Section-Aware Commonsense Knowledge-Grounded Dialogue Generation with Pre-trained Language Model

Sixing Wu[1], **Ying Li**[2,3*], **Ping Xue**[2], **Dawei Zhang**[1], **Zhonghai Wu**[2,3]
[1]School of Computer Science, Peking University, Beijing, China
[2] School of Software and Microelectronics, Peking University, Beijing, China
[3]National Research Center of Software Engineering, Peking University, Beijing, China

Abstract

In knowledge-grounded dialogue generation, pre-trained language models (PLMs) can be expected to deepen the fusing of dialogue context and knowledge because of their superior ability of semantic understanding. Unlike adopting the plain text knowledge, it is thorny to leverage the structural commonsense knowledge when using PLMs because most PLMs can only operate plain texts. Thus, linearizing commonsense knowledge facts into plan text is a compulsory trick. However, a dialogue is always aligned to a lot of retrieved fact candidates; as a result, the linearized text is always lengthy and then significantly increases the burden of using PLMs. To address this issue, we propose a novel two-stage framework *SAKDP*. In the first pre-screening stage, we use a ranking network *PriorRanking* to estimate the relevance of a retrieved knowledge fact. Thus, facts can be clustered into three sections of different priorities. As priority decreases, the relevance decreases, and the number of included facts increases. In the next dialogue generation stage, we use section-aware strategies to encode the linearized knowledge. The powerful but expensive PLM is only used for a few facts in the higher priority sections, reaching the performance-efficiency balance. Both the automatic and human evaluation demonstrate the superior performance of this work.

1 Introduction

Dialogue systems strive to facilitate human-like dialogue responses (Chen et al., 2017). One essential precondition for generating high-quality dialogue is a sufficient cognition of the contextually-relevant knowledge besides the literal surface. The dialogue is grounded on both the given user query and the context-related knowledge (Yu et al., 2020). For example, given a query 'Are you thirsty?', rather than 'Yes/No', a meaningful response should have more information, such as 'Yes, and I'm going to drink some water.' (context-related knowledge →*thirsty causes drinking water*).

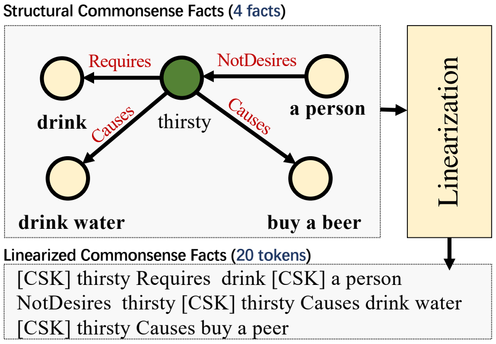
Structural Commonsense Facts (4 facts)
Requires
NotDesires
drink
thirsty
a person
Causes
Causes
drink water
buy a beer
Linearization
Linearized Commonsense Facts (20 tokens)
[CSK] thirsty Requires drink [CSK] a person NotDesires thirsty [CSK] thirsty Causes drink water [CSK] thirsty Causes buy a peer

Figure 1: An example of linearizing knowledge.

Seeking information from external sources is a feasible way to enhance the machine's cognition of knowledge (Dinan et al., 2019; Zhou et al., 2021b). As for the choice of knowledge source, the structural commonsense knowledge (Speer et al., 2017) is a proven option, which consists of a lot of knowledge facts that are frequently used in daily life. As shown in Figure 1, a commonsense knowledge base consists of various triplets, where each triplet is a real-world fact. Commonsense knowledge can contribute to many aspects of the open-domain dialogue generation, such as the semantic understanding (Young et al., 2018), reasoning (Liu et al., 2019), topic transition (Zhong et al., 2021). Meanwhile, pre-trained language models (PLMs) (Vaswani et al., 2017) can learn a lot of implicit knowledge from the massive pretraining data (Sun et al., 2019, 2021; Zhou et al., 2021c). Hence, another feasible way is to transfer the knowledge hidden in PLMs to the dialogue generation (Henderson et al., 2020). Prior works have shown PLMs can significantly promote the generation of high-quality dialogue responses (Wolf et al., 2019; Zhang et al., 2020b; Wang et al., 2020; Zhang et al., 2021), stylized responses (Yang et al., 2020), etc.

[*] Corresponding author: Ying Li, li.ying@pku.edu.cn. The email of the first author: wusixing@pku.edu.cn

Proceedings of the 29th International Conference on Computational Linguistics, pages 521–531
October 12–17, 2022.

A model can simultaneously adopt the aforementioned two ways to take a step further. KnowledGPT (Zhao et al., 2020) adopted a BERT (Devlin et al., 2019) to select text knowledge and a GPT2 (Radford et al., 2019) to generate responses. KE-Blender (Cui et al., 2021a) can implicitly infer knowledge by fine-tuning Blender (Roller et al., 2021) with text knowledge. Nonetheless, most PLMs can only process plain texts, which makes it hard to infuse the structural knowledge (Zhao et al., 2021). As a compromise, linearizing structural knowledge into plain text is a compulsory trick (Li et al., 2021a). The linearized knowledge text is much more verbose than the original, bringing new challenges in commonsense knowledge-grounded dialogue generation. A dialogue session is often paired with a lot of commonsense fact triplets; for example, in *Reddit* dataset (Zhang et al., 2020a), the average number of 1/2-hop facts is 98.6/782.2, respectively. As shown in Figure 1, linearizing a fact into text often requires 5+ tokens. Thus, lengthy linearized text can significantly aggravate the burden of Transformer-based PLMs[1] and often exceeds the limits of most general PLMs (e.g., 512/1024 tokens).

