Improving Open-Domain Dialogue Response Generation with Multi-Source Multilingual Commonsense Knowledge

Sixing Wu[1,2], **Jiong Yu**[1,2], **Jiahao Chen**[1,2], **Xiaofan Deng**[1,2], **Wei Zhou**[1,2*]

[1]National Pilot School of Software, Yunnan University, Kunming, China
[2]Engineering Research Center of Cyberspace, Yunnan University, Kunming, China
{wusixing, zwei}@ynu.edu.cn

Abstract

Knowledge-grounded Dialogue Response Generation (KRG) can facilitate informative and fidelity dialogues using external knowledge. Prior monolingual works can only use the knowledge of the corresponding native language. Thus, due to the prohibitive costs of collecting and constructing external knowledge bases, the limited scale of accessible external knowledge always constrains the ability of KRG, especially in low-resource language scenarios. To this end, we propose a new task, *Multi-Source Multilingual Knowledge-Grounded Response Generation (MMKRG)*, which simultaneously uses multiple knowledge sources of different languages. We notice that simply combining knowledge of different languages is inefficient due to the *Cross-Conflict* issue and *Cross-Repetition* issue. Thus, we propose a novel approach *MMK-BART*, which uses a simple but elegant *Estimate-Cluster-Penalize* mechanism to overcome the mentioned issues and adopts the multilingual language model mBART as the backbone. Meanwhile, based on the recent multilingual corpus *XDailyDialog*, we propose an MMKRG dataset *MMK-DailyDialog*, which has been aligned to the large-scale multilingual commonsense knowledge base ConceptNet and supports four languages (English, Chinese, German, and Italian). Extensive experiments have verified the effectiveness of our dataset and approach in monolingual, cross-lingual, and multilingual scenarios.

Introduction

Open-domain dialogue systems allow users to initiate conversations on any topic of their choice (Sutskever, Vinyals, and Le 2014; Yan 2018). Nevertheless, current dialogue response generation (RG) approaches are susceptible to generating uninformative responses (Li et al. 2016) and hallucinatory information (Shuster et al. 2021). The aforementioned issues can be partly attributed to the inadequate knowledge acquired by the model (Yu et al. 2020). Therefore, knowledge-grounded response generation (KRG) approaches have emerged from recent studies as they are capable of retrieving informative knowledge from external sources and substantially improving the generation quality of dialogue responses (Zhou et al. 2018; Qin et al. 2019; Zhang et al. 2020).

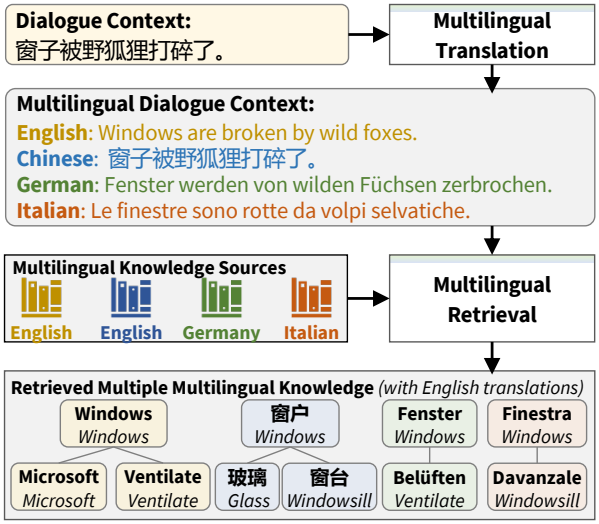

Figure 1: An example of multilingual knowledge retrieval. *Cross-Repetition issue:* it can be the repetition *(Windows → Ventilate/Windowsill)* across languages. *Cross-Conflict issue: (Windows → Microsoft)* is conflicted with the context.

The scale of accessible knowledge is crucial for KRG to handle complex topics (Wu et al. 2022). When there is no available knowledge to retrieve from the given knowledge base, KRG will fall back into traditional RG. However, most prior works are monolingual, and the available knowledge is often limited to a single native language source (Kann et al. 2022), which presents three challenges: *1) Limited Knowledge Coverage:* A single knowledge source suffers from insufficient knowledge coverage. Many conversation contexts may not match any knowledge in practice; *2) Low-resource Dilemma:* There are significant differences in the abundance of knowledge resources across different languages. For example, there are many knowledge resources in high-resource languages such as Chinese and English but only very few bases in low-resource languages such as Indonesian and Icelandic. Consequently, prior monolingual KRG works are challenging to deploy in many low-resource scenarios; *3) Transfer Barrier:* Even if efforts are invested in transferring knowledge from one language to another to address the

*The corresponding author.

above issues, it remains difficult and expensive due to language and cultural differences.

To alleviate these issues, this work proposes to use multiple multilingual knowledge sources (MMKS) simultaneously to ground open-domain dialogue response generation, i.e., MMKRG. Nonetheless, reality is often not as beautiful as the ideal. As illustrated in Figure 1, we find the paradigm of simply fusing MMKS is problematic due to the following issues: *1) Cross-Conflict:* Conflicts (i.e., retrieved facts are irrelevant to the context) between the native language dialogue context and the cross-lingual knowledge are more notable than in the traditional monolingual scenario; *2) Cross-Repetition:* Different sources may provide knowledge of the similar meaning but different languages. Thus, the retrieved knowledge items may be redundant and verbose, causing the backbone model to become hesitant and inefficient in knowledge selection; *3) Corpus Shortage:* To our knowledge, there is no dataset for MMKRG currently.

