the guidance of the identifier.

5.5 Human evaluation

Table 3 shows the human-evaluating results. Results show that GENKS consistently outperforms baselines on all datasets. The Fleiss’ kappa value is above 0.60, indicating substantial agreement among the annotators. GENKS outperforms KnowledGPT by about 0.02 and DukeNet by about 0.20 in terms of response generation evaluation metrics (i.e., *Fluency* and *Context Coherence*). Moreover, for the *Knowledge Relevance*, the annotators agree that GENKS is capable of selecting knowledge that is more relevant to the dialogue and generating more informative responses than baselines. The *Factuality* results show that by explicitly identifying the knowledge snippet used in response, GENKS can reduce the hallucination of response generation.

| Methods | WoW Seen | | | WoW Unseen | | |
|--------------------|----------|------|-----|------------|------|-----|
| | F1 | KF1 | B4 | F1 | KF1 | B4 |
| GenKS | 22.9 | 29.5 | 4.5 | 22.7 | 28.1 | 4.6 |
| GenKS-2 | 22.4 | 29.3 | 4.2 | 22.2 | 27.6 | 4.2 |
| GenKS (5 Snippets) | 22.3 | 27.6 | 4.2 | 21.8 | 25.5 | 4.1 |
| GenKS (3 Snippets) | 21.1 | 29.3 | 3.2 | 20.0 | 20.9 | 2.9 |
| GenKS (128 Tokens) | 21.5 | 25.6 | 3.5 | 20.7 | 22.9 | 3.4 |
| GenKS (64 Tokens) | 20.7 | 23.3 | 3.0 | 20.1 | 20.6 | 2.9 |

Table 5: Analytical experiment results on WoW. The first group compares GENKS and its variant GENKS, which selects two snippets instead of one. The second group includes the results of GENKS with different maximum number of snippets inputs or maximum input tokens.

5.6 Efficiency evaluation

To evaluate the efficiency of GENKS, we compare the model with previous end-to-end models and pipeline models. The results listed in Table 4 show that GENKS is more efficient than previous pipeline models. We infer that this phenomenon is because GENKS jointly models knowledge selection and response generation, avoiding repeated encoding of dialogue history and knowledge. As a pipeline method, we also find that GENKS achieves comparable efficiency compared to end-to-end models like RAG, but benefits from explicit knowledge selection.

5.7 Analytical experiment

Multi-snippets selection GENKS select a single snippet following the experimental setup outlined in the baselines (Dinan et al., 2019), but it can also select multiple snippets by generating multiple identifiers. We test a variant of our GENKS model, GENKS-2, which selects two snippets by generating two identifiers consecutively. We compared its performance with the original GENKS on the WoW dataset. The results are listed in Table 5 group 1. GENKS-2 performs slightly worse than the original GenKS, likely because the WoW dataset only uses one snippet in response annotation and therefore does not benefit from using multiple snippets (Dinan et al., 2019). Nevertheless, the results suggest that the proposed generative knowledge selection approach has the ability to select multiple knowledge.

Hyper-parameter analysis We also conduct ablation experiments on the number of input snippets to the model and maximum input tokens. The results are listed in Table 5 group 2. We find that

| | |
|---------------|---|
| Topic | Budweiser |
| User | I think Budweiser taste terrible. Have you ever had it? |
| Know 1 | Produced in various breweries around the world, Budweiser is a filtered beer available in draft and packaged forms. |
| Res 1 | Yes, I have. It is produced in various breweries around the world! |
| Know 2 | Budweiser is an American-style pale lager produced by Anheuser-Busch, currently part of the transnational corporation Anheuser-Busch InBev. |
| Res 2 | Yes, I have. I know that it is an American-style pale lager produced by Anheuser-Busch. |

Table 6: Examples of GENKS outputs on the WoW.

reducing the number or length of knowledge reduces model effectiveness.

5.8 Case study

To better understand end-to-end baselines and our model, we provide an example in Table 6, which shows that GENKS appropriately changes its response prediction when providing different knowledge snippets⁵. Therefore, GENKS is more controllable and interpretable than end-to-end models, where the end-to-end system is a black box. We provide more case studies in Appendix B.