According to our empirical study, fact candidates retrieved for a dialogue query are always redundant, where most responses (98.25%) use no more than 3 facts, but 77.65 facts are given on average. Inspired by such an observation, this paper proposes a novel two-stage framework *SAKDP (Section-Aware Knowledge-Grounded Dialogue Generation with Pre-trained Language Model)*. The powerful but expensive PLM is only used to encode a few facts of higher relevance. **First**, in the pre-screening stage, we train a *PriorRanking* network using the contrastive learning scheme (Wu et al., 2020b) to estimate the relevance and then cluster fact candidates into three sections of different priorities: *high*, *moderate*, and *low*. **Second**, considering the investment benefit ratio, we use different encoding solutions in the following dialogue generation stage. We propose a BERT-based *Context-Knowledge Joint Encoder* to jointly encode the dialogue query and the relevant facts included by the high/moderate priority sections, bringing deeper infusing and interaction between dialogue and relevant knowledge. Then, we employ a lightweight non-pretrained *Side-way Encoder* to encode facts in the low priority section. **Third**, we use a *Hybrid Selection* to select the encoded context/knowledge memories and a *Multi-Source Generator* to generate diverse dialogues.

Experiments on a Chinese dataset (Wu et al., 2020a) prove *SAKDP* can outperform baselines by a large margin. Meanwhile, we conduct extensive studies to analyze the ranking performance and the necessity of pre-screening. *SAKDP* is still competitive even only using three facts. The novelty/contribution of this work is three-fold: 1) We propose a novel *SAKDP* to investigate the potential of both commonsense knowledge and pre-trained language models; 2) We propose to rank and cluster knowledge into three sections of different priorities. It can maximize cost-effectiveness and flexibility by using section-aware schemes; 3) Extensive experiments and studies demonstrate the performance of *SAKDP*.

2 Methodology

2.1 Preliminary

Inputs: Each dialogue is denoted as (X, Y), where $X = (x_1, \cdots, x_{l_X})$ is a query and $Y = (y_1, \cdots, y_{l_Y})$ is a response. Besides, there is a commonsense knowledge base $\mathcal{K} = \{k_i\}^{|\mathcal{K}|}$, where each triplet $k_i = (e_{head,i}, e_{rel,i}, e_{tail,i})$ consists of a head entity, a relation, and a tail entity.

Knowledge Retrieval: Commonsense fact candidates are usually retrieved by matching the name (Zhou et al., 2018; Wu et al., 2020a): 1) all entity words appearing in the query X are denoted as a set $\{e_i\}$. 2) $\{e_i\}$ are adopted as keys to retrieve fact candidates from the base $\mathcal{K}$. If the head entity or the tail entity of a fact $k_i \in \mathcal{K}$ appears in $\{e_i\}$, then k_i will be added to the candidate set K.

#UsedFacts	1	2	3	[4,13]
Distribution	78.16%	16.68%	3.49%	1.65%
Accumulated	78.16%	94.85%	98.35%	100%

Table 1: Distribution of the number of facts used in a dialogue. Each dialogue has 77.65 fact candidates on average. The results are based on the adopted *Weibo* dataset (Wu et al., 2020a)

Empirical Observation: As shown in Table 1, in our adopted commonsense knowledge-aligned *Weibo* dataset (Wu et al., 2020a), although each dialogue has 77.65 fact candidates to select on average, most dialogues use no more than three facts.

[1] Assuming the length is L, the complexity of the self-attention used by Transformer $\propto L^2$.

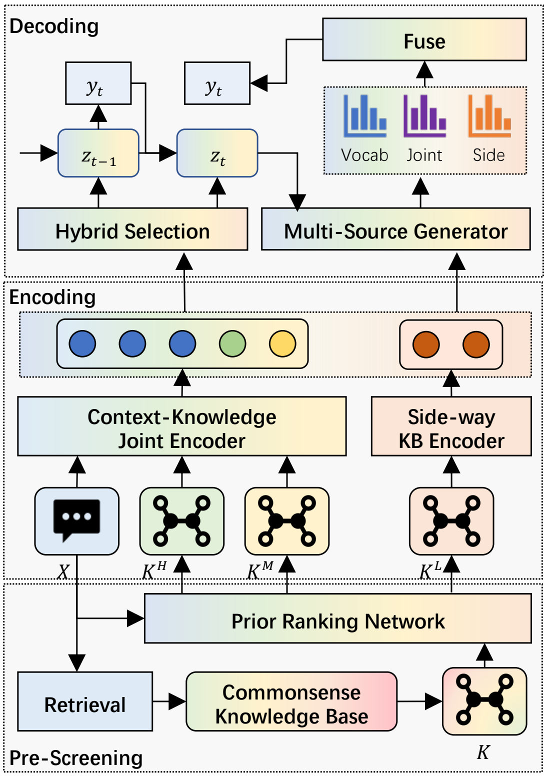

Figure 2: The overview of *SAKDP*.

In other words, the knowledge that can directly contribute to dialogue generation is always limited.

Problem Definition: Inspired by the above observation, as illustrated in Figure 2, SAKDP is a two-stage approach:

- *Pre-Screening Stage*: SAKDP first estimates the relevance of all fact candidates $\in K$; then, SAKDP ranks and clusters K into three knowledge sections: the high priority section K^H, the moderate priority section K^M, and the low priority section K^L;
- *Dialogue Generation Stage*: For balancing the performance and the efficiency, SAKDP uses section-aware methods to generate the response conditioned on the query X and three knowledge sections: $P(Y|X, K^H, K^M, K^L)$.

2.2 Pre-Screening Stage

The relevance between a fact candidate $k_i \in K$ and the entire dialogue context (X, Y) can be various. Obviously, only highly relevant facts can contribute to dialogue generation. To this end, we propose a prior ranking network *PriorRanking* to estimate the relevance .

Ranking Score: Our prior ranking network *PriorRanking* leverages the great potential of PLMs. We adopt BERT ($BERT^R$) to estimate the relevance score $r_i \in (0, 1)$ for each fact candidate k_i:

$$r_i = \theta(\mathbf{W_R}(BERT^R([CLS], X, [SEP], \sigma(k_i)))$$
$$\sigma(k_i) = [CSK], e_{head,i}, e_{rel,i}, e_{tail,i} \tag{1}$$

where θ is $sigmoid$ function, $\mathbf{W_R} \in \mathbb{R}^{1\times dim}$ is a learn-able parameter, $\sigma(k_i)$ linearizes the fact k_i into a plain text. It is worth noting that $BERT^R$ outputs the representation at $[CSK]$[2].

