More is Better: Enhancing Open-Domain Dialogue Generation via Multi-Source Heterogeneous Knowledge

Sixing Wu[1], Ying Li[4,5]*, Minghui Wang[2], Dawei Zhang[1], Yang Zhou[3], Zhonghai Wu[4,5]
[1]School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2]School of Software and Microelectronics, Peking University, Beijing, China
[3]Auburn University, Auburn, Alabama, USA
[4]National Research Center of Software Engineering, Peking University, Beijing, China
[5]Key Lab of High Confidence Software Technologies (MOE),
Peking University, Beijing, China

Abstract

Despite achieving remarkable performance, previous knowledge-enhanced works usually only use a single-source homogeneous knowledge base of limited knowledge coverage. Thus, they often degenerate into traditional methods because not all dialogues can be linked with knowledge entries. This paper proposes a novel dialogue generation model, *MSKE-Dialog*, to solve this issue with three unique advantages: (1) Rather than only one, *MSKE-Dialog* can simultaneously leverage multiple heterogeneous knowledge sources (it includes but is not limited to commonsense knowledge facts, text knowledge, infobox knowledge) to improve the knowledge coverage; (2) To avoid the topic conflict among the context and different knowledge sources, we propose a *Multi-Reference Selection* to better select context/knowledge; (3) We propose a *Multi-Reference Generation* to generate informative responses by referring to multiple generation references at the same time. Extensive evaluations on a Chinese dataset show the superior performance of this work against various state-of-the-art approaches. To our best knowledge, this work is the first to use the multi-source heterogeneous knowledge in the open-domain knowledge-enhanced dialogue generation.

1 Introduction

The rapid developments of knowledge-enhanced techniques have enabled machines to understand the instinct semantics of human conversations further and generate informative responses (Yu et al., 2020). External knowledge, such as commonsense bases (Speer et al., 2017), documents (Zhao et al., 2020), and tables (Wu et al., 2021), can bridge the gap between machines and humans in conversation by generously providing knowledge that is hard to be learned from a conversational corpus (Ghazvininejad et al., 2018).

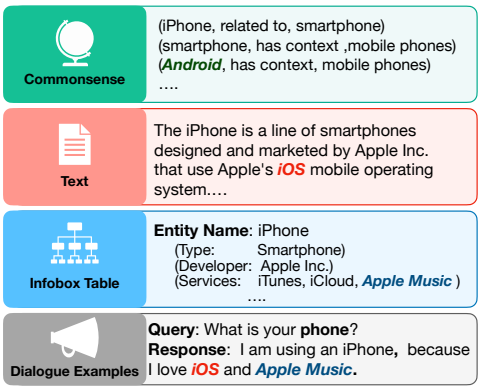

Figure 1: Examples of generating dialogues with multi-source heterogeneous knowledge. The number of knowledge sources can exceed one, and the knowledge structure can vary among sources.

However, previous knowledge-enhanced works are still far from satisfactory because they usually solely rely on a single-source homogeneous knowledge base: (1) Conversations are diverse because humans are free to talk about whatever topics they like (Li et al., 2016; Hu et al., 2020), but the knowledge coverage of a single knowledge base is limited. Thus, only a finite portion of dialogues could benefit from the external knowledge; the remaining can only rely on the given query because no knowledge can be matched. Suffering from the long-tail issue and the cost of a massive workforce, it is not wise to improve the coverage by expanding the number of entries in a single-source knowledge base. (2) Each knowledge source has its advantages and disadvantages (Liu et al., 2019), for example, plain text has richer information than knowledge graph, but it performs worse in logically modeling. No knowledge type can always perform best; the most suitable knowledge only depends on the case.

Human beings can use various kinds of knowledge learned from different sources. Therefore, as shown in Figure 1, we believe using multiple

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

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2286–2300
November 7–11, 2021.

knowledge sources can improve knowledge coverage and have more room to select appropriate knowledge. However, every coin has two sides; using multi-source heterogeneous is more challenging because of the following two conflicts: *(1) Topic Conflict:* given a dialogue, knowledge entries are usually retrieved by the entity name matching technique (Wu et al., 2020a). Thus, knowledge entries retrieved from one source may be irrelevant to the dialogue context and have different topics compared to entries retrieved from other sources. Blindly using such irrelevant/conflicting knowledge entries can confuse the model; *(2) Generation Conflict:* Although dialogue utterances and different knowledge bases are made of words, the word distributions vary among them. It can affect the generation if a model tries to improve the informativeness by copying words from knowledge entries. For example, if a word 'apple' appears in both the dialogue context and the commonsense knowledge, there exist two tokens of 'apple' in the dialogue vocab and the commonsense vocab, respectively. Then, two 'apple' will have two different tokens/probabilities when predicting the next word, making it difficult for a model to judge which one should be the objective. This issue is severe when using multi-source heterogeneous knowledge. With more knowledge sources, there are more chances for conflicts; then, the more conflicts, the lower the response quality.

This paper proposes a novel multi-source heterogeneous knowledge-enhanced dialogue generation model, *MSKE-Dialog*. *MSKE-Dialog* can improve knowledge coverage by integrating more knowledge sources. In this paper, we use commonsense knowledge, text knowledge, and infobox knowledge at the same time. Compared to only use one of them, simultaneously using such three knowledge sources can improve the coverage by $63 \sim 200\%$ in our dataset. To alleviate the impact of topic conflict, we propose a *Multi-Reference Selection* mechanism. It uses a global relevance gate and a dynamic selection gate to select relevant knowledge from different sources. We also propose a *Multi-Reference Generation* mechanism, which will construct a unified dynamic vocab and comprehensively refer to all inputs (i.e., the context and the multi-source knowledge) during the decoding. As a result, *MSKE-Dialog* can avoid the impact of generation conflict as possible and generate an informative response.

Our experiments are conducted on a Chinese Weibo dataset. In both automatic and human evaluations, *MSKE-Dialog* can outperform various state-of-the-art knowledge-enhanced methods by notable margins, as well as can surpass the fine-tuned pre-training system *CDial-GPT* (GPT & GPT2) (Wang et al., 2020b) even with fewer parameters and training corpus. Extensive deep analyses also demonstrate: (1) Compared to simply integrating multiple knowledge bases, *MSKE-Dialog* has better performance because it can alleviate two mentioned challenging conflicts; (2) Even if *MSKE-Dialog* only uses a single-source knowledge, our model can also achieve promising results. It demonstrates the performance gain comes from not only the multi-source knowledge but also the approach itself.

2 Approach

2.1 Problem Statement and Overview

The goal is to generate the dialogue response $Y = (y_1, \ldots, y_{l_Y})$ conditioned on $\mathcal{R}$, where $\mathcal{R} = (R_X, \{R_{K_i}\})$ is a set of given references to guide the generation. $R_X = (r_{X,1}, \ldots, r_{X,l_X})$ represents the dialogue context (history), $\{R_{K_i}\}$ represents a set of multi-source heterogeneous knowledge, where the i-th $R_{K_i} = (r_{K_i,1}, \ldots, r_{K_i,l_{K_i}})$ represents the relevant entries retrieved from the i-th knowledge source. Considering both R_X and $\{R_{K_i}\}$ serve as a type of reference in the response generation stage, we call R_X and $\{R_{K_i}\}$ as dialogue reference and knowledge references, respectively. Thus, $\mathcal{R}$ is called as the reference set.