6 Conclusion

In this paper, we have proposed GENKS, a simple yet effective knowledge-grounded dialogue model. GENKS is a generative model, which learns to select knowledge snippets by generating their identifiers. Benefiting from the modeling of intra-knowledge interaction and dialogue-knowledge interaction, GENKS effectively addresses the challenges of *ambiguity* and *discourse structure*. Our experiments have shown that GENKS establishes a new state-of-the-art on three knowledge-grounded dialogue benchmarks. Notably, GENKS particularly excels at new topics and as the dialogue goes deeper. GENKS also outperforms SOTA end-to-end models. Hence, we believe GENKS reveals a new paradigm for knowledge selection in open-domain dialogue.

⁵Note that this example only aims to show the output of the model. In fact, according to <https://en.wikipedia.org/wiki/Budweiser>, Budweiser is also a famous lager from the Czech Republic, and the American Budweiser being sold and known as Bud through most of the European Union.

Limitations

The limitations of this work include the modular modeling of passage reranks, which reduces the efficiency of the approach. Besides, we only conduct human evaluation on one popular dataset, i.e., Wizard of Wikipedia. Furthermore, the effectiveness of GENKS is only verified in the English dataset. Research on other languages establishes a new challenge, especially for languages with limited knowledge and annotated data. In future work, we would like to explore more efficient passage rerank techniques on knowledge-grounded dialogues. We will also conduct human evaluation for more datasets. Besides, generative knowledge selection can be extended to future studies about conversational recommendation.

Ethics statement

The paper proposes a knowledge-grounded dialogue system to generate a response using external knowledge. The intended use of this system is to perform chit-chat with the user on topics such as books and movies. The system is developed using large pre-trained language models (i.e., BART), who are trained on large-scale web data known to contain biased or discriminatory content. The datasets (i.e., WoW, Holl-E, CMU_DoG) that we train on also include subjective knowledge (comments on movies) that may express the bias of the writers. Although the system is able to reduce the hallucination of response compared to end-to-end models, the outputs from our system may still contain non-factual information and should not be considered as advice for any critical decision-making.

Acknowledgements

This work was supported by the National Key R&D Program of China with grant No. 2020YFB1406704, the Natural Science Foundation of China (62272274, 62202271, 61902219, 61972234, 62072279, 62102234), the Natural Science Foundation of Shandong Province (ZR2021QF129), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

- Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur D. Szlam, and Jason Weston. 2021. [Reason first, then respond: Modular generation for knowledge-infused dialogue](#). *ArXiv*, abs/2111.05204.
- Siqi Bao, H. He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zhengyu Niu. 2021. [Plato-xl: Exploring the large-scale pre-training of dialogue generation](#). *ArXiv*, abs/2109.09519.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. [Autoregressive entity retrieval](#). In *ICLR 2021*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of wikipedia: Knowledge-powered conversational agents](#). In *ICLR 2019*.
- Daniel De Freitas, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a human-like open-domain chatbot](#). *ArXiv*, abs/2001.09977.
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *EACL 2021*.
- Byeongchang Kim, Jae Hyun Ahn, and Gunhee Kim. 2020. [Sequential latent knowledge selection for knowledge-grounded dialogue](#). In *ICLR 2020*.
- 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 2020*, pages 7871–7880.
- Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *NIPS 2020*, volume 33, pages 9459–9474.
- Lin-Xiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. [Zero-resource knowledge-grounded dialogue generation](#). In *NIPS 2020*, pages 8475–8485.
- Sha Li, Madhi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Z. Hakkani-Tür. 2022. [Enhancing knowledge selection for grounded dialogues via document semantic graphs](#). In *NAACL 2022*.
- Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Q. Li, and Jie Zhou. 2019. [Incremental transformer with deliberation decoder for document grounded conversations](#). In *ACL 2019*, pages 12–21.
- Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. [Learning to select knowledge for response generation in dialog systems](#). In *IJCAI 2019*, pages 5081–5087.
- Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. [Knowledge diffusion for neural dialogue generation](#). In *ACL 2018*, pages 1489–1498.
- Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. [Knowledge aware conversation generation with explainable reasoning over augmented graphs](#). In *EMNLP 2019*, pages 1782–1792.
- Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2021. [Teach me what to say and i will learn what to pick: Unsupervised knowledge selection through response generation with pretrained generative models](#). In *EMNLP 2021 | NLP4ConvAI*, pages 254–262.
- Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and M. de Rijke. 2021. [Initiative-aware self-supervised learning for knowledge-grounded conversations](#). In *SIGIR 2021*, page 522–532.
- Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and M. de Rijke. 2020. [Duketnet: A dual knowledge interaction network for knowledge-grounded conversation](#). In *SIGIR 2020*, page 1151–1160.
- Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021. [Joint passage ranking for diverse multi-answer retrieval](#). In *EMNLP 2021*, pages 6997–7008.
- Nikita Moghe, Siddharth Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. [Towards exploiting background knowledge for building conversation systems](#). In *EMNLP 2018*, pages 2322–2332.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with bert](#). *ArXiv*, abs/1901.04085.
- Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy J. Lin. 2020. [Document ranking with a pre-trained sequence-to-sequence model](#). In *Findings of EMNLP 2020*, pages 708–718.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *ACL 2002*, pages 311–318.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *ArXiv*, abs/1910.01108.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of EMNLP 2021*, pages 3784–3803.