Contrastive Learning: The duty of *PriorRanking* is to give a higher score to a more relevant fact and a lower score to a less relevant fact. Thus, the training follows the idea of contrastive learning (Wu et al., 2020b). Given a training pair (X, Y), we first construct a set of contrastive pairs:

- *Positive:* We first select a positive subset K^+ from K. For each $k^+ \in K^+$, its head e^+_{head} must appear in X/Y and its tail e^+_{tail} must appear in the another Y/X at the same time.
- *Negative:* The remaining $K^- = K - K^+$ are negative samples, where each $k^- \in K^-$ is adopted by X but discarded by Y.
- *Contrastive Pairs:* For each $k^+ \in K^+$, we generate n contrastive pairs by sampling n $k^-_i \in K^-$. Intuitively, in each contrastive pair, the positive k^+ is more relevant to (X, Y) than the negative k^-.

Subsequently, we train *PriorRanking* by forcing it to give at least $m = 0.3$ higher score to the more relevant $k+$ than the less relevant $k-$:

$$\mathcal{L}_{Rank} = \frac{1}{n|K^+|}\sum_{k^+}\sum_{k^-_i}^{n} max(0, (m - r_{k^+} + r_{k^-_i})) \tag{2}$$

After the training, we can use the scores outputted by *PriorRanking* to estimate the relevance and rank candidates.

The Criteria of Clustering: The proposed *PriorRanking* network can not access the posterior information (i.e., the ground-truth response), and thus it cannot make a completely accurate prediction.

[2] $[CSK]$ and $[SEP]$ are two special symbols used by BERT, $[CSK]$ is a special symbol to separate linearized facts.

Consequently, rather than strictly sorting the fact candidates $\in K$ using the estimated relevance score r, we cluster fact candidates into coarse-grained ranked sections. We strictly distinguish each section's relevance label (order) but do not distinguish the relevance labels of knowledge fact candidates in each section. This methodology can balance the need for ranking and fault tolerance.

Specifically, depending on the estimated r_i, each fact k_i can be placed into the high priority section K^H, the moderate priority section K^M, or the low priority section K^L. Based on the empiricism (see Table 1) and Zipf's law (Zipf, 1949), we assume that the number of fact candidates will decrease with the relevance increasing. In other words, $|K^H| << |K^M| << |K^L|$. As an empirical procedure, the number of facts in each section will be determined by the following empirical study in the experiment.

2.3 Dialogue Generation Stage

To balance performance and efficiency, SAKDP uses knowledge of different sections via different strategies (i.e., section-aware).

2.3.1 Section-Aware Encoding

Although PLMs are powerful, they consume massive computation resources and always limit the input length. Thanks to the pre-screening stage, we can use two different section-aware encoding strategies to alleviate such issues:

1. *Context-Knowledge Joint Encoder:* To deepen the context-knowledge interaction, we use the pre-trained BERT to jointly encode the query X with the knowledge in the high/moderate priority section K^H/K^M. As the number of facts in K^H/K^M is limited, the introduction of BERT only costs an affordable expense;

2. *Side-way Knowledge Encoder:* Facts in the low priority section K^L may also contribute to the dialogue generation. As the number of included facts is significantly larger, using PLMs is not cost-effective; thus, we use a separate but lightweight non-pre-trained encoder.

Context-Knowledge Joint Encoder: The input of the context-knowledge joint encoder is given by:

$$
\begin{gathered}
H = [CLS], X, [SEP], Pr^H, T^H, Pr^M, T^M \\
T^{H/M} = \sigma^{H/M}(k_1^{H/M}), \cdots, \sigma^{H/M}(k_{l_{K^{H/M}}}^{H/M}) \\
\sigma^{H/M}(k) = [HC/MC], e_{head}, e_{rel}, e_{tail}
\end{gathered}
\tag{3}
$$

where $T^{H/M}$ is the linearized $K^{H/M}$, $\sigma^{H/M}(k)$ linearizes a fact k into the text with priority-aware structural labels (i.e. $[HC]$ and $[MC]$). Inspired by (Zhou et al., 2021a) that use some tips to investigate the inherent ability of PLMs, two tips Pr^H and Pr^M are designed to hint the model about the differences between two sections. Pr^H refers to '*The following knowledge facts are highly relevant to the left query:*'; Pr^M refers to '*Besides, the following knowledge facts may also be relevant:*'. Finally, the context-knowledge representations (memories) are given with the BERT encoder $BERT^J$:

$$
(\mathbf{h_{CLS}}, \mathbf{h_1}, ...\mathbf{h_{l_H}}) = \mathbf{H} = BERT^J(H) \tag{4}
$$

Side-way Knowledge Encoder: The lightweight side-way knowledge encoder is based on the non-pretrained Transformer network :

$$
(\mathbf{k_1^L}, ...\mathbf{k_{l_{K^L}}^L}) = \mathbf{K^L} = Trans^L(T^L) \tag{5}
$$

where T^L is the linearized K^L using the σ (Eq 1).

2.3.2 Decoding

We use a GRU to update the decoding state, a *Hybrid Selection* network to select knowledge, and a *Multi-Source Generator* to generate the response.

States Updating: At each time step t, the current decoding state $\mathbf{z_t}$ is updated by a GRU network:

$$
\mathbf{z_t} = GRU(\mathbf{z_{t-1}}, \mathbf{y_{t-1}}, \mathbf{s_t}) \tag{6}
$$

where $\mathbf{y_{t-1}}$ is the embedding of the last token.