As shown in Figure 2, *MSKE-Dialog* employs three heterogeneous knowledge sources; in other words, $\{R_{K_i}\}$ contains the commonsense knowledge R_{K_C}, the text knowledge R_{K_T}, and the infobox knowledge R_{K_I}. The high-level architecture of *MSKE-Dialog* consists of three parts. **(1) Reference Encoding:** We propose four different encoders to encode the given references $R_X, R_{K_C}, R_{K_T}, R_{K_I}$ into intermediate hidden representations $\mathbf{R_X}, \mathbf{R_{K_C}}, \mathbf{R_{K_T}}, \mathbf{R_{K_I}}$, respectively. **(2) Reference Selection:** In the decoding stage, we update the decoder with not only the last predicted token, but also the context-aware readouts gathered from the encoded reference set $\mathbf{R}$. To obtain conflict-free readouts from the encoded $\mathbf{R}$, we propose a *Multi-Reference Selection* mechanism. **(3) Multi-Reference Guided Generation:** *MSKE-Dialog* can not only generate a word from the fixed vocabulary, but also copy a word from

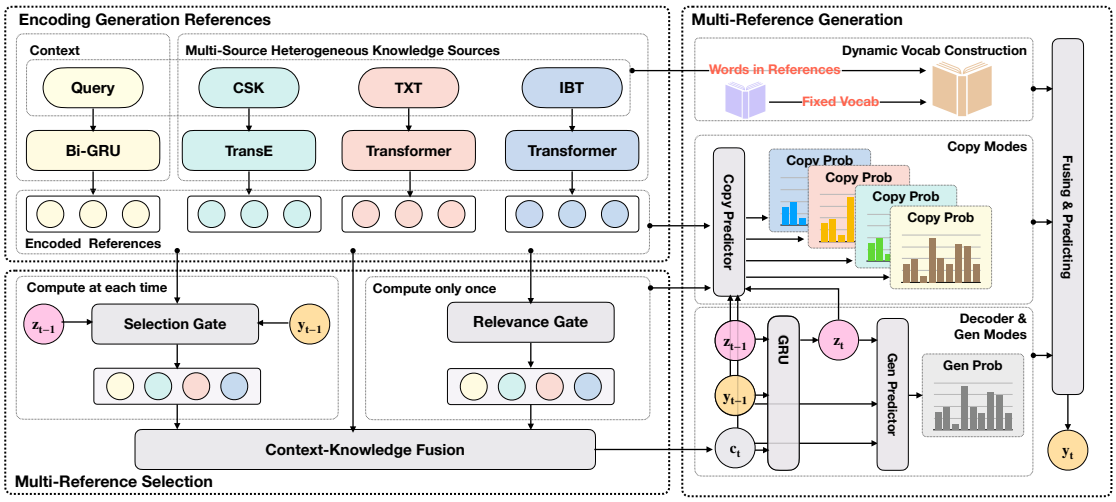

Figure 2: The framework of *MSKE-Dialog*.

$\mathcal{R}$. To avoid the conflicts during the generation, we propose a *Multi-Reference Generation* mechanism and a *Dynamic Copy* mechanism.

2.2 Reference Encoding

Dialogue Reference: Each word $r_{X,t} \in R_X$ is first embedded as $\mathbf{r^w_{X,t}}$, with the fixed vocab $\mathcal{V}_{R_X}$. Then, a bi-directional GRU network (denoted as g) (Cho et al., 2014) is adopted to encode R_X into hidden states $\mathbf{R_X} = (\mathbf{r_{X,1}}, \cdots, \mathbf{r_{X,l_X}})$, $\mathbf{r_{X,t}} = [\mathbf{r^{\leftarrow}_{X,t}}; \mathbf{r^{\rightarrow}_{X,t}}]$, $[\cdot;\cdot]$ indicates the concatenation operation:

$$\begin{aligned} \mathbf{r^{\rightarrow}_{X,t}} &= g^{\rightarrow}(\mathbf{r^w_{X,t}}, \mathbf{r^{\rightarrow}_{X,t-1}}) \\ \mathbf{r^{\leftarrow}_{X,t}} &= g^{\leftarrow}(\mathbf{r^w_{X,t}}, \mathbf{r^{\leftarrow}_{X,t+1}}) \end{aligned} \tag{1}$$

Commonsense Reference: Each entry $r_{K_C,n} \in R_{K_C}$ is a fact triplet $r_{K_C,n} = (e_{h,n}, e_{r,n}, e_{t,n})$, where $e_{h/t}$ is the head/tail entity, e_r is the relation. Following Zhang et al. (2020), we adopt TransE[1] (Bordes et al., 2013) to learn the embedding $\mathbf{e_{h/r/t,n}}$ with the vocab $\mathcal{V}_{R_{K_C}}$. TransE learns the translation-based embedding as:

$$\mathbf{e_{h,n}} + \mathbf{e_{r,n}} \approx \mathbf{e_{t,n}} \tag{2}$$

Thus, $\mathbf{r_{K_C,n}} = [\mathbf{e_{h,n}}; \mathbf{e_{r,n}}; \mathbf{e_{t,n}}]$ is the encoded entry, and the encoded commonsense knowledge entry set is denoted as $\mathbf{R_{K_C}} = \{\mathbf{r_{K_c,n}}\}$.

Text Reference: Each text reference is a word sequence $R_{K_T} = (r_{K_T,1}, \ldots, r_{K_T,l_{K_T}})$. Thus, each token $r_{K_T,n}$ is first embedded as $\mathbf{r^w_{K_T,n}}$ with the vocab $\mathcal{V}_{R_{K_T}}$. Considering R_{KT} is a long text paragraph, we use a 2-layer Transformer (Vaswani et al., 2017) to encode the sequence efficiently:

$$\mathbf{R_{K_T}} = \{\mathbf{r_{K_T,n}}\} = Transformer_{T2}(\{\mathbf{r^w_{K_T,n}}\}) \tag{3}$$

Infobox Reference: Following Liu et al. (2018), each infobox table R_{K_I} is first regarded as a set of key-value attributes $\{(a^k_n, a^v_n)\}$, where each key a^k_n is an noun phrase, and each value $a^v_n = (a^w_{n,1}, \cdots, a^w_{n,|a^v_n|})$ is a short text. Thus, R_{K_I} can be subsequently decomposed to a set of key-word pairs $\{a^{kw}_{n,m}\}$,where each key-word pair $a^{kw}_{n,m}$ includes the n-th key a^k_n and the m-th word in the n-th value $a^w_{n,m}$. Then, $a^{kw}_{n,m}$ is embedded as:

$$\mathbf{a^{kw}_{n,m}} = [\mathbf{a^k_n}; \mathbf{a^w_{n,m}}; \mathbf{pos_{n,m}}] \tag{4}$$

where the attribute key embedding $\mathbf{a^k_n}$ uses the vocab $\mathcal{V}_{R_{K_{I,K}}}$, the attribute word embedding $\mathbf{a^w_{n,m}}$ uses the vocab $\mathcal{V}_{R_{K_I}}$, the positional embedding $\mathbf{pos_{n,m}}$ is appended to indicate the position (i.e, n, m). After decomposing key-value pairs to key-word pairs, the number of pairs will significantly increase. Therefore, for the efficiency, we use a 2-layer Transformer to encode key-word pairs:

$$\mathbf{R_{K_I}} = \{\mathbf{r_{K_I,n,m}}\} = Transformer_{I2}(\{\mathbf{a^{kw}_{n,m}}\}) \tag{5}$$

2.2.1 Scalability

Now, the reference set $\mathcal{R}$ has been encoded to $\mathbf{R} = (\mathbf{R_X}, \{\mathbf{R_{K_i}}\})$[2]. Each encoded $\mathbf{R_j}$ can be

[1]TransE is not the STOA method. However, this paper does not focus on embedding learning. For comparing models accurately, we use TransE as previous works do.