- Tianxiang Sun, Xiangyang Liu, Xipeng Qiu, and Xuanjing Huang. 2022. [Paradigm shift in natural language processing](#). *Int. J. Autom. Comput.*, 19:169–183.
- Yi Tay, Vinh Quang Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. [Transformer memory as a differentiable search index](#). *ArXiv*, abs/2202.06991.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. [Lamda: Language models for dialog applications](#). *ArXiv*, abs/2201.08239.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NIPS 2017*, volume 30.
- Zequi Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. [Dialki: Knowledge identification in conversational systems through dialogue-document contextualization](#). In *EMNLP 2021*.
- Chenxu Yang, Zheng Lin, JiangNan Li, Fandong Meng, Weiping Wang, Lan Wang, and Jie Zhou. 2022. [Take: Topic-shift aware knowledge selection for dialogue generation](#). In *COLING 2022*.
- Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, Nishit Shirish Keskar, and Caiming Xiong. 2022. [Modeling multi-hop question answering as single sequence prediction](#). In *ACL 2022*, pages 974–990.
- Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021a. [Colv: A collaborative latent variable model for knowledge-grounded dialogue generation](#). In *EMNLP 2021*, pages 2250–2261.
- Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021b. [Augmenting knowledge-grounded conversations with sequential knowledge transition](#). In *NAACL 2021*, pages 5621–5630.
- Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. [Low-resource knowledge-grounded dialogue generation](#). In *ICLR 2020*.
- Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. [Knowledge-grounded dialogue generation with pre-trained language models](#). In *EMNLP 2020*, pages 3377–3390.
- Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. [Difference-aware knowledge selection for knowledge-grounded conversation generation](#). In *Findings of EMNLP 2020*, pages 115–125.
- Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. [A dataset for document grounded conversations](#). In *EMNLP 2018*, pages 708–713.

| | WoW | Holl-E | CUM_DoG |
|------------------------|-----------|--------|---------|
| Training size | 18,430 | 7,228 | 3,373 |
| Validation size | 1,948 | 930 | 229 |
| Test size | 965 / 968 | 913 | 619 |
| Number of topics | 1,365 | 858 | 30 |
| Avg. Turn per dialogue | 9.0 | 10.1 | 22.2 |
| Avg. Num of snippets | 62.5 | 57.3 | 36.3 |
| Avg. Num of passages | 11.6 | 5.9 | 4.0 |

Table 7: Statistics of two experimental datasets, Wizard of Wikipedia (WoW), Holl-E, and CMU_DoG. The two numbers in WoW indicate the size of seen and unseen test set, respectively.

| | F1 | B4 | KF1 | RG1 | RG2 |
|----------------------------|-------------|-------------|-------------|-------------|-------------|
| BART (Lewis et al., 2020a) | 34.7 | 22.3 | 29.1 | 38.0 | 27.9 |
| CoLV (Zhan et al., 2021a) | - | 20.3 | - | 32.0 | 25.8 |
| MIKe (Meng et al., 2021) | 32.1 | 21.1 | - | 38.0 | 25.2 |
| Graph (Li et al., 2022) | - | - | - | 42.5 | 34.4 |
| GenKS | 36.7 | 24.3 | 31.3 | 42.3 | 35.2 |

Table 8: Results on Holl-E in term of response quality. RG1 and RG2 denote ROUGE-1 and ROUGE-2 respectively. Best results are heighten with **bold**.