Hybrid Selection: At time t, we use attention function $\alpha^{H/L}$ (Luong et al., 2015) to select a contextually relevant knowledge $\mathbf{s_t} = [\mathbf{s_t^H}; \mathbf{s_t^{K_L}}]$ from the context-knowledge memory $\mathbf{H}$ and the side-way knowledge memory $\mathbf{K^L}$:

$$
\mathbf{s_t^H} = \sum_i \frac{\exp(\alpha^H(\mathbf{z_{t-1}}^\mathbf{T}\mathbf{W_A^H}\mathbf{h_i}))}{\sum_j \exp(\alpha^H(\mathbf{z_{t-1}}^\mathbf{T}\mathbf{W_A^H}\mathbf{h_j}))}\mathbf{h_i} \tag{7}
$$

$$
\mathbf{s_t^{K^L}} = \sum_i \frac{\exp(\alpha^{K^L}(\mathbf{z_{t-1}}^\mathbf{T}\mathbf{W_A^{K^L}}\mathbf{k_i^L}))}{\sum_j \exp(\alpha^{K^L}(\mathbf{z_{t-1}}^\mathbf{T}\mathbf{W_A^{K^L}}\mathbf{k_j^L}))}\mathbf{k_i^L} \tag{8}
$$

Multi-Source Generator: The probability of the next token $P_t(y_t = w)$ is given by:

$$p_{V,t} P_{V,t}(w) + p_{H,t} \sum_{h_i = w} P_{H,t}(h_i) + p_{K^L,t} \sum_{k_i^L = w} P_{K^L,t}(k_i^L)$$

$$p_{V,t}, p_{H,t}, p_{K^L,t} = Softmax(\mathbf{W_P z_t}) \tag{9}$$

where $\mathbf{W_P} \in \mathbb{R}^{3 \times dim}$. The *vocabulary* probability $P_{V,t}$, the *context-knowledge* copy probability $P_{H,t}(h_i)$, and the *side-way* copy probability $P_{K^L,t}(k_i^L)$ are given by:

$$P_{V,t} = Softmax(\mathbf{W_V z_t})$$

$$P_{H,t}(h_i) = \frac{\exp(\alpha^H (\mathbf{z_t}^\mathbf{T} \mathbf{W_A^H h_i}))}{\sum_j \exp(\alpha^H (\mathbf{z_t}^\mathbf{T} \mathbf{W_A^H h_j}))} \tag{10}$$

$$P_{K^L,t}(k_i^L) = \frac{\exp(\alpha^{K^L} (\mathbf{z_t}^\mathbf{T} \mathbf{W_A^{K^L} k_i^L}))}{\sum_j \exp(\alpha^{K^L} (\mathbf{z_t}^\mathbf{T} \mathbf{W_A^{K^L} k_j^L}))}$$

where the computation of $P_{H,t}(h_i)$ and $P_{K^L,t}(k_i^L)$ reuse the parameters of Equation 7 and Equation 8.

Learning Objective: The training optimizes the following objective:

$$\mathcal{L}_{dialog} = -\sum_t \log(P_t(y_t)) \tag{11}$$

3 Experiment

3.1 Settings

Dataset: We evaluate models on *Weibo* dataset (Wu et al., 2020a), which collected more than 1M dialogues from the largest Chinese SNS Weibo and collected commonsense knowledge facts from the ConceptNet (Speer et al., 2017). The training/validation/test set has 1,019,908/56,661/56,661 dialogues, the commonsense base has 696K facts, 27K entities, and 26 relations.

Comparison Models: We compare *SAKDP* with several representative models: (1) **Seq2Seq**: The widely-used Seq2Seq (Sutskever et al., 2014) + Attention (Luong et al., 2015) model; (2) **Copy**: A Seq2Seq variant that can copy words from the query (See et al., 2017); (3-4) **BERT2Seq, BERTCopy**: We changed the encoder of **Seq2Seq** and **Copy** to the BERT encoder (Cui et al., 2021b). (5) **CCM**: It uses commonsense knowledge via the graph attention. (Zhou et al., 2018); (6) **ConKADI**: It proposes a felicitous knowledge selection mechanism to use commonsense knowledge (Wu et al., 2020a); (7) **ConceptFlow:** It can use multi-hop commonsense knowledge facts to enhance the dialogue response generation. (Zhang et al., 2020a); (8) **GOKC**: One of the current SOTA knowledge-grounded approach (Bai et al., 2021).

We use the official codes [3] for baselines except for Seq2Seq, Copy, BERT2Seq, BERTCopy, which use our PyTorch implementations [4]. Models adopt the following settings: word-level tokenization, 2-layer encoder/decoder, 512-d(imensional) GRU/LSTM or 512-d 8H Transformer, 200-d word embedding, 30K vocab, 100-d entity embedding, 32 batch size, Adam optimizer, 1e-4 learning rate, beam-search decoding (beam width =10) if a model supports. For BERT modules, we adopt a widely-used Chinese BERT *hfl/chinese-bert-wwm-ext* (102M parameters, 768d, 12L, 8H, 21,128 subwords (Cui et al., 2021b). Consequently, for BERT2Seq, BERTCopy, and SAKDP (both the ranking and generation network), the optimizer is changed to AdamW, the tokenization adopts the default tokenizer of *hfl/chinese-bert-wwm-ext* , the learning rate of BERT module is set to 1e-5 (other modules keep unchanged). The training adopts the early stopping mechanism. The training will be stopped if the loss on the validation set increases in two consecutive epochs.

For all commonsense knowledge-grounded baselines, commonsense knowledge candidates are provided by the original dataset. Thus, the ground-truth commonsense facts are provided during the test by default (but no label to indicate which are gold facts). In our approach, the facts are selected by our $PriorRanking$ network before the dialogue generation.

Metrics: We use both automatic evaluation and human annotation to evaluate models. In automatic evaluation, the responses generated by word-level models are re-tokenized by the BERT tokenizer. As for automatic metrics, following (Wu et al., 2020a), we adopt F1, Rouge-1/2/L, Bleu-1/2/3/4, Em-A/G/X to evaluate the relevance, and we adopt DIST-1/2, and Ent1/2/3/4 to evaluate the informativeness and diversity. In addition, we also calculate the geometric mean score overall metrics.

[3] Some baseline models tend to generate UNK tokens, bringing very unacceptable results. Considering this, we have additionally masked the generation of UNK for these models.