Considering such challenges, we propose an MMKRG dataset *MMK-DailyDialog* and an MMKRG approach *MMK-BART*. To our knowledge, both the corpus and approach are the first of their kind. Specifically, we first construct our corpus *MMK-DailyDialog* by aligning the large-scale multi-lingual commonsense knowledge base ConceptNet (Speer, Chin, and Havasi 2016) to a recent multilingual conversational corpus *XDailyDialog* (Liu et al. 2023). *XDailyDialog* extends the monolingual English *DailyDialog* (Li et al. 2017) into four languages, including *English (En)*, *Chinese (Zh)*, *German (De)* and *Italian (It)*, where all four languages are fully covered by the ConceptNet.Regarding methodology, we propose a *Estimate-Cluster-Penalize* mechanism for *MMK-BART* to reduce the redundant and conflict items across different languages. The first *Estimate* step estimates the possible relevance between the fact candidate and the dialogue context via learnable scoring functions. The next *Cluster* step tries to cluster commonsense knowledge facts across languages of a similar meaning into a cluster with the unsupervised K-Means algorithm. The last *Penalize* step penalizes the estimated relevance scores via the in-cluster rank label (alleviating the *Cross-Repetition* issue) and then only keeps the facts with high penalized scores (alleviating the *Cross-Conflict* issue). Finally, we take the multilingual Seq2Seq mBART (Liu et al. 2020) as the backbone language model, which has memorized massive multilingual knowledge in the pre-training.

Extensive experiments have been conducted to evaluate the effectiveness of our *MMK-DailyDialog* dataset and our *MMK-BART*. In the first monolingual evaluation, we show that our dataset can take advantage of the commonsense knowledge graph *ConceptNet*, bringing notable performance improvement in response generation. The next multilingual evaluation proves that 1) simply using knowledge facts retrieved from multiple multilingual sources may even result in worse results due to the *Cross-Repetition* issue and *Cross-Conflict* issue, and 2) our *MMK-BART* and *Estimate-Cluster-Penalize* can effectively address such two issues. We also conduct a cross-lingual evaluation to explore the potential in the zero-native-knowledge scenario.

The contribution of this work is four-fold: 1) To our knowledge, we are the first to explore using multi-source multilingual knowledge in the context of KRG; 2) We collected and constructed the first MMKRG dataset *MMK-DailyDialog*; 3) The proposed *MMK-BART* can effectively use multi-source multilingual knowledge with the novel *Estimate-Cluster-Penalize* mechanism. 4) Extensive experiments verified the effectiveness of our dataset and approach.

Related Work

Knowledge-Grounded Dialogue Generation Compared with the traditional dialogue response generation (RG) (Sutskever, Vinyals, and Le 2014), knowledge-grounded dialogue response generation (KRG) (Yu et al. 2020) uses external knowledge to alleviate the issue of generating non-informative dialogues (Li et al. 2016) and hallucination information (Shuster et al. 2021). Depending on the choice of knowledge base, previous KRG works can be text-based (Lin et al. 2020a; Kim et al. 2021), graph-based (Wu et al. 2020; Zhou et al. 2022), and many others (Moghe et al. 2020; Li et al. 2022). Nonetheless, most prior works only use one monolingual knowledge source, which always suffers from the insufficient knowledge issue and limits the performance in low-resource languages (Kann et al. 2022). Some works try to retrieve knowledge from multiple sources (Liang et al. 2021; Wu et al. 2021, 2022) or use multiple types of external information (Jang et al. 2022) to enrich the knowledge. However, such works are still focusing on the monolingual scenario. Compared to such works, this work uses knowledge of multiple languages simultaneously.

Multilingual Dialogue Generation With the development of multilingual language models such as mBART (Liu et al. 2020) and mT5 (Xue et al. 2021), much attention has been paid to multilingual applications. However, only a few RG works are multilingual. MulZDG (Liu et al. 2022) and XDailyDialog (Liu et al. 2023) studies the multilingual response generation, X-Persona (Lin et al. 2020b) learns to generate the multilingual personalized dialogues, and ToD (Majewska et al. 2023) learns the multilingual task-oriented dialogue systems. However, such works do not access external knowledge. Unlike them, KoWoW (Kim et al. 2021) can leverage English and Korean text-based knowledge in one model. Unfortunately, this work is limited to only two languages. Compared to KoWoW, this work can use multiple multilingual knowledge sources simultaneously.

Cross-lingual Learning How to use cross-lingual knowledge is crucial in multilingual learning. To address this issue, there are three typical paradigms (Huang, Yu, and Allan 2023): The first two are 1) translating the cross-lingual resource into the current language before feeding them into the model and 2) translating the output of the model. The last 3) does not explicitly conduct any translation but uses end-to-end training. The construction of *MMK-DailyDialog* follows the first paradigm. The next *MMK-BART* needs to manipulate inputs/outputs of several languages simultaneously; thus, it uses the last paradigm.

	English (En)	*Chinese (Zh)*	*German (De)*	*Italian (It)*
#Training	10.5K Sessions and 39.7K Dialogues			
#Valid/Test	995/996 Sessions and 3.83K/3.69K Dialogues			
#Entities	1.17M	0.13M	0.52M	0.51M
#Relations	47	25	11	16
#Facts	3.28M	0.37M	1.00M	0.58M

Table 1: The statistics of our *MMK-DailyDialog* .

MMK-DailyDialog

Background

There are only three existing multilingual RG corpora. The first non-knowledge-grounded *XPersona* (Lin et al. 2020b) translates the English *Persona-Chat* (Zhang et al. 2018) into seven languages and faces two crucial issues: 1) non-English dialogues are not fluent because it lacks enough human corrections, and 2) this dataset is not fully parallel, and thus we are hard to study the latent relationship among languages. The next non-knowledge-grounded dataset is the recent *XDailyDialog* (Liu et al. 2023). Although it is still translated from the English *DailyDialog* (Li et al. 2017) and only supports four languages, it is fully parallel and with significantly higher quality because of a more careful human correction process. The last is knowledge-grounded KoWoW (Kim et al. 2021). Nonetheless, it only supports English and Korean and uses text-based knowledge rather than graph-based commonsense knowledge. Thus, this work presents a multi-source multilingual KRG dataset *MMK-DailyDialog*, by extending the *XDailyDialog*.