[2]The index $i \in \{C, T, I\}$ is only used to index a knowledge source, $j \in \{X, K_C, K_T, K_I\}$ is only used to index a generation reference.

similarly regarded as a set of embedding $\{\mathbf{r_{j,n}}\}$. The remaining part of *MSKE-Dialog* does not have any knowledge-type-specific module. Consequently, *MSKE-Dialog* has the superior scalability because it can remove a knowledge R_{K_i} by simply removing it from $\mathcal{R}$, or add a new knowledge source to $\mathcal{R}$ by adding a corresponding encoder.

2.3 Reference Selection

State Updating: At each decoding step t, the decoder state $\mathbf{z_t}$ is firstly updated with a GRU unit g_d, the embedding of the last generated token $\mathbf{y_{t-1}}$, and the context-aware reference readout $\mathbf{c_t}$:

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

Multi-Reference Selection: The reference readout $\mathbf{c_t}$ is obtained by fusing local reference readouts $\{\mathbf{r^c_{R_j,t}}\} = \{\mathbf{r^c_{X,t}}, \mathbf{r^c_{K_C,t}}, \mathbf{r^c_{K_T,t}}, \mathbf{r^c_{K_I,t}}\}$ with relevance gates $\{\alpha^{rel}_{R_j}\}$ and selection gates $\{\alpha^{sel}_{R_j,t}\}$:

$$\mathbf{c_t} = \sum_{R_j \in \mathcal{R}} \frac{\alpha^{rel}_{R_j} \cdot \alpha^{sel}_{R_j,t}}{\sum_{R_m \in \mathcal{R}} \alpha^{rel}_{R_m} \cdot \alpha^{sel}_{R_m,t}} \mathbf{r^c_{R_j,t}} \tag{7}$$

where the dialogue reference readout $\mathbf{r^c_{X,t}}$ gathers from the encoded $\mathbf{R_X} = (\mathbf{r_{X,1}}, \cdots, \mathbf{r_{X,l_X}})$[3]:

$$\sum_n \frac{\exp((\mathbf{W^{R_k}_X r_{X,n}})^\top (\mathbf{W^{R_q}_X z_{t-1}})) \mathbf{W^{R_v}_X r_{X,n}}}{\sum_m \exp((\mathbf{W^{R_k}_X r_{X,m}})^\top (\mathbf{W^{R_q}_X z_{t-1}}))} \tag{8}$$

and each knowledge reference readout $\mathbf{r^c_{R_{K_i},t}} \in \{\mathbf{r^c_{K_C,t}}, \mathbf{r^c_{K_T,t}}, \mathbf{r^c_{K_I,t}}\}$ gathers from the encoded $\mathbf{R_{K_{C/T/I}}}$, respectively:

$$\mathbf{r^c_{K_i,t}} = \sum_n \frac{\exp(\mathbf{r_{K_i,n}}^\top \mathbf{W^{R_a}_{K_i} z_{t-1}}) \mathbf{W^{R_v}_{K_i} r_{K_i,n}}}{\sum_m \exp(\mathbf{r_{K_i,m}}^\top \mathbf{W^{R_a}_{K_i} z_{t-1}})} \tag{9}$$

Relevance Gate: Each reference $R_j \in \mathcal{R}$ may have various importance, and may have conflicts with other references. Thus, we employ a global *Relevance Gate* $\alpha^{rel}_{R_j} \in (0, 1)$ to control the participation of each reference. Each relevance gate $\alpha^{rel}_{R_j}$ is given before the decoding:

$$\alpha^{rel}_{R_j} = \sigma(\mathbf{W^{G_2}_{R_j}}\, ELU(\mathbf{W^{G_1}_{R_j}}[\mathbf{W^G_X r_{X,l_X}}; \mathbf{s_{R_j}}])) \tag{10}$$

where σ is the $sigmoid$ activation function, ELU is another activation function (Clevert et al., 2016), $\mathbf{s_{R_j}}$ is the reference summary of R_j.

Each reference summary $\mathbf{s_{R_j}}$ is given by taking the last dialogue reference state $\mathbf{r_{X,l_X}}$ as attention query, and the encoded reference $\mathbf{R_j} = \{\mathbf{r_{j,n}}\}$ as keys/values :

$$\sum_n \frac{\exp((\mathbf{W^{S_k}_{R_j} r_{j,n}})^\top (\mathbf{W^{S_q}_{R_j} r_{X,l_X}}))}{\sum_m \exp((\mathbf{W^{S_k}_{R_j} r_{j,m}})^\top (\mathbf{W^{S_q}_{R_j} r_{X,l_X}}))} \mathbf{W^{S_v}_{R_j} r_{j,n}} \tag{11}$$

Selection Gate: During each decoding step, we employ a dynamic context-aware *Selection Gate* $\alpha^{sel}_{R_j,t}$ to control the fine-grained usage of R_j :

$$\mathbf{a^{sel}_t} = \epsilon(\mathbf{W^D}[\mathbf{z_{t-1}}; \{\mathbf{r^c_{R_j,t}}\}; \mathbf{y_{t-1}}]) \tag{12}$$

where ϵ is the $softmax$ operation, $\mathbf{a^{sel}_t} \in \mathbb{R}^{|\mathcal{R}|}$; thus, each local selection gate is $\alpha^{sel}_{R_j,t} = \mathbf{a^{sel}_t}[j]$.

2.4 Multi-Reference Guided Generation

Copying words besides the fixed vocabulary has shown great potential in promoting OOV-free, informative and diverse responses (Lin et al., 2020). However, token distributions are various among multiple references $\mathcal{R}$. It poses a great challenge to avoid conflicts. We propose a *Multi-Reference Generation* mechanism to address this issue.

Word Prediction: To predict the next token y_t, we first compute a generation probability over the fixed vocab $\mathcal{V}_{R_X}$ by a two-layer MLP f_{gen} :

$$\mathbf{p^{gen}_t} = softmax(f_{gen}(\mathbf{z_t}, \mathbf{c_t}, \mathbf{y_{t-1}})) \tag{13}$$

then, for each reference $R_j \in \mathcal{R}$, we compute a probability distribution to estimate the probability to copy a token from the corresponding reference:

$$\mathbf{p^{copy}_{R_j,t}} = f^{R_j}_{copy}([\mathbf{z_t}; \mathbf{c_t}; \mathbf{y_{t-1}}], \mathbf{R_j}) \tag{14}$$

where each $f^{R_j}_{copy}$ is a *General* attention function (Luong et al., 2015), $[\mathbf{z_t}; \mathbf{c_t}; \mathbf{y_{t-1}}]$ is the attention query, the encoded $\mathbf{R_j}$ serves as the attention key.

Dynamic Vocab: To eliminate conflicts brought by different word distributions of given references, a dynamic vocab $\mathcal{V}^d$ is built, which consists of all words that appear in both the reference set $\mathcal{R}$ and the fixed vocab $\mathcal{V}_{R_X}$[4]:

$$\mathcal{V}^d = \Phi(\mathcal{R}) \cup \mathcal{V}_{R_X} \tag{15}$$

Then, a projection matrix $\mathbf{M_{\mathcal{V}_{R_X}}} \in \mathbb{R}^{|\mathcal{V}^d| \times |\mathcal{V}_{R_X}|}$ to map the computed generation distribution $\mathbf{p^{gen}_t}$

[3] the shape of vectors/matrices is defined as $\mathbb{R}^{n\times 1}/\mathbb{R}^{n\times m}$

[4] where $\Phi(\cdot)$ outputs the token set.

to the dynamic vocab space. Similarly, for each copy distribution $\mathbf{P}^{\mathbf{copy}}_{\mathbf{R_j,t}}$ of reference R_j, we construct a projection matrix $\mathbf{M}_{\mathbf{R_j}} \in \mathbb{R}^{|\mathcal{V}^d| \times |R_j|}$, which maps the copying distribution of R_j to the dynamic vocab space.