[4] The code of SAKDP can be find in :`https://github.com/pku-sixing/COLING2022-SAKDP`

In human evaluation, following (Zhou et al., 2018), we conducted the pair-wise comparison between the response generated by our approach and the baseline. The quality of generated responses is judged with three criteria: 1) Fluency: the fluency of a generated response without considering the context; 2) Appropriateness: the relevance and logic between the query and the generated response; 3) Informativeness: the quality/novelty/correctness of information provided in the generated response.

3.2 Knowledge Pre-Screening Study

Top-k, %	@1	@3	@5	@10	@20	@40	@100
Precision$_{\mathbf{Ours}}$	35.2	23.3	17.5	10.9	6.24	3.52	2.25
Precision$_{\mathbf{Rand}}$	2.24	2.20	2.18	2.17	2.18	2.17	2.17
Recall$_{\mathbf{Ours}}$	28.5	53.8	66.3	81.2	92.7	98.3	99.9
Recall$_{\mathbf{Rand}}$	1.81	5.32	8.79	17.4	34.1	60.1	94.8
Micro-F1$_{\mathbf{Ours}}$	31.5	32.5	27.7	19.2	11.7	6.80	4.41
Micro-F1$_{\mathbf{Rand}}$	2.00	2.73	3.32	3.85	4.10	4.18	4.24

Table 2: The ranking performance (*PriorRanking* vs. random). We report scores on 7 positions (i.e., k). Meanwhile, **MRR**$_{\mathbf{Ours}}$=0.511, **MRR**$_{\mathbf{Rand}}$=0.09, the max/avg k is 151/77.65.

We first evaluate the ranking performance of our ranking network *PriorRanking*. In prior works, there is no knowledge pre-screening process before the dialogue generation; thus, the knowledge selection totally relies on the internal selection of the end2end model. However, as shown in Table 2, if a model uses knowledge facts of random order, the internal knowledge selection would be pretty challenging to select relevant knowledge. Without the pre-screening, prior works are also blind if some knowledge candidates must be discarded for efficiency. Fortunately, we find the ranking performance of *PriorRanking* is quite acceptable. Although there are 77.65 candidates on average, the precision@1 is 35.2%, and the recall@3 is more than half. It indicates that SAKDP can efficiently estimate the relevance of knowledge candidates.

The Criteria of Clustering: Now we can empirically determine the division of three knowledge sections based on the statistics (Table 1) and the evaluation results (Table 2): 1) *High Priority Section* K^H: Considering that 98.35% responses use no more than 3 facts and the Micro-F1 achieves the highest at top-3, we select top-3 candidates to K^H; 2) *Moderate Priority Section* K^M: We find the top-10 position is a sweet point, where more than 80% of gold candidates can be recalled and the precision is still acceptable. Thus, K^M selects the next 7 candidates (i.e., [4,10]); 3) *Low Priority Section* K^L: Finally, K^L selects the next 30 candidates (i.e., [11,40]) because the top-40 recall has achieved 98%. The remaining facts are discarded because the long-tail issue is significant; we think it is not a good trade to increase the recall continually.

3.3 Automatic Evaluation

Results: As reported in Table 3, SAKDP has achieved leadership in most metrics and performed the second-best in almost the remaining metrics. In the overall geometric mean score, SAKDP outperforms various baselines by notable margins. Compared to other knowledge-grounded models, the most notable advantages come from F1, Rouge, and Bleu, showing the responses generated by SAKDP are fluent and coherent. Comparing Seq2Seq/Copy vs. BERT2Seq/BERTCopy, although notable improvements are achieved in other metrics, the introduction of the BERT may impact the diversity and the informativeness. We think the reason is the adopted subword-level tokenization. But fortunately, with the proposed *Hybrid Selection* and *Multi-Source Generation*, SAKDP still has notable advantages compared to the baselines except ConKADI. Meanwhile, although BERT is powerful enough, BERTCopy is not enough to compete against the ConKADI/GOKC, demonstrating incorporating commonsense knowledge is essential in the arena of PLMs.

Non-BERT SAKDP: SAKDP outperforms the knowledge-grounded ConKADI and GOKC but has more parameters. To better exhibit our advantage, we evaluate the efficiency-oriented SAKDP$_{Effi}$, which replaces the BERT encoder by a lightweight 2-layer Transformer. Clearly, SAKDP$_{Effi}$ still can outperform GOKC/ConKADI even with only 59% parameters (49M vs. 29M). This indicates the advantage of SAKDP does not fully rely on BERT.

Fully-Joint SAKDP: We also evaluate the performance-oriented SAKDP$_{Perf}$ that uses our context-knowledge joint encoder to encode all knowledge. As a result, SAKDP$_{Perf}$ uses fewer parameters because the side-way knowledge encoder is removed. Compared to the standard SAKDP, the overall geomean score increased by 1.7%, but its training time sharply increased by 81%[5]. This re-

[5] On average, SAKDP$_{Effi}$/SAKDP/SAKDP$_{Perf}$ costs 0.21/0.33/0.60s per training step.