Conversations *MMK-DailyDialog* keeps all conversations of *XDailyDialog*. Thus, as reported in Table 1, there are four languages and each has about 13K sessions and about 50K dialogues. Dialogues are fully parallel among four languages; namely, one dialogue has four language versions.

Commonsense Knowledge We choose the large-scale multilingual commonsense knowledge base *ConceptNet* (Speer, Chin, and Havasi 2016). We regard knowledge facts of **each language** as an **independent** knowledge source. As shown in Table 1, we can find the knowledge abundance of each language is various, where English facts are the most.

Single-Source Alignment

Given a dialogue session $(H^{l_{src}}, R^{l_{src}})$ of a language l_{src}, where $H^{l_{src}}$ is the dialogue context (history) and $R^{l_{src}}$ is the response, the task is to retrieve an aligned commonsense knowledge fact set $K^{l_{src}} = \{k_i^{l_{src}}\}$ from the corresponding knowledge source $\mathcal{G}^{l_{src}}$ of the same language.

We design $Retrieve(\cdot)$ to retrieve coarse-grained facts for the given utterance. $Retrieve(\cdot)$ first tokenizes and lemmatizes the given utterance into an n-gram list with NLTK, and all stopwords are removed. Then, for a fact $k_i^{l_{src}} = (e_{head}^{l_{src}}, e_{rel}^{l_{src}}, e_{tail}^{l_{src}}) \in \mathcal{G}^{l_{src}}$, if the head entity $e_{head}^{l_{src}}$ or the tail entity $e_{head}^{l_{tail}}$ appears in the filtered n-gram list, this fact $k_i^{l_{src}}$ will be retrieved. Then, we use SentenceBERT (Reimers and Gurevych 2020) to compute the embedding and filter out the facts retrieved by $Retrieve(H^{l_{src}})$. Specifically, for each $k_i^{l_{src}} \in Retrieve(H^{l_{src}})$, we estimate its prior relevance and the posterior relevance by:

$$s_{prior,i}^{l_{src}} = Cosine(SBERT(k_i^{l_{src}}), SBERT(H^{l_{src}}))$$

$$s_{post,i}^{l_{src}} = Cosine(SBERT(k_i^{l_{src}}), SBERT(R^{l_{src}}))$$

where the prior score $s_{prior,i}$ considers the cosine similarity between the candidate fact and the dialogue history, and the posterior s_{post} considers the similarity between the candidate fact and the ground-truth response.

Finally, we first use the posterior scores s_{post} to select the top 20 most relevant facts. Then, to avoid the impacts of the posterior information during the training, we use the prior scores s_{prior} to re-order the obtained 20 most relevant facts and obtain $K^{l_{src}}$. Thus, the KRG models still need to select the knowledge from $K^{l_{src}}$, acting more like the practical scenario. Meanwhile, we also construct a golden set $K_{Gold}^{l_{src}} = Retrieve(H^{l_{src}}) \cap Retrieve(R^{l_{src}})$.

Multi-Source Multilingual Alignment

To retrieve knowledge of other languages, we assume there is a *Multilingual Translation System* $MTS(\cdot)$, which can translate $(H^{l_{src}}, R^{l_{src}})$ to the other language version $(H^{l_{tgt}}, R^{l_{tgt}})$; then, we can get the corresponding $K^{l_{tgt}}$ as discussed in previous Section . Thus, for each dialogue session $(H^{l_{src}}, R^{l_{src}})$, the corresponding multi-source multilingual knowledge set K^M can be given by:

$$K^M = K^{l_{src}} \cup K^{l_{tgt_1}} \cup K^{l_{tgt_2}} \cup \cdots \quad (1)$$

Methodology

Problem Formulation and Overview

Suppose the constructed *MMK-DailyDialog* corpus is $\mathcal{D} = \{(H^{l_{src}}, R^{l_{src}}, K^M)\}$ that covers languages $\mathcal{L} = \{l_*\}$. Thus, our task is defined as $P(R^{l_{src}}|H^{l_{src}}, K^M)$.

As shown in Figure2, this work proposes a *MMK-BART* approach, which uses a novel *Estimate-Cluster-Penalize* mechanism to leverage K^M in a more efficient way.

Estimate-Cluster-Penalize

Empirically, we find blindly use multi-source multilingual knowledge set K^M suffers from two issues:

- *Cross-Conflict:* Due to the inherent drawbacks of the retrieval process, K^M often involves many contextually irrelevant facts that can impact the dialogue generation.
- *Cross-Repetition:* Some facts that share a very similar meaning but in different languages may be retrieved at the same time. For example, as shown in Figure 1, the English fact *(Windows, UsedFor, Ventilate)* and the German fact *(Fenster, UsedFor, Belüften)*. Thus, many retrieved facts will be redundant with the use of multilingual knowledge sources, making the dialogue generation model hesitant and inefficient in knowledge selection.

To this end, this work proposes a novel and elegant three-stage mechanism *Estimate-Cluster-Penalize* to remove both redundant and irrelevant knowledge and optimize the arrangement of knowledge facts.