Multi-Reference Generation: The probability of the next word y_t is given by infusing all distributions with a generation gate γ^{gen}_t and several copy gates $\gamma^{copy}_{R_j,t}$:

$$P_t(y_t) = \frac{\gamma^{gen}_t \mathbf{M}_{\mathcal{V}_{\mathbf{R_X}}} \mathbf{p}^{\mathbf{gen}}_{\mathbf{t}} + \sum_{R_j} \gamma^{copy}_{R_j,t} \mathbf{M}_{\mathbf{R_j}} \mathbf{p}^{\mathbf{copy}}_{\mathbf{R_j,t}}}{\gamma^{gen}_t + \sum_{R_j} \gamma^{copy}_{R_j,t}} \tag{16}$$

we not only use a mode weight $\alpha^{mode}_{*,t}$ to control the participation of each distribution, but also adopt the previous relevance gate $\alpha^{rel}_{R_j}$ to help the infusing of copy distributions:

$$\gamma^{gen}_t = \alpha^{mode}_{gen,t}, \ \ \gamma^{copy}_{R_j,t} = \alpha^{rel}_{R_j} \cdot \alpha^{mode}_{R_j,t} \tag{17}$$

where mode gates $\alpha^{mode}_{gen,t} = \mathbf{a^m_t}[0]$ and $\alpha^{mode}_{R_j,t} = \mathbf{a^m_t}[j+1]$ are given by:

$$\mathbf{a^m_t} \in \mathbb{R}^{1+|\mathcal{R}|} = softmax(\mathbf{W^M}[\mathbf{z_t}; \mathbf{c_t}; \mathbf{y_{t-1}}]) \tag{18}$$

Training: Finally, $P_t \in \mathbb{R}^{|\mathcal{V}^d|}$ can be used to predict the next token. The model can subsequently be optimized by minimizing the following negative log-likelihood:

$$\mathcal{L} = -\sum_t \log P_t(y_t | y_{1:t-1}, \mathcal{R}) \tag{19}$$

3 Experiment

3.1 Experiment Methodology

Dataset: It is built upon three open-released Chinese Weibo corpora (Shang et al., 2015; Ke et al., 2018; Cai et al., 2019). We adopt a ConceptNet (Speer et al., 2017) base released by (Wu et al., 2020b) as the commonsense knowledge. It contains about 696K triples, 27K entities, and 26 relations. For the text knowledge, we collect introduction paragraphs of 1,663K entities from Chinese Wikipedia. Besides, we also collect infobox tables of 1,581K entities from Chinese Wikipedia. All texts are tokenized by Jieba [5]. Following (Wu et al., 2020b), entity words $\in R_X$ are used as queries to retrieve knowledge queries from knowledge bases. For each dialogue, we retrieve up to 200 most relevant[6] commonsense triplets, up to 1 relevant text paragraph, and up to 1 infobox table. After the pre-processing and the dialogue-knowledge alignment, the statistics have been reported in Table 1. As reported in Table 1, using all three knowledge sources can improve the coverage by 63∼200%.

Set:	Training	Validation	Test
#Total:	700,000	40,000	40,000
Commonsense (%)	48.6%	48.8%	48.8%
Text (%)	24.7%	24.2%	24.4%
Infobox (%)	26.9%	26.9%	27.0%
Any of them(%)	79.6%	79.8%	79.8%

Table 1: Dataset Statistics. (%) indicates the coverage. The coverage of commonsense/text/infobox in the raw three corpora is 14.9/7.6/8.3%. In this paper, we let about 80% of dialogues can be aligned with at least one type of knowledge.

Models: There are 5 groups: **(1) None**: the widely-used attentive *Seq2Seq* (Luong et al., 2015), and its variant *Copy* that can copy words from the query (See et al., 2017); **(2) Commonsense**: the first *CCM* leverages the commonsense knowledge with two graph attention mechanism (Zhou et al., 2018). The next *ConceptFlow* (Zhang et al., 2020) and *ConKADI* (Wu et al., 2020b) are two latest SOTA commonsense knowledge-enhanced methods; **(3) Text**: we use one of the latest SOTA text knowledge-enhanced methods *RefNet*, which proposes a reference-aware network to access the background text (Meng et al., 2020a); **(4) Infobox**: we adapt two data-to-text works to dialogue models by adding dialogue encoding/attention/copy modules (from *Copy*). The first *SA-S2S* proposes a structure-aware seq2seq to use the infobox knowledge (Liu et al., 2018), the next *TransInfo* is one of the latest SOTA infobox knowledge-aware text generation approach with a Transformer Encoder (Bai et al., 2020); **(5) Pre-training**: *CDial-GPT* (Wang et al., 2020b) proposes a GPT-based and a GPT2-based dialogue model (Radford et al., 2019), where GPT_{base} and $GPT2_{base}$ have been pre-trained on a Chinese Novel dataset (1.3B words) and about 6.8M dialogue sessions. Both *GPT*s have 95.5M parameters (*MSKE-Dialog* has 59.14M parameters), and are fine-tuned on our dataset. The implementation details have been listed in Appendix A. The code is open released (https://github.com/pku-sixing/EMNLP2021-MSKE_Dialog).

Metrics: We use the embedding-based Embed-A/G/X (Average/Greedy/Extreme (Liu et al.,

[5] https://pypi.python.org/pypi/jieba/

[6] here, relevance is defined as the word overlap between the response and the candidate knowledge entry.

Aspects	Relevance								Diversity			Knowledge				Overall ↑
Metrics	Embed			ROUGE	BLEU				DIST		Ent	Entity				Mean
Configs	A	G	X	L	1	2	3	4	Uni	Bi	4	CSK	TXT	IBT	AVG	Geo.
Seq2Seq	0.825	0.686	0.630	11.68	12.79	4.60	1.82	0.76	1.31	6.75	7.44	0.35	0.16	0.14	0.21	1.00
Copy	0.822	0.688	0.629	12.44	13.49	5.13	2.13	0.91	3.64	15.00	8.24	0.30	0.21	0.17	0.23	1.23
CCM	0.840	0.697	0.635	13.03	14.16	4.97	1.98	0.82	1.42	9.01	8.88	**0.53**	0.18	0.17	0.29	1.16
ConceptFlow	0.845	0.696	0.637	12.82	14.95	5.10	2.00	0.84	1.56	9.89	8.90	0.34	0.15	0.15	0.21	1.10
ConKADI	0.844	0.683	0.630	13.35	15.62	5.61	2.33	1.03	3.53	19.21	10.94	0.50	0.23	0.20	0.31	1.43
SA-S2S	0.824	0.690	0.636	12.83	14.24	5.42	2.26	0.99	3.22	12.70	7.77	0.30	0.20	0.24	0.25	1.24
TransInfo	0.825	0.689	0.638	13.16	14.18	5.45	2.26	1.01	3.78	15.34	8.38	0.29	0.22	0.248	0.25	1.29
RefNet	0.829	0.682	0.622	11.92	14.25	4.67	1.62	0.59	2.75	14.53	10.16	0.42	**0.48**	0.17	0.359	1.32
GPT_{base}	0.836	0.678	0.631	12.88	15.03	5.96	2.86	1.56	5.07	23.97	**11.03**	0.41	0.25	0.21	0.29	1.52
$GPT2_{base}$	0.833	0.680	0.630	12.71	14.75	5.71	2.67	1.40	4.16	20.07	10.77	0.38	0.23	0.19	0.27	1.42
MSKE-Dialog	**0.854**	**0.700**	**0.653**	**16.14**	**15.73**	**6.82**	**3.40**	**1.92**	**6.04**	**27.50**	10.82	0.47	0.36	**0.253**	**0.363**	**1.72**

Table 2: Automatic evaluation results. **Score** means a model outperforms others in the corresponding metric.