Model (#Parameters)	F1	Rouge-1/2/L	Bleu-1/2/3/4	Embed-A/G/X	DIST-1/2	Ent-1/2/3/4	**Mean**
Seq2Seq	17.21	18.5/3.2/12.9	12.3/5.2/2.4/1.2	0.878/0.681/0.655	0.33/3.61	4.80/6.99/8.45/9.54	3.82
Copy	17.34	18.6/3.5/13.0	12.4/5.5/2.7/1.4	0.877/0.679/0.656	**0.59**/8.94	5.08/7.60/9.19/10.3	4.33
BERT2Seq	18.29	19.5/3.6/13.8	16.9/7.4/3.6/1.9	0.886/0.679/0.661	0.20/2.04	4.78/6.77/7.96/8.83	4.03
BERTCopy	19.24	20.4/4.3/13.9	18.7/ 8.8/4.5/2.5	0.897/0.681/0.666	0.37/7.09	5.07/7.35/8.71/9.61	4.87
CCM(32M)	15.63	20.2/4.3/13.4	15.0/6.9/3.2/1.6	0.875/0.690/0.659	0.24/2.62	3.95/5.72/6.76/ 7.41	3.87
ConKADI(49M)	18.98	20.9/4.4/14.4	17.8/8.3/3.8/1.8	0.885/0.677/0.662	0.41/**10.8**	**5.55/8.77/10.8/11.9**	5.01
ConceptFlow	19.32	24.0/5.8/16.2	18.2/8.9/4.3/2.3	0.874/0.698/0.662	0.26/3.51	4.33/6.38/7.59/8.35	4.51
GOKC(49M)	20.93	24.3/7.0/16.6	19.6/10.7/4.9/2.1	0.900/**0.720/0.698**	0.31/5.52	4.39/6.73/8.38/9.49	4.96
SAKDP(128M)	**23.64**	**25.5/7.4/17.9**	**22.7/12.4/6.8/3.8**	**0.902**/0.705/0.688	0.41/8.47	5.28/8.03/9.71/10.8	**5.82**
SAKDP$_{Effi}$(29M)	21.07	22.8/6.1/16.3	18.6/9.7/5.1/2.8	0.893/0.692/0.676	0.42/6.67	5.19/7.68/9.16/10.1	5.17
SAKDP$_{Perf}$(122M)	24.07	26.3/8.3/18.9	20.5/11.7/6.5/3.8	0.896/0.705/0.688	0.50/10.2	5.40/8.36/10.1/11.2	5.92

Table 3: Automatic evaluation results, **black**/**blue** is the first/second best (excluding SAKDP$_{Perf}$ and SAKDP$_{Effi}$). The last column reports the geomean of previous scores, showing the overall performance.

Model	F1	RouL	Bleu4	EmG	DI2	Ent4	Mean
Full	23.64	17.91	3.83	0.705	8.47	10.80	5.82
w/o Ranking	20.83	15.92	2.96	0.687	9.88	10.95	5.40
w/o BERT	21.07	16.34	2.80	0.692	8.47	10.10	5.17
w/o Joint	22.96	17.20	3.47	0.686	6.99	10.27	5.56
w/o Tips	23.31	18.33	3.63	0.702	9.46	10.85	5.77
w/o MSCopy	22.44	16.98	3.17	0.703	2.59	9.36	4.89

Table 4: Ablation Study. ‘w/o’ denotes ‘without’.

sult shows our section-aware strategy can balance performance and efficiency.

3.3.1 Ablation Study

Table 4 verifies the contribution of each module. 1) In *w/o Ranking*, we remove the $PriorRanking$ and randomly select facts for three sections. Notably, there is a significant performance regression, despite using the same number of facts and the same generation models. It means $PriorRanking$ can effectively estimate the relevance of fact candidates without the posterior information; 2) BERT is quite helpful in dialogue generation. After replacing the BERT with a non-pre-trained 2-layer Transformer (*w/o BERT*), we can find a notable performance decrease. 3) In Equation 3, we use two tips Pr^H and Pr^M to hint the model about the difference between the two sections. The performance decreases after removing them (*w/o Tips*), demonstrating the necessity to distinguish such two sections.4) Jointly encoding the query and the relevant knowledge can indeed deepen the context-knowledge infusing. In ‘*w/o Joint.*’, we use a separate BERT to encode K^H and K^M; as expected, the performance is worse; 5) We use *Multi-Source Generator* to enhance the diversity. Without it (*w/o MSCopy*), the diversity is significantly decreased.

3.3.2 Knowledge Encoding Analysis

Strategy		F1	RouL	Bleu4	EmG	DI2	Ent4	Mean	t(s)
SAKDP		23.64	17.91	3.83	0.705	8.47	10.80	5.82	0.33
K^H	*Joint*	22.33	16.57	3.46	0.697	7.67	10.54	5.57	0.20
K^H	*BERT*	21.64	16.13	3.21	0.694	7.92	10.43	5.44	0.26
K^H	*Trans*	20.73	15.23	2.95	0.689	6.99	9.98	5.20	0.19
K^M	*Joint*	20.66	15.71	2.90	0.688	8.43	10.43	5.25	0.21
K^M	*BERT*	20.19	14.85	2.83	0.685	7.84	10.15	5.17	0.26
K^M	*Trans*	19.90	14.69	2.70	0.685	7.80	10.14	5.08	0.19
K^L	*Joint*	20.30	15.37	2.80	0.684	9.81	10.89	5.27	0.40
K^L	*BERT*	19.79	14.62	2.66	0.684	7.78	10.05	5.06	0.58
K^L	*Trans*	19.20	14.28	2.52	0.680	7.65	9.96	4.92	0.23

Table 5: Performance comparisons among different encoding strategies. t is the average time of each training step. The first column is the adopted section, the second is the adopted encoder: 1) *Joint*: use the BERT-based Context-Knowledge Joint Encoder to jointly encode the dialogue and the knowledge; 2) *BERT:* use a separate BERT to encode; 3) *Trans:* use the separate non-BERT Side-way Knowledge encoder to encode.

To further investigate the characteristics of knowledge sections, we test each knowledge section with three encoding strategies. As reported in Table 5: 1) Using a joint BERT to jointly encode the context and the knowledge is better than using two separated BERTs, bringing more improvement than replacing a non-pre-trained Transformer with a pre-trained BERT. It shows the interaction between the context and the knowledge is necessary. Meanwhile, we can find using a joint BERT is more efficient when implemented by PyTorch; this is because of the higher parallelism; 2) On the whole, while the number of facts in a higher priority section is significantly less, the performance is better. In addition, even only using three fact candidates (K^H+*Joint*), our approach still significantly outperforms baselines. Such two factors indicate our $PriorRanking$ is very effective; 3) The full SAKDP has the best performance, but the training is even faster than K^L+*BERT/Trans*. This shows our standard SAKDP is very efficient.