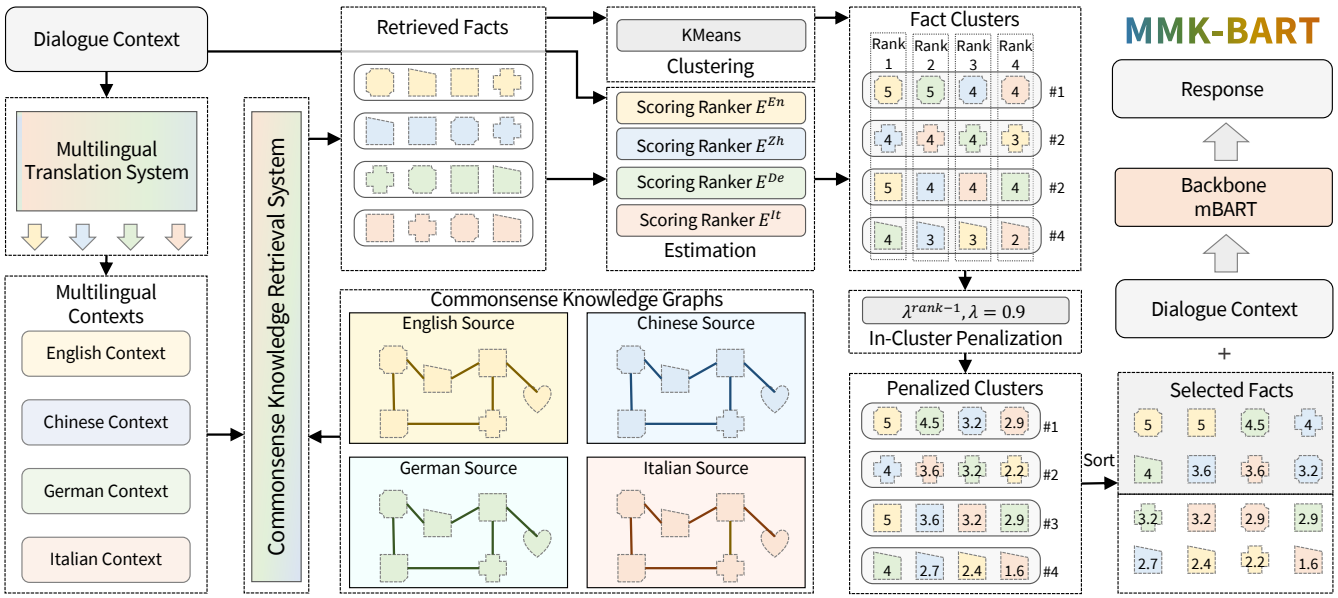

Figure 2: An illustration of the proposed *MMK-BART*. Different colors indicate different languages. Different small shapes indicate facts of different meanings.

Estimating Relevance To identify the possible conflict between the dialogue context and each knowledge fact, we first train a language-specific scoring function $s_i^{l_{src}} = E^{l_{src}}(H^{l_{src}}, k_i^{l_{src}})$ for each language l_{src} to predict the relevance score $s_i^{l_{src}}$ between the target response $R^{l_{src}}$ and the knowledge fact $k_i^{l_{src}}$ based on the dialogue history $H^{l_{src}}$:

$$E^{l_{src}}(H^{l_{src}}, k_i^{l_{src}}) = \theta\left(\mathbf{W_E^{l_{src}}} SBERT^{l_{src}}(I_i^{l_{src}})\right) \quad (2)$$
$$I_i^{l_{src}} = [CLS], H^{l_{src}}, [SEP], \delta(k_i^{l_{src}}), [SEP]$$

where θ is $Sigmoid$, $\mathbf{W_E^{l_{src}}}$ is a linear layer, $SBERT^{l_{src}}$ is the pre-trained multilingual SentenceBERT Cross-Encoder (Urbanek et al. 2019), $\delta(\cdot)$ linearizes a fact into plain text.

Then, we adopt the contrastive learning paradigm to train each $E^{l_{src}}$. We let the corresponding golden set $K_{Gold}^{l_{src}}$ as the positive set $L_+^{l_{src}}$ and leave the facts $L_-^{l_{src}} = K^{l_{src}} - L_+^{l_{src}}$ as the negative set. Then, we force the $E^{l_{src}}$ to faithfully estimate the relevance between the $k_{i,j}^{l_{src}}$ and $R^{l_{src}}$ based on the $H^{l_{src}}$. The corresponding objective is given by:

$$\mathcal{L}_E = \frac{1}{|K^{l_{src}}|} \sum_{k_i \in K^{l_{src}}} -\Big[U_i \cdot \log E^{l_{src}}\left(H^{l_{src}}, k_i^{l_{src}}\right) + (1-U_i) \cdot \log\left(1 - E^{l_{src}}\left(H^{l_{src}}, k_i^{l_{src}}\right)\right)\Big] \quad (3)$$

where $U_i = 1$ if $k_i^{l_{src}} \in L_+^{l_{src}}$ else $U_i = 0$.

Clustering To identify the repetitive facts across languages, we place facts with similar meanings into the same cluster in spite of the language. Given the retrieved facts $K^M = \{k_i^M\}^n$, we first compute the feature matrix $\mathbf{K_F^M}$:

$$\mathbf{K^M} \in \mathbb{R}^{n \times dim} = SBERT^{uni}(\delta(K^M)) \quad (4)$$

$$\mathbf{K_F^M} \in \mathbb{R}^{n \times n} = \mathbf{K^M}(\mathbf{K^M})^T \quad (5)$$

where unlike the Equation 2, this $SBERT^{uni}$ is frozen and shared by all languages[1], which can 1) encode each linearized knowledge fact into a fixed dimension embedding; and 2) ensure the embedding is language-agnostic as possible. Thus, if two facts of two different languages say the same thing, their embeddings would be similar to each other.

Then, we assume that if the features of the two facts are closer, they may say the same thing and should be placed into the same cluster. To this end, we adopt the unsupervised clustering method $KMeans$ to cluster K^M into clusters:

$$\{C_i = \{k_{i,j}\}\} = KMeans(\mathbf{K_F^M}) \quad (6)$$

where C_i is the i-th cluster, $\{k_{i,j}\}$ is corresponding fact set.

Penalized Selection This stage filters out the multilingual fact candidates and rearranges their orders. For each cluster $C_i = \{k_{i,j}\}$, we first rank all the elements according to their predicted relevance scores $\{s_{i,j}\}$. The output is given by:

$$C_i^R = \{(k_{i,j}, r_{i,j}, s_{i,j})\} \quad (7)$$

where $r_{i,j} \in [1, |C_i^R|]$ is the in-cluster rank order of $k_{i,j}$ and $s_{i,j}$ is the original ranking score of $k_{i,j}$.

Subsequently, we compute a repetition-penalized relevance score $s_{i,j}^p$ for each $k_{i,j}$:

$$s_{i,j}^p = \lambda^{(r_{i,j}-1)} \times s_{i,j} \quad (8)$$

where the penalty factor λ is a hyper parameter. $\lambda = 1$ denotes non-penalty, $\lambda \in (0, 1)$ penalizes the scores of repetitive facts according to its in-cluster rank.

[1]each $SBERT^{l_{src}}$ in Equation 2 will be fine-tuned on the language-specific relevance estimation task, while $SBERT^{uni}$ is directly loaded from the pre-trained checkpoint and without tuning.

Method	*Knowledge*	*BERT*	*F1*	*ROUGE-L*	*BLEU-1*	*BLEU-2*	*BLEU-3*	*BLEU-4*	*DIST-1*	*DIST-2*	*Ent*	*Mean*
mT5	-	0.393	46.86	24.60	23.20	12.03	7.63	5.19	5.42	20.68	8.82	9.32
mBART	-	0.407	45.85	24.55	21.32	11.15	7.23	5.04	7.01	27.35	9.12	9.64
mT5	Mono-20	0.456	49.49	**26.99**	**25.16**	**13.82**	**9.02**	**6.23**	5.60	21.91	9.00	10.28
mBART	Mono-20	**0.469**	**49.88**	26.20	24.21	13.12	8.54	5.97	**6.81**	**27.86**	**9.34**	**10.58**

Table 2: Monolingual Evaluation. Each model only considers one language. Scores are the average scores of all languages.

Finally, we re-rank all knowledge facts $\{k_{i,j}\}$ according to its penalized relevance score $\{s^p_{i,j}\}$ and keep the first half re-ranked facts as the output K^P.

Response Generation

To generate the target response $R^{l_{src}}$ based on the dialogue context $H^{l_{src}}$ and the multilingual knowledge set K^P outputted by the above *ECP* procedure. We adopt the multilingual Encoder-Decoder model *mBART* (Liu et al. 2020) as our backbone. The input is formatted as follows:

$$I = [H_1^{l_{src}}, H_2^{l_{src}}, \cdots, \theta_{MT}(k_1^p), \theta_{MT}(k_2^p), \cdots] \quad (9)$$

where $H_i^{l_{src}}$ is the i-th turn dialogue and $\theta_{MT}(k_j^p)$ is the linearized fact, which uses the pattern $[l_* : e_{head}, e_{rel}, l_* : e_{tail}]$, where l_* is a label to indicate the language type.

Training Considering mBART is a multilingual model, we use one uniform mBART to train all dialogues and adopt the Maximum Likelihood Estimation in training, and the corresponding objective function is defined as:

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{i \in |\mathcal{D}|} \sum_{r_{i,t} \in R_i} -log(P(R_{i,1:t-1}|H_i, K_i^P)) \quad (10)$$

Experiments

Experimental Setting

Dataset All experiments are conducted on the previously constructed *MMK-DailyDialog*.

Automatic Metrics We report the BERT-based Embedding Score (BERT) (Reimers and Gurevych 2020), F1 (Liu et al. 2023), BLEU1-4, ROUGE-L, DIST-1/2 (Li et al. 2016), Ent (Entropy-4) (Mou et al. 2016), and the geomean score of previous metrics to measure the overall performance. To avoid impacts from different tokenizers, all responses are tokenized by the tokenizer of mT5 before the evaluation.

Methods Two representative multilingual models *mBART* and *mT5* are adopted as baseline models. Both the baseline BART and our proposed *MMK-BART* are initialized from the *mBART-Large (611M parameters)*, while mT5 is initialized from the *mT5-Base (582M parameters)*. All methods are implemented with PyTorch and Huggingface library. Similar to (Liu et al. 2023), we use the mini-batch of 32 in fine-tuning. Depending on the 24GB V-GRAM of the Nvidia RTX-3090 GPU and the input length, the gradient acclimation step is set to either 4 (when there are 20*4 facts) or 2 (in other scenarios). We use the Adam optimizer and 500 warming steps. We search the learning rate and the epoch number based on the English subset. For mT5, we set the learning rate to 3e-4 and train 5 epochs. For the mBART(MMK-BART), we set the learning rate to 2.5e-5 and train 3 epochs. In the inference, we select the last epoch and use the beam width of 5. For our *MMK-BART*, *ECP* clusters facts into 10 groups and sets the penalty factor λ to 0.99 by default.

Monolingual Evaluation

Prior *XDailyDialog* (Liu et al. 2023) did not consider KRG tasks, and *MMK-DailyDialog* can only do the knowledge alignment in a post-processing manner. Thus, it is not trivial to verify the knowledge-alignment effectiveness in our extended *MMK-DailyDialog*. As reported results in Table 2, we can find after using the aligned commonsense knowledge, two backbone models can significantly improve the performance in every dimension. It can undoubtedly prove our knowledge alignment procedure is rational and efficient. Meanwhile, the two backbone models have different tendencies. mT5 has higher performance in BLEU scores, while mBART has higher diversity and overall score.

Multilingual Evaluation

In this evaluation, each method uses one uniform model to handle dialogues of all languages. The corresponding results have been reported in Table 3. The proposed *MMK-BART* has undoubtedly achieved the best performance among all models. Besides, we also have several findings as follows:

Uniform model is better Although the first four models in Table 3 can not use multiple multilingual knowledge sources at the same time, they can implicitly benefit from other languages via the shared uniform backbone language model. Compared to using several separated versions in Table 2, we can find all four methods have gained notable improvements without explicitly accessing more external knowledge. It shows that knowledge can be implicitly transferred from one language to another language via sharing a backbone language model. Thus, developing and using multilingual backbone models is crucial in all KRG scenarios.