2016)) and the word overlap-based BLEU1-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and to evaluate the relevance to the ground-truth responses. Following (Zhang et al., 2018), we use DIST-Uni/Bi (the ratio of distinct 1/2-grams among all generated tokens), and the 4-gram entropy Ent4 to evaluate the diversity. In addition, we use the entity score (i.e., the number of the generated entity/knowledge words per sentence) to measure the knowledge utilization. We count the entity score on each type of knowledge (CSK, TXT, IBT : commonsense, text, infobox), and compute the averaged entity score (AVG). Finally, to fairly compare the overall performance, we report the overall geometric mean scores relative to *Seq2Seq*, Appendix B has elaborated the detail. When comparing different approaches, we do not use the $perplexity$, because the definitions and computations vary among approaches. We will report the $perplexity$ in the ablation study, because all model variants share the same computation.

3.2 Experimental Results

Automatic Evaluation: As reported in Table 2, *MSKE-Dialog* wins the best in 12 metrics, wins second/third place in 3 metrics, and the best overall performance. In the aspect of relevance, *MSKE-Dialog* beats baselines in all related metrics, indicating responses generated by *MSKE-Dialog* are closer to the topic. Thanks to the proposed *Multi-Reference Generation* mechanism, *MSKE-Dialog* has the best performance in DIST-Uni/Bi and the second-best performance in Ent-4, showing *MSKE-Dialog* can generate diverse and informative responses. *MSKE-Dialog* slightly loses to GPT_{base} in Ent-4, we think the reason is GPT_{base} has already been pre-trained by a large amount of dialogues. Moving to the aspect of knowledge, *MSKE-Dialog* undoubtedly beats other baselines in the overall score with the cooperation of three heterogeneous knowledge sources. *MSKE-Dialog* loses to *RefNet*/*CCM* in terms of text/commonsense entity score. The reason is such two baselines only use one knowledge source, but our approach uses three sources; thus, our approach would not only focus on using one source.

Aspects	Appropriateness			Informativeness		
vs.	Win	Tie	Loss	Win	Tie	Loss
Seq2Seq	**56.0%**	13.8%	30.2 %	**55.5%**	12.8%	31.7%
TransInfo	**58.2%**	15.5%	26.3%	**58.2%**	13.8%	28.0%
ResNet	**60.3%**	11.5%	28.2%	**58.3%**	10.2%	31.5%
ConKADI	**49.8%**	14.3%	35.8%	**46.0%**	14.7%	39.3%
GPT_{base}	**47.0%**	16.8%	36.2%	**45.2%**	16.7%	38.2%

Table 3: Human annotation results. **Score** means our approach significantly outperforms baselines (sign test, p-value < 0.05, ties are removed). The agreement among volunteers have been reported in Appendix C .

Human Evaluation: We conduct the pair-wise evaluation. Baselines include *ConKADI*, *TransInfo*, *RefNet*, GPT_{base} (the best in the corresponding group), and the naive *Seq2Seq*. We employ 3 well-educated native speakers to annotate the sampled 200 test cases. There are two criteria: (1) Appropriateness evaluates the fluency, and the relevance to the context; (2) Informativeness evaluates how much new knowledge is provided.

As reported in Table 3, *MSKE-Dialog* can also outperform baselines in human evaluation. Compared with the automatic results, *Seq2Seq* has better performance in human evaluation. This is due to humans always have a high tolerance for boring/generic but fluent responses. The remaining results are roughly in line with the automatic results.

Geo. Aspect	Rel.	Diver.	Know.	Overall	Perplexity ↓
Base	1.08	1.39	1.19	1.18	94.89
Base+CSK	1.19	1.89	1.38	1.38	88.30
Base+TXT	1.19	1.86	1.52	1.41	89.27
Base+IBT	1.09	1.31	1.29	1.19	91.47
Context	1.21	3.22	1.33	1.58	90.14
Context+CSK	1.24	**3.35**	1.48	1.66	86.14
Context+TXT	1.22	3.10	1.43	1.60	88.69
Context+IBT	1.24	3.06	1.38	1.60	87.02
Full	**1.30**	3.01	**1.71**	**1.72**	**81.10**

Table 4: Knowledge Ablation Study. We report the geomean relative score for each aspect, the full results have also been reported in Appendix D. Both *Base* and *Context* do not use external knowledge, but *Context* can copy words from the context X. *+CSK/TXT/IBT* uses the commonsense/text/infobox knowledge.

Geo. Aspect	Rel.	Diver.	Know.	Overall	Perplexity↓
Full	**1.30**	3.01	1.71	**1.72**	**81.10**
-$\mathcal{V}^d$	1.21	2.58	**1.86**	1.63	86.46
-Multi R.Gen.	1.23	1.36	1.33	1.29	85.77
-Multi R.Select.	1.22	2.50	1.81	1.61	82.87
-K.Copy	1.28	2.92	1.52	1.64	84.12
-K.Attn	1.25	**3.23**	1.62	1.69	82.94

Table 5: Model Ablation Study. *-*$\mathcal{V}^d$ does not build the dynamic vocab. *-Multi R.Gen.* does not use the Multi-Reference Generation; thus, all words are predicted based on the generation probability. *-Multi R.Select.* does not use the Multi-Reference Selection to control the participation of each generation reference. *-K.Copy/Read* does not copy/read external knowledge (copying/reading the context X is still allowed). The full results have also been reported in Appendix E

It is worth noting that, *MSKE-Dialog* can outperform GPT$_{base}$, while *MSKE-Dialog* only uses 62% of parameters (59.14M vs. 95.5M) and less than 10% of training data (700K vs.1.3B words+6.8M pre-training+ 700K fine-tune). It verifies the advantage of using multi-source heterogeneous knowledge and the effectiveness of our model.

3.3 Ablation Analysis

To investigate what makes the most contribution to *MSKE-Dialog*, we conduct extensive studies.

Knowledge Contribution: We design a set of single-source variants of *MSKE-Dialog* to explore which knowledge brings the most improvement. As reported in Table 4, compared to *Base*, which neither uses external knowledge nor copies word, all three single-source variants have improvements in both the overall performance and the perplexity. Previous works (Gu et al., 2016; Vinyals et al., 2015) have shown that copying words from the context R_X can significantly improve the performance. Our models can also benefit from this factor, the *Context+** outperforms *Base+** by notable margins. Evidently, among the three knowledge sources, commonsense knowledge and text knowledge bring more contributions. The perplexity of *Context/Base+IBT* have notable improvements, but the improvement of the overall score (i.e., the quality of the generated responses) is not notable [7]. We guess the employed beam-search decoding may be a bottleneck, we leave it as future work.

It is worth noting that our approach can also beat the best knowledge-enhanced baselines without using more source knowledge sources. The best commonsense/text/infobox knowledge-enhanced baselines and their overall scores are *CinKADI*/1.43, *RefNet*/1.32, *TransInfor*/1.29, which are lower than our single-source variants: *Context+CSK*/1.66, *Context+TXT*/1.60, *Context+IBT*/1.60. It proves that our *MSKE-Dialog* not only has the ability to use multi-source heterogeneous knowledge, but also has more efficiency in model design.

Model Contribution In this part, all variants use all three knowledge sources. We check the performance contribution by removing a module from the *Full* model, namely, *MSKE-Dialog*. As reported in Table 5, we first remove the use of dynamic vocab $\mathcal{V}^d$. While the knowledge score increases, the relevance score and the diversity score sharply decrease. This is due to *-*$\mathcal{V}^d$ tends to copy words from external knowledge without considering the context. Meanwhile, we propose a Multi-Reference Selection mechanism to solve the topic conflict and propose a Multi-Reference Generation mechanism to generate informative responses without the impact of generation conflict. *-Multi R.Gen./R.Select* prove such two mechanisms are effective, especially the Multi-Reference Generation. Comparing *-K.Copy* and *-K.Attn*, *-K.Copy* has more degeneration, indicating copying knowledge words brings more improvements.

Full does not achieve the best in each aspect but has the best overall performance and perplexity, which indicates using multi-source knowledge is quite challenging. It is crucial to fuse the knowledge sources into the context without the impact of the possible conflicts.

[7] Perplexity reflects the difficulty to generate the ground-truth, lower is better.

Query#1:	突然想和女朋友分手了。 I suddenly want to break up with my girlfriend.
ConKADI	突然想和女朋友分手了... I suddenly want to break up with my girlfriend...
GPT_{base}	分手快乐 祝你快乐。 Happy breakup, I wish you happiness
MSKE-Dialog	我也突然想和[男友]分手。 I also suddenly want to break up with my **boyfriend**.
Query#2:	怎样打好篮球? How to play basketball well?
ConKADI	打好[前锋] 打[羽毛球] 打[羽毛球]. To be a good **playmaker**, play **badminton**, play **badminton**.
GPT_{base}	打篮球 打篮球 打篮球 打篮球 打篮球 打篮球。 Play basketball, play basketball, play basketball.
MSKE-Dialog	多练习[运球]。 Practice **dribbling** more.
Query#3:	大家都用什么防晒霜? What kind of sunscreen cream do you use?
ConKADI	[ZA]的。防晒的时候用防晒霜。 I use **ZA**('s product). Use sunscreen when you need it
GPT_{base}	我用的是[雅漾]的,感觉还不错,不过不是很[油]。 I use **Avene**('s product). It feels good, but it's not **oily**.
MSKE-Dialog	[露得清]的防晒效果不错。 **Neutrogena**'s sunscreen cream has a good effect.

Table 6: Some cases generated by models. **entity**/[实体词] means the new information (new knowledge word).

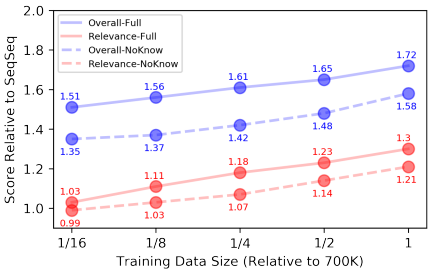

Figure 3: Evaluations on different data sizes. We test the full model and a variant that does not use multi-source knowledge. We report the overall score and the geomean relative score in the aspect of relevance. Compared to the diversity/knowledge, relevance is a more representative aspect in the low-resource evaluation.

3.4 More Studies

Low Resource Evaluation: We train *MSKE-Dialog* and a non-knowledge-enhanced variant on only a part of the dataset. As illustrated in Figure 3, with the incorporation of multi-source knowledge, *MSKE-Dialog* with only $\frac{1}{2} \sim \frac{1}{4}$ conversational data can archive comparable performance with the non-knowledge-enhanced variant. It indicates the multi-source knowledge can indeed help the dialogue generation if the conversational data is not enough. This can be quite useful when constructing a system in a low-resource language/scenario.

Case Study: We show three cases generated by our *MSKE-Dialog* and two better baselines in the human evaluation in Table 6. In Case #1, only *MSKE-Dialog* provided the new information, demonstrating our multi-source heterogeneous knowledge-enhanced approach is able to generate more informative responses with the improved knowledge coverage. In the next Case #2, although *ConKADI* also provided new information, it failed to generate a fluent response. It indicates it is crucial to alleviate the conflict between knowledge and context. In the last Case #3, although all three models have generated fluent and informative responses, GPT_{base} generated a more natural response and brought more information, which can be attributed to GPT_{base} was trained by more training data. It tells us the potential to investigate the pre-training methods; we leave it as future work. This work only focuses on the non-pre-training method because pre-training models have expensive costs in training/using.

4 Related Work

The vanilla Seq2Seq tends to generate generic responses, such as 'I don't know' (Chen et al., 2017). Many efforts have been devoted to diversifying the generations (Li et al., 2016; Gao et al., 2019), etc. One crucial factor leading to this issue is the lack of sufficient knowledge. During the conversation, the vanilla Seq2Seq model can only access the given query, which only contains limited knowledge (Ghazvininejad et al., 2018). The insufficient knowledge makes it hard for a model to understand the context and generate an informative response. To this end, knowledge-enhanced approaches have been proposed and demonstrated promising performances (Yu et al., 2020). The knowledge can be texts (Ren et al., 2020; Zhao et al., 2020; Kim et al., 2020; Tam, 2020), the structured graphs/tables/bases (Zhou et al., 2018; Qin et al., 2019; Wu et al., 2020b,c; Zhang et al., 2020), the semi-structured infobox (Wu et al., 2021), the pre-trained models (Devlin et al., 2019; Radford et al., 2019; Moghe et al., 2020), and many other external knowledge components (Wang et al., 2018; Xu et al., 2019).

However, most previous works can only use single-source homogeneous knowledge. Solely re-

lying on only one type of knowledge greatly limits the performance in the real scenario. Some previous works have also noticed this issue. For example, augmenting the knowledge graph with an external text comprehension module (Liu et al., 2019) or a KBQA module (Wang et al., 2020a), introducing multi-modal visual features (Liang et al., 2020) for emotional conversation or visual conversation (Meng et al., 2020b). Our work is different from them because we focus on the open-domain knowledge-enhanced dialogue response generation, rather than the emotional/visual conversation, etc. To our best knowledge, few works have studied this topic in this area. In addition, *MSKE-Dialog* has a salable framework. A new knowledge source can easily be integrated by simply adding a knowledge encoder.

5 Conclusion & Future Work

This paper proposes a novel multi-source heterogeneous knowledge-enhanced dialogue generation approach, *MSKE-Dialog*, which outperforms competitive knowledge-enhanced baselines and pre-training models. It verifies the advantages of using multi-source heterogeneous knowledge and the advantages of our approach.

We will continue to investigate the advantages of knowledge-enhanced dialogue generation. We notice the current decoding strategy may be a bottleneck of knowledge-enhanced works and the potential of multi-source knowledge + pre-training. We will also pay more attention to such topics.

Acknowledgements

This work is partly supported by ICBC Technology.

Ethical Considerations

In this work, the dataset involves both conversational data and knowledge data. All three involved Chinese Weibo (weibo.com, an open SNS in China) conversational datasets are open-released by previous works for research (Shang et al., 2015; Ke et al., 2018; Cai et al., 2019). Including but not limited to the involved three datasets, conversational data crawled from Weibo are widely used in training/evaluating in the research of Chinese dialogue generation research and other NLP researches, for example, (Wang et al., 2020b; Su et al., 2020). All data crawled from Weibo are open-accessed posts/responses that everyone can see; no privacy-related data (such as gender, nickname, birthday, etc.) are used. But if it needs commercial use, it may need to ask for additional permission from the original author/copyright owner. We use the commonsense knowledge from ConceptNet (Speer et al., 2017); according to its description, it is allowed to reuse them in research (see conceptnet.io). We also collect text knowledge and infobox knowledge from Wikipedia (under the license *CC BY-SA 3.0*); it is allowed to reuse them in both research and commercial. To summary, as research work, this work has no concern on the dataset and other aspects.

References

Yang Bai, Ziran Li, Ning Ding, Ying Shen, and Hai-Tao Zheng. 2020. Infobox-to-text generation with tree-like planning based attention network. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3773–3779.

Antoine Bordes, Nicolas Usunier, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-Relational Data. In *NIPS*, pages 2787–2795.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1866–1875.

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

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (ELUs). In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN,*

USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. 2019. A discrete CVAE for response generation on short-text conversation. In *EMNLP*, pages 1898–1908.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*, pages 5110–5117.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning.

Wenpeng Hu, Ran Le, Bing Liu, Jinwen Ma, Dongyan Zhao, and Rui Yan. 2020. Translation vs. dialogue: A comparative analysis of sequence-to-sequence modeling. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4111–4122, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan. 2018. Generating Informative Responses with Controlled Sentence Function. *Proceedings of ACL*, pages 1499–1508.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*, pages 110–119.

Yunlong Liang, Fandong Meng, Ying Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2020. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. *CoRR*, abs/2012.04882.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. 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*, pages 41–52.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pages 2122–2132.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4881–4888.

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*, pages 1412–1421.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020a. Refnet: A reference-aware network for background based conversation. pages 8496–8503.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020b. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *CoRR*, abs/2012.15015.

Nikita Moghe, Priyesh Vijayan, Balaraman Ravindran, and Mitesh M. Khapra. 2020. On incorporating structural information to improve dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 11–24, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with KB retriever. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 133–142. Association for Computational Linguistics.

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, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local

knowledge selection for background based conversation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8697–8704.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1577–1586.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 175–180. Association for Computational Linguistics.

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

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7087–7097. Association for Computational Linguistics.

Yik-Cheung Tam. 2020. Cluster-based beam search for pointer-generator chatbot grounded by knowledge. *Comput. Speech Lang.*, 64:101094.

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.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020a. Improving knowledge-aware dialogue generation via knowledge base question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9169–9176. AAAI Press.

Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. 2018. Chat More: Deepening and Widening the Chatting Topic via A Deep Model. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*, pages 255–264, New York, New York, USA. ACM Press.

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

Sixing Wu, Ying Li, Dawei Zhang, and Zhonghai Wu. 2020a. Improving knowledge-aware dialogue response generation by using human-written prototype dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 1402–1411.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020b. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020c. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3766–3772.

Sixing Wu, Minghui Wang, Dawei Zhang, Yang Zhou, Ying Li, and Zhonghai Wu. 2021. Knowledge-aware dialogue generation via hierarchical infobox accessing and infobox-dialogue interaction graph network. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3964–3970. ijcai.org.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. In *ACL*, pages 5416–5426.

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.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 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, pages 2031–2043.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1815–1825.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

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.

A Model Implementations

We re-implement Seq2Seq, Copy, SA-S2S, and TransInfo by using the PyTorch, and the remaining use the official implementations and decoding strategies.

Hyper-Parameters : The word embedding dimension is 200, the commonsense entity/relation dimension is 100, the GRU dimension is 512. We use the Adam optimizer with the initial learning rate of 0.0001, and the batch size of 32. After each training epoch, we will check a model's performance (perplexity) on the validation set, if the perplexity starts to increase, the learning rate will be halved; if the epoch number reaches 20 or the perplexity increases in two successive epochs, the training will be stopped. During the inference, we select the hyper-parameter of the lowest perplexity on the validation set. The official implementations of CCM, ConceptFlow, RefNet, and GPTs use the greedy search decoding and do not support beam-search; thus, we keep the official settings. For the remaining approaches, we apply the beam-search decoding strategy (beam width $= 10$). Under such settings, the training on a single Nvidia Geforce RTX Titan roughly costs 2 days. In addition, we use a pre-trained Chinese word embedding released (Song et al., 2018) to initialize (if support) and evaluate.

B Overall Score

We evaluate models with more than 10 metrics, it is confusing to judge the overall performance by only checking the different discrete scores. For comparing the overall performance, we report the overall geometric mean relative scores. The performance baseline is Seq2Seq (i.e., its relative score is defined as 1.0).

In detail, each metric is defined as $M_{i,j,k}$, where i is the index of the evaluation aspects (*Relevance*, *Diversity*, and *Knowledge*), j is the index of the evaluation method in the i-th aspect (for example, the aspect *Relevance* includes *Embed*, *ROUGE*, and *BLEU*), the last k indicate the specific metric variant of $M_{i,j}$ (for example, the embedding-based metric Embed has three settings: average, greedy, extreme.). Subsequently, the computation of the overall geometric relative score can be described as:

- 1. For each $M_{i,j,k}$, we first compute the performance rate relative to Seq2Seq. For example, if Seq2Seq achieves 10.0 in terms of the metric $M_{i,j,k}$, and *MSKE-Dialog* achieves 15.0; then, the relative performance rate of *MSKE-Dialog* is 1.50. The relative performance rate of $M_{i,j,k}$ is denoted as $R_{i,j,k}$.
- 2. Each evaluation method $M_{i,j}$ may have different metric variants, but the number of metric variants should not affect the overall score, so for each evaluation method $M_{i,j} \in$ (Embed, ROUGE, BLEU, DIST), we compute the geometric mean relative score among its variants: $R_{i,j} = GeoMean(\{R_{i,j,k}\})$. Meanwhile, in this part, we use the averaged entity score AVG instead of the geomean of three sub entity scores.
- 3. For each aspect M_i, we compute the geometric mean relative score among its evaluation methods: $R_i = GeoMean(\{R_{i,j}\})$.
- 4. The overall geometric mean relative score is given by: $R = GeoMean(\{R_i\})$.

The computed R can be used to compare different models easily.

C Human Evaluation

Following (Wu et al., 2020b), we conduct the pair-wise evaluation. The competitors include *ConKADI*, *TransInfo*, *RefNet*, *GPT*$_{base}$ (the best baselines in the corresponding group), and the widely-used *Seq2Seq*. We employ three well-educated native speakers to annotate the sampled 200 test cases (1,200 pairs in total). There are two criteria: (1) Appropriateness evaluates the relevance to the dialogue context, and fluency; (2) Informativeness evaluates how much new knowledge is provided in a generated response.

Aspects	Appropriateness			Informativeness		
vs.	Win	Tie	Loss	Win	Tie	Loss
Seq2Seq	**56.0%**	13.8%	30.2 %	**55.5%**	12.8%	31.7%
TransInfo	**58.2%**	15.5%	26.3%	**58.2%**	13.8%	28.0%
ResNet	**60.3%**	11.5%	28.2%	**58.3%**	10.2%	31.5%
ConKADI	**49.8%**	14.3%	35.8%	**46.0%**	14.7%	39.3%
GPT$_{base}$	**47.0%**	16.8%	36.2%	**45.2%**	16.7%	38.2%

Table 7: Human annotation results. **Score** means our approach significantly outperforms baselines (sign test, p-value < 0.05, ties are removed).

As reported in Table 7, *MSKE-Dialog* can also outperform baselines in human evaluation. Compared with the automatic results, *Seq2Seq* has better performance in human evaluation. This is due

Aspects	Relevance								Diversity			Knowledge				Overall	
Metrics	Embed			ROUGE	BLEU				DIST		Ent	Entity				Mean	PPL↓
Configs	A	G	X	L	1	2	3	4	Uni	Bi	4	CSK	TXT	IBT	AVG	Geo.	-
Base	0.851	0.696	0.641	13.97	14.44	5.19	2.06	0.88	1.74	10.31	9.87	0.41	0.19	0.17	0.25	1.18	94.87
Base+CSK	0.853	0.696	0.645	14.72	14.97	5.96	2.66	1.32	2.46	17.16	10.44	0.47	0.22	0.18	0.29	1.38	88.30
Base+TXT	0.856	0.700	0.649	14.78	15.02	5.97	2.65	1.30	2.67	15.56	10.16	0.36	0.40	0.19	0.32	1.41	89.27
Base+IBT	0.854	0.700	0.647	14.43	14.75	5.28	2.07	0.86	1.81	8.57	9.54	0.40	0.18	0.24	0.27	1.19	91.47
Context	0.846	0.689	0.641	14.68	14.30	6.02	2.92	1.60	6.81	29.97	10.80	0.35	0.28	0.21	0.28	1.58	90.14
Context+CSK	0.846	0.689	0.642	15.03	14.57	6.27	3.12	1.77	7.04	31.94	11.03	0.43	0.28	0.22	0.31	1.66	86.14
Context+TXT	0.848	0.693	0.645	14.94	14.25	6.10	2.99	1.66	6.47	28.06	10.77	0.33	0.36	0.22	0.30	1.60	88.69
Context+IBT	0.850	0.694	0.647	15.23	14.73	6.25	3.09	1.73	6.52	27.53	10.52	0.34	0.27	0.25	0.29	1.60	87.02
Full	0.854	0.700	0.653	16.14	15.73	6.82	3.40	1.92	6.04	27.50	10.82	0.47	0.36	0.25	0.36	**1.72**	**81.10**

Table 8: Knowledge Ablation Study. Both *Base* and *Context* do not use external knowledge, but *Context* can copy words from the context X. *+CSK/TXT/IBT* means using the commonsense/text/infobox knowledge, respectively.

Aspects	Relevance								Diversity			Knowledge				Overall	
Metrics	Embed			ROUGE	BLEU				DIST		Ent	Entity				Mean	PPL↓
Configs	A	G	X	L	1	2	3	4	Uni	Bi	4	CSK	TXT	IBT	AVG	Geo.	-
Full	0.854	0.700	0.653	16.14	15.73	6.82	3.40	1.92	6.04	27.50	10.82	0.47	0.36	0.25	0.36	**1.72**	**81.10**
-$\mathcal{V}^d$	0.857	0.705	0.652	16.02	16.22	6.29	2.68	1.23	5.15	20.96	10.44	0.57	0.36	0.25	0.39	1.63	86.46
-Multi R.Gen.	0.840	0.700	0.650	13.97	15.52	6.44	3.02	1.55	1.78	11.37	8.23	0.33	0.28	0.23	0.28	1.29	85.77
-Multi K.Select.	0.856	0.705	0.655	16.05	15.96	6.37	2.78	1.32	4.87	20.40	10.40	0.50	0.38	0.25	0.38	1.61	82.87
-K.Copy	0.853	0.699	0.652	15.85	15.33	6.58	3.27	1.85	5.94	25.86	10.64	0.39	0.33	0.25	0.32	1.64	84.12
-K.Read	0.851	0.696	0.648	15.58	15.16	6.42	3.10	1.70	6.69	30.43	10.89	0.43	0.35	0.26	0.34	1.69	82.94

Table 9: Model Ablation Study. *-$\mathcal{V}^d$* does not construct the dynamic vocab. *-Multi R.Gen.* does not use the Multi-Reference Generation; thus, all words are predicted based on the generation probability. *-Multi K.Select.* does not use the Multi-Reference Selection to control the participation of each generation reference. *-K.Copy/Read* does not copy/read external knowledge (copying/reading the context X is still allowed).

to humans always have a high tolerance for boring/generic but fluent responses. The remaining results are roughly in line with the automatic results. It is worth noting that, compared to GPT$_{base}$, although only using 62% of parameters (59.14M vs. 95.5M) and less than 10% of training data (700K vs.1.3B Words+6.8M pre-training+ 700K fine-tune), *MSKE-Dialog* still can outperform GPT$_{base}$. It demonstrates the advantage of using multi-source heterogeneous knowledge and the effectiveness of our model design.

Following (Wu et al., 2020b), We count the agreement among volunteers, for the appropriateness, 2/3 agreement (the percentage of the cases that at least 2 volunteers give the same label) is 94.2%, the 3/3 agreement is 53.7%; for the informativeness, 2/3 agreement is 94.4%, the 3/3 agreement is 52.4%.

D Knowledge Contribution

We design a set of single-source variants of *MSKE-Dialog* to explore which knowledge brings the most improvement. As reported in Table 8, compared to *Base*, which neither uses external knowledge nor copies word, all three single-source variants have improvements in both the overall performance and the perplexity. Previous works (Gu et al., 2016; Vinyals et al., 2015) have shown that copying words from the context R_X can significantly improve the performance. Our models can also benefit from this factor, the *Context+** outperforms *Base+** by notable margins. Evidently, among the three knowledge sources, commonsense knowledge and text knowledge bring more contributions. The perplexity of *Context/Base+IBT* have notable improvements, but the improvement of overall score (i.e., the quality of the generated responses) is not notable [8]. We guess the employed beam-search decoding may be a bottleneck, we leave it as future work.

It is worth noting that our approach can also beat the best knowledge-enhanced baselines without using more source knowledge sources. The best commonsense/text/infobox knowledge-enhanced baselines and their overall scores are *CinKADI*/1.43, *RefNet*/1.32, *TransInfor*/1.29, which are lower than our single-source variants: *Context+CSK*/1.66, *Context+TXT*/1.60, *Context+IBT*/1.60. It proves

[8] Perplexity reflects the difficulty to generate the ground-truth, lower is better.

that our *MSKE-Dialog* not only has the ability to use multi-source heterogeneous knowledge, but also has more efficiency in model design.

E Model Contribution

In this part, all variants use all three knowledge sources. We check the performance contribution by removing a module from the *Full* model, namely, *MSKE-Dialog*. As reported in Table 9, we first remove the use of dynamic vocab $\mathcal{V}^d$. While the knowledge score increases, the relevance score and the diversity score sharply decrease. This is due to -$\mathcal{V}^d$ tend to copy words from external knowledge without considering the context. We propose a Multi-Reference Selection mechanism to solve the topic conflict and propose a Multi-Reference Generation mechanism to generate informative responses without the impact of generation conflict. *-Multi R.Gen./R.Select* prove such two mechanisms are effective, especially the Multi-Reference Generation. Comparing *-K.Copy* and *-K.Attn*, *-K.Copy* has more regression, indicating copying knowledge words brings more improvements.

Full does not achieve the best in each aspect but has the best overall performance and perplexity, which indicates using multi-source knowledge is quite challenging. It is crucial to fuse the knowledge sources into the context without the impact of the possible conflicts.