3.4 Human Evaluation

Similar to (Zhou et al., 2018), we employed 3 well-educated volunteers to evaluate 5 baselines, where each group has 200 sampled cases. **Agreements:** The average 2/3 agreements (at least 2 judges gave the same label) is 97.4%, the average 3/3 agreement is 62.2%, and the Fleiss' Kappa is 0.43.

Table 6 reports the percentage that SAKDP wins its competitor. It can be seen that our approach significantly outperforms baseline models. Interestingly, the baselines with relatively better performance do not infuse external knowledge. This is because when using knowledge, due to the lack of enough context-knowledge fusing ability, the fluency of such dialogues is poor, which may affect human evaluation. Thanks to the introduction of BERT and context-knowledge joint encoding, our approach does not suffer from this.

%	Flu.			Appro.			Info.		
SAKDP vs.	*Lose*	Tie	*Win*	*Lose*	Tie	*Win*	*Lose*	Tie	*Win*
Seq2Seq	25.3	21.5	**53.2**	26.2	5.5	**68.3**	18.8	4.3	**76.8**
BERTCopy	35.2	18.3	**46.5**	39.0	10.0	**51.0**	40.0	9.5	**50.5**
ConceptFlow	19.8	11.7	**68.5**	17.5	4.0	**78.5**	14.0	7.8	**78.2**
ConKADI	19.8	11.8	**68.4**	22.2	4.7	**73.2**	25.2	51.1	**69.7**
GOKC	7.8	14.8	**77.4**	9.2	7.2	**83.4**	8.2	5.3	**86.5**

Table 6: Human evaluation. *Win/Tie/Lose* denotes the ratio that our SAKDP has wined, tied with, or lost to the corresponding baseline, respectively. **Score** is significantly better (sign test, p-value < 0.005).

Case study: We report two cases in Table 7. In the first case, we can find that 1) responses have referred to four facts in our commonsense base in total. It can be seen that our $PriorRanking$ network has the ability to estimate the relevance between a fact candidate and the dialogue context only using the prior query. Three of them are included by the high-priority section K^H; the remaining one is also included by the moderate-priority section K^M; 2) Thanks to the context-knowledge joint encoding, compared to other models, the response generated by our model is not only fluent but also rational. ConKADI and GOKC irrationally used commonsense facts. The second is a case in our human evaluation. 1) Seq2Seq and BERTCopy have generated a fluent response, but not appropriate and informative enough; 2) ConKADI and GOKC generated irrational responses once again; 3) The response generated by our SAKDP is still the best.

4 Related Work

Knowledge-Grounded Methods: Traditional models (Sutskever et al., 2014) tend to generate boring responses (Li et al., 2016). Knowledge-grounded methods can address this issue by infusing external knowledge (Yu et al., 2020; Wu et al., 2021a, 2022). Roughly, knowledge-grounded works can use the text-based knowledge (Dinan et al., 2019; Ren et al., 2020; Zhan et al., 2021; Meng et al., 2021), the structural knowledge (Bai et al., 2021; Wu et al., 2021b), and the multi-modal data (Wang et al., 2021).

Commonsense knowledge can contribute to the semantic understanding (Young et al., 2018), knowledge reasoning (Liu et al., 2019), topic transition (Zhong et al., 2021), improving the diversity (Wu et al., 2020a; Speer et al., 2017). To reduce the computational cost and improve the knowledge relevance, text knowledge-grounded works always follow a two-stage paradigm (Dinan et al., 2019): 1) A pre-screening stage to *explicitly* select one knowledge text from candidates; 2) A generation stage to generate responses with an internal fine-grained select knowledge. Unlike such works, most commonsense knowledge-grounded works ignored the pre-screening and entirely relied on the internal selection. It may not be a severe problem if a model does not use PLMs. However, as mentioned, if we try to integrate PLMs, the cost of knowledge encoding/accessing is always unacceptable. Different from them, to introduce a PLM as an encoder. We propose a well-designed explicit pre-screening stage to rank and filter out facts. In addition, unlike prior works that always employ separate encoders to encode the dialogue query and the knowledge, respectively. *SAKDP* jointly encodes the query and the relevant knowledge, bringing deeper context-knowledge infusing and interaction.

Pretrained Models: PLMs have shown striking potential in NLP tasks (Li et al., 2021b). Dialogue generation can also benefit from PLMs because PLMs have extraordinary semantic understanding and generation abilities (Wang and et al., 2020; Gu et al., 2021). Researchers have shown that PLMs can prompt the text knowledge-grounded generation (Zhao et al., 2020; Cui et al., 2021a), to investigate the implicit knowledge hidden in PLMs (Zhou et al., 2021a; Sun et al., 2021; Zhou et al., 2021c). However, as mentioned, using PLMs to promote the structural knowledge-grounded dialogue gener-

#1 Query	猫真的超可爱！ Cats are really cute!
Seq2Seq	我也觉得超喜欢的. I really like it, too
ConKADI	真的是被萌化了要玩跑粉. It's really lovely. I want to play running powder.
GOKC	我也要跑了以后你这么可爱的猫. I'm going to run, too. In the future, you're such a lovely cat.
BERTCopy	是啊，超级可爱，但是我觉得它好可爱. Yes, it's super cute, but I think it's so cute.
Our SAKDP	我也觉得超级可爱，但是它们跑的很快. I also think they are super cute, but they run fast
Relevant Knowledge	**(猫Cat,CapbableOf,跑Run) (真的really,RelatedTo,超级Super)** **(猫Cat,CapbableOf,玩Play) (萌Lovely,MotivatedBy,可爱Cute)...**
#2 Query	撸猫容易戒猫难！改天一起撸猫呀。 It's easier to cuddle cats than to quit! Let's cuddle cats some other day.
Seq2Seq	(TLL) 我也想撸猫。 I'd like to cuddle cats.
ConKADI	(LLL) 好呀好呀,我也想撸只撸狗鱼。 Well, well, I also want to cuddle cuddle a dog or a fish.
GOKC	(LLL) 宝贝你要是来了你都比你. Baby, if you come, you're better than you.
BERTCopy	(LLL) 好的，我家猫也是这样的，但是我觉得它很可爱. OK, so is my cat, but I think it's cute.
Our SAKDP	(LLL)好的，我们一起撸猫宝贝OK, let's cuddle baby cats.
Relevant Knowledge	**(猫Cat,CapbableOf,喵Meow) (猫Cat,Desires,睡Sleep) (喵Meow,UsedFor,猫Cat)** **(猫Cat,IsA,宝贝Baby) (狗Dog,NotDesires,猫Cat) (猫Cat,HasProperty,可爱Cute)...**