More knowledge is not always better For the backbone model mBART, we can find that if directly using all multilingual knowledge (i.e., Multi-80), the performance is even worse than only using the corresponding monolingual knowledge (i.e., Mono-20). In other words, it uses more knowledge and more computational expense but outputs less performance. Such results can vividly demonstrate our previous statement that simply combining knowledge retrieved

Method	*Knowledge*	*BERT*	*F1*	*ROUGE-L*	*BLEU-1*	*BLEU-2*	*BLEU-3*	*BLEU-4*	*DIST-1*	*DIST-2*	*Ent*	*Mean*
mT5	-	0.410	48.38	25.61	24.48	12.98	8.29	5.66	5.05	19.84	8.91	9.62
mBART	-	0.422	48.46	25.28	23.23	12.46	8.19	5.78	6.54	27.27	9.36	10.17
mT5	Mono-20	0.487	51.42	28.54	26.83	15.23	10.12	7.11	5.41	22.06	9.18	10.87
mBART	Mono-20	0.499	51.90	28.88	26.20	15.14	10.31	7.47	6.89	29.40	9.59	11.61
mT5	Multi-80	0.512	52.34	29.71	27.45	15.67	10.52	7.52	5.29	22.19	9.31	11.15
mBART	Multi-80	0.508	51.72	29.32	25.71	14.77	10.08	7.34	**6.91**	29.70	9.62	11.56
MMK-BART	Multi-40	**0.553**	**54.75**	**30.97**	**28.78**	**17.18**	**11.89**	**8.75**	6.89	**30.69**	**9.80**	**12.59**

Table 3: Multilingual Evaluation Results. All languages share one model (i.e., uniform models). *Mono/Multi-k* uses up to k monolingual/multilingual commonsense knowledge facts. Scores are the average scores of all languages.

Method	*Know.*	*BERT*	*F1*	*ROUGE-L*	*BLEU-1*	*BLEU-2*	*BLEU-3*	*BLEU-4*	*DIST-1*	*DIST-2*	*Ent*	*Mean*
Full	Multi-40	0.553	54.75	30.97	28.78	17.18	11.89	8.75	6.89	30.69	9.80	12.59
-w/o. Penalization	Multi-40	0.548	54.80↑	30.27	28.73	17.01	11.75	8.63	6.60	29.54	9.77	12.40
-w/o. Estimation	Multi-40	0.534	53.87	29.85	27.71	16.19	11.06	8.03	6.91↑	30.15	9.73	12.14
-w/o. ECP	Multi-40	0.470	50.39	27.47	24.62	13.74	9.20	6.60	6.86	28.85	9.52	10.97
-w/o. Uni	Multi-40	0.469	49.00	26.31	23.72	12.90	8.41	5.82	7.08↑	28.63	9.34	10.55

Table 4: Ablated Evaluation Results. Except for ↑, the full *MMK-BART* is better than ablated models.

from multiple multilingual sources would cause several issues such as *Cross-Repetition* and the *Cross-Conflict*.

Estimate-Cluster-Penalize is efficient To address the mentioned two issues, *MMK-BART* uses a novel *Estimate-Cluster-Penalize (ECP)* mechanism to remove the redundant and irrelevant facts, as well as rearrange the order of facts. As shown in Table 3, our *MMK-BART* can significantly outperform the multilingual mBART and mT5 with only half the scale of the knowledge, demonstrating the superior effectiveness and efficiency of our *MMK-BART* and *ECP*.

Ablation Study

To decompose the contribution of *MMK-BART*, we conduct a ablation study by removing each module in Table 4:

- *-w/o. Penalization* does not consider the in-cluster rank by setting the penalty λ to 1.0, and thus more repetitive facts may be selected. The decreased performance can prove the existence of the *Cross-Repetition* issue and the effectiveness of our solution.
- *-w/o. Estimation* does not use the proposed cross-encoder relevance estimation scoring function but uses the prior embedding similarity. By removing this, the decreased performance shows the existence of the *Cross-Conflict* issue and the effectiveness of our solution.
- *-w/o. ECP* fully removes our *Estimate-Cluster-Penalize* mechanism. The magnitude of performance degradation is greater than removing a portion separately, demonstrating our *ECP* is an efficient pipeline.
- *-w/o. Uni* uses one separate model to train each language. In this case, there has been a sharp decline in performance, even lower than using only the monolingual knowledge (see the mBART+Mono-20 in Table 2). It reveals understanding and leveraging multilingual knowledge is a challenging job, whose training needs enough multilingual conversational instances at the same time.

Zero Native Knowledge Evaluation

Once no available native-language knowledge can be used, previous monolingual KRG models will fall back to the traditional RG models. A possible solution is to use cross-lingual knowledge as an alternative. To verify the feasibility, we have conducted extensive experiments in Table 5.

Separated Models If each language uses one separated model, we find using the cross-lingual knowledge results in worse performance (*mBART+None* v.s. *mBART+Cross-60*). By using the *Estimate-Cluster-Penalized (ECP)* mechanism, *MMK-BART* is comparable to using the native knowledge (*mBART+Mono-20*), demonstrating the *Cross-Repetition* issue and *Cross-Conflict* issue are also serious in this scenario.

Uniform Models Once we use one uniform model to handle all languages, all methods have notable improvements. It demonstrates that knowledge can be implicitly transferred via the backbone language model. Our *MMK-BART* can beat the native *mBART+Mono-20* in this scenario. The native *mBART+Mono-20* has the best result in English because English has the most knowledge abundance (see Table 1).

Parameter Sensitivity Analysis

Estimate-Cluster-Penalize has two hyper-parameters, the number of clusters and the penalize factor λ. We design a parameter sensitivity analysis in Figure 3. We first freeze the penalize factor $\lambda = 0.99$, and check the performance of $\{2, 5, 10, 20, 40\}$ clusters. It can be seen that the performance reaches the best when clustering 80 facts into 10 clusters. Fewer clusters would bring more significant penalties to the repetitive facts while more clusters bring smaller penalties; thus, a moderate cluster number is crucial. Next, we freeze the cluster number to 10 and check the performance when $\lambda \in \{0.91, 0.93, 0.95, 0.97, 0.99\}$. Unlike the previous, no obvious pattern can be found in this test. This means we have to select λ according to the practical experiments.

Method	*Know.*	*BERT*	*F1*	*B4*	*D2*	*BERT*	*F1*	*B4*	*D2*	*BERT*	*F1*	*B4*	*D2*	*BERT*	*F1*	*B4*	*D2*	*Mean*
		English				*Chinese*				*German*				*Italian*				*Avg.*
Separated Models																		
mBART	-	0.38	53.7	6.28	27.4	0.43	30.3	5.01	34.1	0.42	53.3	4.36	26.9	0.40	50.9	4.69	21.5	9.64
mBART	Mono-20	0.46	57.4	8.00	27.2	0.47	32.9	5.69	36.7	0.44	57.3	5.15	23.4	0.51	57.4	6.73	25.8	10.58
mBART	Cross-60	0.38	53.3	6.12	24.5	0.44	30.9	5.26	34.6	0.42	52.8	3.92	26.4	0.37	49.5	3.85	21.0	9.53
MMK-BART	Cross-30	0.40	56.9	6.57	21.0	0.47	32.3	5.74	36.0	0.48	58.9	4.60	23.1	0.44	53.4	5.38	25.3	10.05
Uniform Models																		
mBART	None	0.39	56.5	7.43	24.0	0.45	31.8	5.48	36.9	0.44	56.6	5.2	26.0	0.41	54.2	5.27	24.0	10.17
mBART	Mono-20	**0.48**	60.4	**9.53**	**27.4**	0.50	34.8	6.58	37.9	0.49	58.4	6.4	27.7	**0.52**	59.1	7.76	**25.9**	11.61
mBART	Cross-60	0.39	55.7	7.17	24.8	0.44	31.3	5.30	36.8	0.42	53.8	4.61	28.5	0.43	53.3	5.45	25.5	10.10
MMK-BART	Cross-30	0.47	**61.0**	8.71	25.0	**0.55**	**36.6**	**7.55**	**40.0**	**0.56**	**62.1**	**7.43**	**29.1**	0.49	**60.2**	**9.98**	24.0	**11.79**

Table 5: Cross-lingual Evaluation Results. *Cross-k* uses up to k non-native cross-lingual commonsense knowledge facts. It *must be noted* that, to avoid knowledge leakage via the shared uniform model, we have actually constructed four different three-sources(languages) knowledge uniform models for each *Uniform* case.

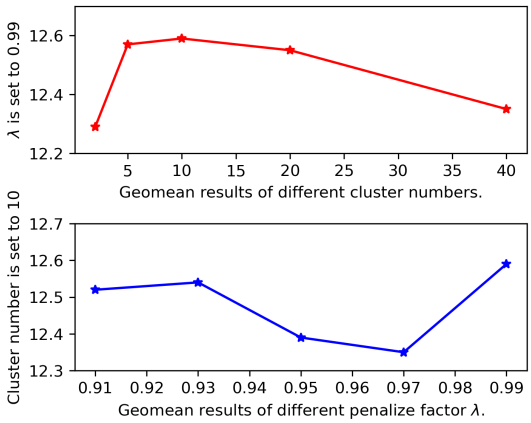

Figure 3: Parameter Sensitivity Analysis

	Fluency		*Rel.*		*Infor.*		*Mean*	
-	*G*	*H*	*G*	*H*	*G*	*H*	*G*	*H*
Vanilla	4.53	4.63	2.58	3.44	2.01	2.72	3.04	3.60
Mono-20	4.48	**4.67**	2.74	3.44	2.17	2.52	3.13	3.54
Multi-80	**4.55**	4.52	2.77	3.08	2.22	2.49	3.18	3.36
MMK-BART	4.46	4.63	**3.04**	**3.55**	**2.29**	**2.78**	**3.26**	**3.65**

Table 6: LLM/Human Evaluation. Except for *MMK-BART*, other methods use a uniform *mBART*. *G* scores are given by the GPT3.5 Turbo, *H* scores are given by human.

LLM and Human Evaluation

Human evaluation is necessary for evaluating generated responses. However, this work studies the multilingual KRG; it is hard to organize native-speaker volunteers for every language. Thus, we only employed three native speakers to evaluate the Chinese responses. To fully evaluate responses of all languages, we also conducted a human-like evaluation on the GPT3.5 Turbo LLM (Azure version), which has outstanding multilingual understanding ability and less subjective preferences. Such two evaluations consider three aspects: *1) Fluency*, *2) Relevance*, and *3) informativeness*. We sampled 100 dialogues from each model and designed an instruction prompt pattern to guide LLM on how to evaluate a given dialogue response.

As reported in Table 6, our *MMK-BART* has the best overall performance than baselines. The only exception is fluency. We think the reason is that higher informativeness also means more complicated content, which may impact fluency. Meanwhile, we conducted more analysis to check the validity of these evaluations. We began by categorizing the evaluation scores into three levels: $\{1, 2\}$ as *Negative*, $\{3\}$ as *Neutral*, $\{4, 5\}$ as *Positive*, respectively. Then, we count the agreements among human annotators. In the evaluation of *MMK-BART*, the 2/3 and 3/3 agreement proportions were 1.00/0.86 for *Fluency*, 0.96/0.64 for *Relevance*, and 0.92/0.54 for *Informativeness*. This result can indicate that annotators have given relatively consistent evaluation results. Further, we investigated the correlation between the LLM's assessments and human judgments using the same 3-scale labels. We calculated Cohen's Kappa ($k \approx 0.4426$) by analyzing the confusion matrix, demonstrating that the agreement between the LLM annotations and human annotations is moderate beyond chance. More details of our LLM instruction prompt pattern and some sampled dialogue cases can be found in our GitHub project https://github.com/Y-NLP/Chatbots/tree/main/AAAI2024_MMK-BART.

Conclusion

To enrich the scale of accessible knowledge in the context of Knowledge-Grounded Response Generation (KRG), this work proposes to use multi-source multilingual knowledge sources. To this end, this work first constructs a dataset *MMK-XDailyDialog* by aligning the multilingual corpus *XDailyDialog* to the multilingual knowledge base *ConceptNet*. Then, we propose an approach *MMK-BART*, which uses the novel *Estimate-Cluster-Penalize* mechanism to alleviate the *Cross-Conflict* issue and the *Cross-Repetition* issue. Finally, extensive experiments have verified the effectiveness of our dataset and approach in scenarios that use monolingual, cross-lingual, or multilingual knowledge. To our best knowledge, both the corpus *MMK-XDailyDialog* and the approach *MMK-BART* are the first of their kind.

Ethics Statement

We did not introduce any new ethical statements/considerations since all the techniques, datasets, and research topics involved are verified by the public community. Meanwhile, all involved researchers/volunteers have been well paid.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 62162067 and 62101480, Research and Application of Object Detection based on Artificial Intelligence, in part by the Yunnan Province expert workstations under Grant202305AF150078.

References

Huang, Z.; Yu, P.; and Allan, J. 2023. Improving Cross-Lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, 1048–1056. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394079.

Jang, Y.; Lim, J.; Hur, Y.; Oh, D.; Son, S.; Lee, Y.; Shin, D.; Kim, S.; and Lim, H. 2022. Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 10803–10812. AAAI Press.

Kann, K.; Ebrahimi, A.; Koh, J.; Dudy, S.; and Roncone, A. 2022. Open-domain Dialogue Generation: What We Can Do, Cannot Do, And Should Do Next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, 148–165. Dublin, Ireland: Association for Computational Linguistics.

Kim, S.; Jang, J. Y.; Jung, M.; and Shin, S. 2021. A Model of Cross-Lingual Knowledge-Grounded Response Generation for Open-Domain Dialogue Systems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 352–365. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, Y.; Peng, B.; Shen, Y.; Mao, Y.; Liden, L.; Yu, Z.; and Gao, J. 2022. Knowledge-Grounded Dialogue Generation with a Unified Knowledge Representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 206–218. Seattle, United States: Association for Computational Linguistics.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Liang, Y.; Meng, F.; Zhang, Y.; Chen, Y.; Xu, J.; and Zhou, J. 2021. Infusing Multi-Source Knowledge with Heterogeneous Graph Neural Network for Emotional Conversation Generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 13343–13352. AAAI Press.

Lin, X.; Jian, W.; He, J.; Wang, T.; and Chu, W. 2020a. Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 41–52.

Lin, Z.; Liu, Z.; Winata, G. I.; Cahyawijaya, S.; Madotto, A.; Bang, Y.; Ishii, E.; and Fung, P. 2020b. XPersona: Evaluating Multilingual Personalized Chatbot. *CoRR*, abs/2003.07568.

Liu, Y.; Feng, S.; Wang, D.; and Zhang, Y. 2022. MulZDG: Multilingual Code-Switching Framework for Zero-shot Dialogue Generation. In Calzolari, N.; Huang, C.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.; Ryu, P.; Chen, H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S., eds., *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 648–659. International Committee on Computational Linguistics.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Liu, Z.; Nie, P.; Cai, J.; Wang, H.; Niu, Z.-Y.; Zhang, P.; Sachan, M.; and Peng, K. 2023. XDailyDialog: A Multilingual Parallel Dialogue Corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12240–12253. Toronto, Canada: Association for Computational Linguistics.

Majewska, O.; Razumovskaia, E.; Ponti, E. M.; Vulić, I.; and Korhonen, A. 2023. Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation. *Transactions of the Association for Computational Linguistics*, 11: 139–156.

Moghe, N.; Vijayan, P.; Ravindran, B.; and Khapra, M. M. 2020. On Incorporating Structural Information to improve Dialogue Response Generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 11–24. Online: Association for Computational Linguistics.

Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. In Calzolari, N.; Matsumoto, Y.; and Prasad, R., eds., *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, 3349–3358. ACL.

Qin, L.; Liu, Y.; Che, W.; Wen, H.; Li, Y.; and Liu, T. 2019. Entity-Consistent End-to-end Task-Oriented Dialogue System with KB Retriever. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 133–142. Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Speer, R.; Chin, J.; and Havasi, C. 2016. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *CoRR*, abs/1612.03975.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 3104–3112.

Urbanek, J.; Fan, A.; Karamcheti, S.; Jain, S.; Humeau, S.; Dinan, E.; Rocktäschel, T.; Kiela, D.; Szlam, A.; and Weston, J. 2019. Learning to Speak and Act in a Fantasy Text Adventure Game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 673–683.

Wu, S.; Li, Y.; Wang, M.; Zhang, D.; Zhou, Y.; and Wu, Z. 2021. More is Better: Enhancing Open-Domain Dialogue Generation via Multi-Source Heterogeneous Knowledge. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2286–2300. Association for Computational Linguistics.

Wu, S.; Li, Y.; Zhang, D.; Zhou, Y.; and Wu, Z. 2020. Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 5811–5820. Association for Computational Linguistics.

Wu, S.; Wang, M.; Li, Y.; Zhang, D.; and Wu, Z. 2022. Improving the Applicability of Knowledge-Enhanced Dialogue Generation Systems by Using Heterogeneous Knowledge from Multiple Sources. In Candan, K. S.; Liu, H.; Akoglu, L.; Dong, X. L.; and Tang, J., eds., *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, 1149–1157. ACM.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498. Online: Association for Computational Linguistics.

Yan, R. 2018. "Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 5520–5526.

Yu, W.; Zhu, C.; Li, Z.; Hu, Z.; Wang, Q.; Ji, H.; and Jiang, M. 2020. A Survey of Knowledge-Enhanced Text Generation. *CoRR*, abs/2010.04389.

Zhang, H.; Liu, Z.; Xiong, C.; and Liu, Z. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2031–2043.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 4623–4629.

Zhou, P.; Gopalakrishnan, K.; Hedayatnia, B.; Kim, S.; Pujara, J.; Ren, X.; Liu, Y.; and Hakkani-Tur, D. 2022. Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1237–1252. Dublin, Ireland: Association for Computational Linguistics.