A Implementation details

We use gradient clipping with a maximum gradient norm of 0.1. We optimize the model for up to 5 epochs with a batch size of 16 on 4 3090 GPUs with 24G memory. We choose the model checkpoints by evaluating the metrics on the validation set for each epoch. During inference, the responses are decoded using a greedy search. We have tried some advanced decoding algorithms (e.g., nucleus sampling) and found no improvement. The training of the model can be completed within 5h, and the latency of the model inference for one example is within 0.1s. The passage rerank model gets Recall@1 of 75.5%, 76.5%, 61.0% for WoW test seen, WoW test unseen, and Holl-E, respectively.

| | PPL | F1 | Avg. | Ext. | Greedy |
|---------------------------|-------------|-------------|--------------|--------------|--------------|
| ITDD (Li et al., 2019) | 26.0 | 10.4 | 0.748 | 0.390 | 0.587 |
| DRD (Zhao et al., 2020a) | 46.1 | 10.8 | 0.791 | 0.406 | 0.613 |
| TMN (Dinan et al., 2019) | 75.2 | 9.9 | 0.789 | 0.399 | 0.615 |
| KGPT (Zhao et al., 2020b) | 20.6 | 13.5 | 0.837 | 0.437 | 0.654 |
| GenKS | 16.1 | 14.1 | 0.847 | 0.442 | 0.668 |

Table 9: Results on CMU_DoG in term of response quality. The best results are highlighted with **bold**.

| | |
|----------------|---|
| Topic | Nickelback |
| User | Do you like Nickelback? |
| ... | |
| User | Yes, what more can you tell me about Nickelback? |
| System | Chad Kroeger is the leading vocalist of the group. |
| User | Amazing. What about the other group members |
| Passage | The band is composed of guitarist and lead vocalist <i>Chad Kroeger</i> , guitarist, keyboardist and backing vocalist <i>Ryan Peake</i> , bassist <i>Mike Kroeger</i> , and drummer <i>Daniel Adair</i> . |
| TMN | i think the song is a very popular song |
| DukeNet | chad kroeger is a great band |
| KGPT | the lead vocalist is chad kroeger. he is also the drummer. |
| GenKS | Chad Kroeger, Ryan Peake, Mike Kroeger and Daniel Adair. |
| Human | Ryan Peake is the keyboardist an Mike Kroeger is the bassist. |

Table 10: Case study on the Wizard Test Unseen dataset. This table shows an example where both GENKS and baselines select the proper knowledge.

| | |
|----------------|--|
| Topic | List of national parks of the United States |
| User | I love national parks dont you |
| ... | |
| User | I live in Arkansas and love the Hot Springs National Park and have been there many times, really it is beautiful |
| DukeNet | i have heard the city is located in the ouachita mountains among the us |
| KGPT | i've been to the ouachita mountains, too! i've been to the ouachita mountains in the ouachita mountains. |
| GenKS | I've never been to Hot Springs, but I've always wanted to go to there. |
| Human | I've never been to that one! I bet its beautiful! |

Table 11: Case study on the Wizard Test Unseen dataset. This table shows an example where both models select the wrong knowledge.

B Case study

To better understand baselines and our model, we present two examples in Table 10 and Table 11. Table 10 show example where both GENKS and baselines select out the proper knowledge (i.e., the knowledge snippet shown in green). We see that the response generated by GENKS is more appropriate to the dialogue context than baselines, while KnowledGPT's response does not answer User2's question and is also factually incorrect. In Table 11, we observed that although neither GENKS nor the baselines selected the label knowledge, the response generated by GENKS is still more natural and coherence. We also find that KnowledGPT is more colloquial than GENKS but has problems with hallucinations.