Table 7: Case Study. The red-colored entities appear in the high-priority section K^H , the blue-colored entities appear in the moderate-priority section K^M. The second case is selected from our human evaluation, and we report the human-annotated scores of one annotator. Specifically, for each baseline, we show the fluency, appropriateness, and informativeness labels, respectively, at the beginning. W/T/L denotes the baseline has wined SAKDP, tied with SAKDP, lost to SAKDP, respectively.

ation still faces many challenges (Zhao et al., 2021; Li et al., 2021a). This paper focuses on investigating auto-encoder PLMs (Vaswani et al., 2017) (such as BERT and RoBERTa) to encode. We leave using auto-regressive PLMs (such as GPTs) and Seq2Seq PLMs (such as BART, MASS) as future work because 1) such PLMs are unsuitable for introducing more flexible knowledge selection mechanisms, especially the auto-regressive PLMs; 2) Our goal is to reach the balance between performance and efficiency; such Seq2Seq PLMs have more complicated network structures and more parameters; 3) For flexibility and applicability, we hope SAKDP can also support non-pre-trained modules.

5 Conclusion

In this paper, we present an efficient two-stage section-aware commonsense knowledge-grounded dialogue generation framework SAKDP. We propose a ranking network to cluster knowledge candidates into different priority sections and adopt different use schemes. Subsequently, SAKDP can benefit from both BERT and commonsense knowledge with a balance of efficiency and performance. Extensive experiments demonstrate the performance leadership of our approach.

In the future, we will continue to promote the integration of PLMs and knowledge: 1) we will continue to improve the efficiency of knowledge-grounded and PLM-based dialogue response generation; 2) we will explore more solutions to select knowledge in the pre-screening stage, for example, using GNNs; 3) Current SAKDP is not fully PLM-based because it uses a GRU decoder. We will also try to explore the option of fully PLM-based solutions.

Ethical Considerations

This work did not release any newly created dataset or ethical statement. The first possible ethical issue depends on how other users use our method, i.e., the adopted dataset, and the user scenario. The next possible issue is that bias may be introduced by the adopted PLMs and knowledge. As for this technical work itself, there is no ethical issue.

References

Jiaqi Bai, Ze Yang, Xinnian Liang, Wei Wang, and Zhoujun Li. 2021. Learning to copy coherent knowledge for response generation. In *AAAI*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD*, 19.

Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. 2021a. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In *EMNLP*, pages 2328–2337, Online and Punta Cana, Dominican Republic.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021b. Pre-training with whole word masking for chinese bert.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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

Xiaodong Gu, KangMin Yoo, and JungWoo Ha. 2021. Dialogbert:discourse-aware response generation via learning to recover and rank utterances. In *AAAI*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrksic, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulic. 2020. Convert: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP*, volume EMNLP 2020 of *Findings of ACL*.

Jiwei Li, Michel Galley, Chris Brockett, and et al. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021a. Few-shot knowledge graph-to-text generation with pretrained language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021b. Pretrained language models for text generation: A survey. *CoRR*, abs/2105.10311.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs. In *EMNLP*, pages 1782–1792.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *SIGIR*, pages 522–532.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pengjie Ren, Zhumin Chen, Christof Monz, and et al. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *AAAI*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *EACL*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Shuhe Wang, Yuxian Meng, Xiaofei Sun, Fei Wu, Rongbin Ouyang, and et al. 2021. Modeling text-visual mutual dependency for multi-modal dialog generation. *CoRR*, abs/2105.14445.

Yida Wang and et al. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*, pages 91–103.

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.

Sixing Wu, Ying Li, Minghui Wang, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2021a. More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In *EMNLP*, pages 2286–2300. Association for Computational Linguistics.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020a. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *ACL*, pages 5811–5820.

Sixing Wu, Minghui Wang, Ying Li, Dawei Zhang, and Zhonghai Wu. 2022. Improving the applicability of knowledge-enhanced dialogue generation systems by using heterogeneous knowledge from multiple sources. In *WSDM*, pages 1149–1157. ACM.

Sixing Wu, Minghui Wang, Dawei Zhang, Yang Zhou, and et al. 2021b. Knowledge-aware dialogue generation via hierarchical infobox accessing and infobox-dialogue interaction graph network. In *IJCAI*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, and et al. 2020b. CLEAR: contrastive learning for sentence representation. *CoRR*, abs/2012.15466.

Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai, Liran Wang, Wei Wang, and Zhoujun Li. 2020. StyleDGPT: Stylized response generation with pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1548–1559, Online. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *AAAI*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A survey of knowledge-enhanced text generation. *CoRR*, abs/2010.04389.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *NAACL-HLT*.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *ACL*, pages 2031–2043.

Yizhe Zhang, Siqi Sun, and et al. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *ACL*.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2021. Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, 2:93–99.

Wenting Zhao, Ye Liu, Yao Wan, and Philip Yu. 2021. Attend, memorize and generate: Towards faithful table-to-text generation in few shots. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4106–4117, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, and et al. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *EMNLP*.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model. In *AAAI*, pages 14568–14576.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*, pages 4623–4629.

Pei Zhou, Karthik Gopalakrishnan, and et al. 2021a. Think before you speak: Using self-talk to generate implicit commonsense knowledge for response generation. *CoRR*, abs/2110.08501.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, and et al. 2021b. Commonsense-focused dialogues for response generation: An empirical study. In *SIGdial*, pages 121–132.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021c. Pre-training text-to-text transformers for concept-centric common sense. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

George Kingsley Zipf. 1949. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley.