KSAM: Infusing Multi-Source Knowledge into Dialogue Generation via Knowledge Source Aware Multi-Head Decoding

Sixing Wu¹, Ying Li²,³, Dawei Zhang¹, Zhonghai Wu²,³
¹School of Computer Science, Peking University, Beijing, China
²National Research Center of Software Engineering, Peking University, Beijing, China
³Key Lab of High Confidence Software Technologies (MOE), Peking University, Beijing, China

Abstract

Knowledge-enhanced methods have bridged the gap between human beings and machines in generating dialogue responses. However, most previous works solely seek knowledge from a single source, and thus they often fail to obtain available knowledge because of the insufficient coverage of a single knowledge source. To this end, infusing knowledge from multiple sources becomes a trend. This paper proposes a novel approach Knowledge Source Aware Multi-Head Decoding, KSAM, to infuse multi-source knowledge into dialogue generation more efficiently. Rather than following the traditional single decoder paradigm, KSAM uses multiple independent source-aware decoder heads to alleviate three challenging problems in infusing multi-source knowledge, namely, the diversity among different knowledge sources, the indefinite knowledge alignment issue, and the insufficient flexibility/scalability in knowledge usage. Experiments on a Chinese multi-source knowledge-aligned dataset demonstrate the superior performance of KSAM against various competitive approaches.

1 Introduction

Conversational AIs play an indispensable role in the human-computer interaction (Chen et al., 2017). Humans can use their learned knowledge to understand the context, reason the intrinsic semantic, and generate informative responses. However, traditional dialogue generation methods can only use dialogue history that carries limited knowledge to generate responses (Sutskever et al., 2014), bringing meaningless responses and frustrating user experience (Li et al., 2016; Ghazvininejad et al., 2018). To bridge such a gap, incorporating external knowledge into the dialogue generation is a feasible way (Zhou et al., 2018).

Compared to the traditional non-knowledge-enhanced methods, the advantages of knowledge-enhanced methods come from the adopted external knowledge source (Wu et al., 2022). If a knowledge-enhanced model fails to seek available knowledge from the given knowledge source, it can only degenerate into a traditional manner. However, most previous works (Zhang et al., 2020a; Yu et al., 2020) only seek knowledge from a single source. The knowledge coverage ¹ of a single knowledge source is always insufficient (Wu et al., 2021a); thus, dialogues often can not benefit from the given knowledge source. Meanwhile, a single knowledge source is also difficult to meet the various requirements in the real scenarios (Liu et al., 2019). Recently, researchers began to seek knowledge from multiple sources to alleviate such issues. GOKC generates dialogues conditioned on both the background knowledge and the goal knowledge (Bai et al., 2021); The recent MSKE leverages heterogeneous knowledge from multiple sources (Wu et al., 2021a). With more knowledge sources, they have successfully improved the performance of knowledge-enhanced dialogue generation.

Nonetheless, as illustrated in Figure 1, many challenges in infusing multi-source knowledge into

¹In other words, how many dialogues can be aligned to a knowledge source.
dialogue generation have not been well solved: 1) Knowledge Diversity: Notable differences inevitably appear among different types of knowledge sources, which can be attributed to the different structures (i.e., text knowledge (Dinan et al., 2019) vs. commonsense knowledge graph (Zhang et al., 2020a)), different domains (i.e., open-domain (Speer et al., 2017) vs. specific-domain (Liu et al., 2018a)), and many other factors (Yu et al., 2020). Previous works only considered the difference in the encoding stage by using different knowledge-specific encoders, but failed to handle the difference in the decoding stage; 2) Indefinite Alignment: Due to the limitation of knowledge coverage, a single dialogue usually cannot fully use all n provided knowledge sources. Depending on the situation, each case may use an arbitrary number of sources, bringing more complexities; 3) Insufficient Flexibility and Scalability: A model itself should not be limited to a knowledge combination of specific types and a specific amount.

This paper proposes a KSAM (Knowledge Source Aware Multi-Head Decoding) approach for multi-source knowledge-enhanced dialogue generation, which explicitly considers the three challenges mentioned above. Besides the dialogue history, KSAM uses three different knowledge sources, i.e., plain text knowledge, commonsense fact knowledge, and table attribute knowledge, to generate the target response. We propose four Source-Specific Encoders to encode such four input sources\(^2\). In the decoding stage, unlike previous works that only adopt a single-head decoder, we assign an independent Source-Aware Decoder Head for each input source. Each decoder head is a source-aware and fully functional decoder network, generating a source-aware response independently. Thus, we can handle the differences among multiple sources by tuning the source-specific encoder or the source-aware decoder head without impacting other encoders/heads. Subsequently, we propose a Source Fusion Network (SFN) to make the final prediction by collecting and fusing the outputs from decoder heads. With source-aware decoder heads and the fusion gates outputted by SFN, KSAM can alleviate the issue of indefinite knowledge-alignment. Meanwhile, SFN does not limit the number of decoder heads or the type of a decoder head; thus, KSAM theoretically supports the use of any combination of knowledge sources.

We evaluate KSAM and baseline models on a previously released Chinese dataset (Wu et al., 2021a), which is aligned to three knowledge sources, i.e., a plain text knowledge base, a commonsense knowledge base, and a table knowledge base. Both the automatic and human evaluation results demonstrate the superior performance of KSAM against various competitive baselines. We also conduct extensive experiments to analyze KSAM further.

2 Approach

2.1 Problem Statement and Overview

The goal is to generate a response \(Y\) conditioned the dialogue history \(X\) and a set of knowledge \(\{K_i\}\). Each \(X = (x_1, \cdots, x_l)\) \(Y = (y_1, \cdots, y_T)\) is a word sequence; each knowledge \(K_i = (k_{i,1}, \cdots, k_{i,l_{K_i}})\) is a set/list of entries that are retrieved from the \(i\)-th knowledge source.

As illustrated in Figure 2, this paper proposes a novel Knowledge Source Aware Multi-Head Decoding approach (KSAM), which consists of three parts: 1) Source-Specific Encoders: We propose a history encoder \(Enc_X\) and several knowledge encoders \(\{Enc_{K_i}\}\) to encode the \(X\) and \(K\) into \(H_X\) and \(\{K_i\}\); 2) Source-Aware Decoder Heads: For alleviating the interference among sources and improving the scalability, each \(X\) or \(K_i\) has an independent and fully functional decoder head \(Dec_{H/K_i}\); 3) Source Fusion Network: It stepwisely collects the predicted outputs from decoder heads and makes the final prediction.

2.2 Source-Specific Encoders

2.2.1 Dialogue History Encoder

Dialogue history encoder \(Enc_X\) aims at encoding the dialogue history \(X\) into hidden states; thus, a bi-directional GRU (\(g\) (Cho et al., 2014)) is adopted. At each time step \(t\), the forward/backward GRU reads \(x_t\) and the last state \(h_{t-1}^f/h_{t+1}^b\):

\[
h_t = [h_t^f; h_t^b] = [\vec{g} (x_t, h_{t-1}^f); \bar{g} (x_t, h_{t+1}^b)]
\]

(1)

where \(x\) is the word embedding of \(x\), \([\cdot; \cdot]\) is the concatenation. The result is \(H = (h_1, \cdots, h_N)\).

2.2.2 Knowledge Encoders

This paper studies three knowledge sources: plain text knowledge \(K_P\), commonsense fact knowledge \(K_C\), and table key-value attribute knowledge \(K_T\).
**Plain Text:** Each text \( K_P \) is a word sequence \((k_{P,1}, \cdots, k_{P,l_p})\); therefore, we embed \( K_P \) to \( K^e_P = (k^e_{P,1}, \cdots, k^e_{P,l_p}) \) using word embedding.

**Commonsense Facts:** Each \( K_C \) is a set of facts \( \{k_{C,i}\} \), where each \( k_{C,i} \) has a head entity \( e_{C,i,h} \), a relation \( e_{C,i,r} \), and a tail entity \( e_{C,i,t} \). Thus, \( k_{C,i} \) can be embed as \( k^e_{C,i} = [e_{C,i,h}; e_{C,i,r}; e_{C,i,t}] \) using embedding pretrained by TransE (Bordes et al., 2013) or other learning approaches. Finally, \( K_C \) is embed to a set of embedding, i.e., \( K^e_C = \{k^e_{C,i}\} \).

**Table Attributes:** Each table \( K_T \) is a set of key-value attribute pairs \( \{k_{T,i} = (a^k_i, a^v_i)\} \), where the key \( a^k_i \) is a word and the value \( a^v_i = (a^v_{i,1}, \cdots, a^v_{i,j}, \cdots) \) is a text sequence. Such a structure is inconvenient for the encoding. Thus, following (Wu et al., 2021a), \( K_T \) is first decomposed into to a set of key-word pairs \( \{k^{kw}_{T,i,j}\} \), and each key-word pair is embedded as:

\[
k^{kw}_{T,i,j} = [a^k_i; a^v_{i,j}; \text{pos}_{i,j}] \tag{2}
\]

where \( a^k_i \) is the word embedding of the \( i \)-th key, \( a^v_{i,j} \) is the word embedding of the \( j \)-th word of the \( i \)-th value, \( \text{pos}_{i,j} \) is the positional embedding to indicate the structure information. Finally, \( K_T \) is embed to a set of embedding, \( K^e_T = \{k^{kw}_{T,i,j}\} \).

**Encoders:** Three knowledge encoders \( Enc_{K_P}, Enc_{K_C}, Enc_{K_T} \) are implemented as three independent Transformer networks (Vaswani et al., 2017):

\[
K_P = Enc_{K_P}(\text{POS}(K^e_P))
\]

\[
K_C = Enc_{K_C}(K^e_C)
\]

\[
K_T = Enc_{K_T}(K^e_T)
\]

where an output \( K_s \) can be viewed as a list or a set of vectors depending on the input \( K^e \). We use a set to pack \( K^e_{K_P}/K^e_{K_T} \) because no strong sequential correlation appears; thus, their encoders do not use the positional layer \( \text{POS} \). While the plain text \( K^e_P \) is a sequence, unlike encoding the dialogue history \( X \), we use a Transformer with \( \text{POS} \) because \( K^e_P \) has a significantly longer length.

### 2.3 Multi-Head Decoding

Previous knowledge-enhanced works (Wu et al., 2020; Bai et al., 2021) often use the single decoder paradigm. However, when using multiple sources, a single decoder always faces more challenges; namely, Knowledge Diversity, Indefinite Alignment, and Insufficient Flexibility and Scalability.

Thus, we propose to use multiple source-aware decoder heads, allocating one independent and fully functional decoder head for using the dialogue history or each knowledge source. The results of such decoder heads are subsequently fused by Source Fusion Network. The advantages can be summarized as 1) Each decoder head is independent; we can easily tune each network based on the source-specific characteristics without impacting other heads; 2) Each head does not need to
consider the complex combinations of knowledge usage. Each head only considers the usage of the corresponding input source. Thus, we can employ the more professional Source Fusion Network to alleviate the impact of indefinite alignment more efficiently; 3) Higher flexibility and scalability because Source Fusion Network does not limit the number and the knowledge-type of heads.

2.3.1 Source-Aware Decoder Head

Each decoder head $\text{Dec}_s \in \{\text{Dec}_H, \text{Dec}_{K_P}, \text{Dec}_{K_C}, \text{Dec}_{K_T}\}$ uses the corresponding source-specific dialogue/knowledge memory $M_s \in \{H, K_P, K_C, K_T\}$ to predict the target response with own networks/parameters $\theta_s$:

$$\text{Dec}_s(M_s; \theta_s), s \in \{H, K_P, K_C, K_T\} \quad (4)$$

**State Updating:** At time $t$, each $\text{Dec}_s$ first updates its state $z_s$ with a GRU network ($g_s$):

$$z_{s,t} = g_s(z_{s,t-1}, y_{t-1}, a_{s,t}), z_{s,0} = h_{x_s} \quad (5)$$

where each initial state $z_{s,0}$ is universally initialized by the last dialogue history state $h_{x_s}$, $y_{t-1}$ is the embedding of the last generated token, $a_{s,t}$ is the attentive readout of the corresponding $M_s$.

**Memory Selection:** To obtain the attentive readout $a_{s,t}$ by selecting the memory $M_s$, we propose a Single-Side Attention for selecting the $X$ or $K_P$ (i.e., $M_s = H/K_P$), and a Dual-Side Attention for selecting the $K_{C/T}$ (i.e., $M_s = K_{C/T}$).

1. **Single-Side Attention:** we use a distribution $\text{align}_{s,t}$ to measure the relevance between each memory slot $m$ and the current context:

$$\text{align}_{s,t} \in R^{1 \times |M_s|} = F^S([y_{t-1}; z_{s,t-1}]W^A_M^T) \quad (6)$$

where $F^S$ is softmax, $W^A_m$ is a parameter, $\text{align}_{s,t}$ is an align distribution, i.e., weights. Then, the attentive readout $a_{s,t} = \text{align}_{s,t}M_s$ is the weighted sum of the memory $M_s \in R^{l_{M_s} \times \text{dim}}$.

2. **Dual-Side Attention:** The commonsense knowledge $K_C$ and the table knowledge $K_T$ have two value sides (head/tail entities, attribute key/words). Considering this, similar to the multi-head attention, the corresponding attentive readout adopts two different side-aware align distributions:

$$a_{K_C,t} = [\text{align}^\text{head}_{K_C,t}K_C; \text{align}^\text{tail}_{K_C,t}K_C]$$

$$a_{K_T,t} = [\text{align}^\text{key}_{K_T,t}K_P; \text{align}^\text{value}_{K_T,t}K_P] \quad (7)$$

The computations of $\text{aligns}$ in Equation 7 still follow the same way in Equation 6. In each head, two $\text{aligns}$ use the same network but the different parameters; the differences and the uniqueness of two $\text{aligns}$ come from the following copy mechanism.

**Token Perdition:** Each source-aware decoder head in $\text{KSAM}$ can generate a complete probability distribution to predict the next token.

First, a decoder head can generate a token by selecting a word from the fixed vocabulary $V$ using the distribution $P_{v_t}(w)$, which is given by:

$$F^S(\text{tanh}([z_{s,t}; y_{t-1}; a_{s,t}]W^V_1)W^V_2) \quad (8)$$

Next, to address the OOV issue and improve the informativeness, a decoder head can also copy a word from the corresponding source by using the previous attentive distribution $\text{align}_{s,t}$. In $\text{Dec}_H$ and $\text{Dec}_{K_P}$, $\text{align}_{H/K_P,t}(w)$ points out the probability to copy the word $w$ from $X/K_P$. In $\text{Dec}_{K_C}$, $\text{align}_{K_C,t}(w)$ points to the head/tail entity of the corresponding commonsense fact $w$. In $\text{Dec}_{K_T}$, $\text{align}_{K_T,t}(w)$ points to the attribute key/word of the corresponding attribute key-word pair $w$.

Finally, we use the following fusion gate $f_s$ to fuse all generation modes for each head:

$$f_s \in R^{1 \times 2/3} = F^S([z_{s,t}; y_{t-1}; a_{s,t}]W^M_1)(9)$$

Then, the aggregated probability is given by:

$$P_{s,t}(w) = \sum_i f_s[i] \text{align}_s^t([w] + f_s[1]P_{v_t}(w) \quad (10)$$

2.3.2 Source Fusion Network

Each head takes the responsibility for a single-source-aware prediction. For generating multi-source knowledge-enhanced responses, we propose a Source Fusion Network, which uses two gates, $f^h$ and $f^p$, to fuse the probability distributions outputted by decoder heads:

$$P_t(w) = \sum P_{t,i} f^h[i]f^p[i]P_{i,t}(w) \quad (11)$$

where the decoder head gates $f^h \in R^{1 \times 4}$ are:

$$F^S(\text{tanh}([a_{s,t}]W^{H1}W^{H2})) \quad (12)$$

and the step-wise gates $f^p \in R^{1 \times 4}$ are given by:

$$F^S(\text{tanh}([y_{t-1}; z_{s,t}; a_{s,t}]W^{S1}W^{S2})) \quad (13)$$
Training: The objective function has two terms:
\[
L = L_{\text{fused}} + \sum_{i} L_{\text{head},i} \quad (14)
\]

The first adopts the aggregated \( P_t(w) \) to compute the overall negative log-likelihood (NLL) :
\[
L_{\text{fused}} = -\sum_{t} \log P_t(y_t|y_{1:t-1}, X, \{K\}) \quad (15)
\]

The next term sums the NLLs of all heads:
\[
\sum_{\text{head},i} L_{\text{head},i} = -\sum_{\text{head},i} \sum_{t} \log P_{1,1}(y_t|y_{1:t-1}, X, M_i) \quad (16)
\]

The first \( L_{\text{fused}} \) can optimize the whole model, and the second \( \sum_{\text{head},i} L_{\text{head},i} \), makes sure that each head can move towards a better direction.

3 Experiment

3.1 Settings

Dataset: As reported in Table 1, we use a multi-source knowledge-aligned conversational dataset\(^5\) released by Wu et al. (2021a), which collected dialogues from three weibo datasets (Shang et al., 2015; Ke et al., 2018; Cai et al., 2019), commonsense knowledge from ConceptNet (Speer et al., 2017), and plain text/table knowledge from the Wikipedia. The vocab size is 21,924.

Baselines: Depending on the knowledge source:

1. Traditional: The attentive Seq2Seq (Luong et al., 2015), and the improved Pointer-Generator Network (PGN) (See et al., 2017); a GPT-based model CDial-GPT (Wang et al., 2020b), which has been pre-trained on 1.3B words+6.8M dialogues.
2. Plain Text: RefNet uses a reference network to use the text-based knowledge (Meng et al., 2020).
3. Commonsense: The first work CCM (Zhou et al., 2018), and two prior STOAs ConceptFlow (Zhang et al., 2020a), ConKADI (Wu et al., 2020).
4. Table: SA-S2S (Liu et al., 2018b) and TransInfo (Bai et al., 2020) use table knowledge via a SA-LSTM/Transformer encoder, respectively.
5. Heterogeneous: GOKC is a recent knowledge-enhanced approach (Bai et al., 2021). It supports a variety of knowledge types. We disable the use of goal knowledge because we study the open-domain dialogue and no goal is provided in the dataset. MSKE (Wu et al., 2021a) is a multi-source knowledge-enhanced approach, which supports to use multiple sources at the same time.

Implementations: For Seq2Seq and PGN, we use our re-implemented PyTorch codes; for ConKADI, GOKC, and MSKE, we use the official codes; for the remaining baselines, the experimental results are collected from MSKE (Wu et al., 2021a). Therefore, in our (re-)implementations, we keep the same hyper-parameter setting as MSKE if available. In short, all dialogue history encoders are a 512-dimensional bi-GRU, all Transformers of knowledge encoders are 2-layer 8-head and 512-dimensional, all decoders are a 512-dimensional GRU. We use a 200-dimensional pretrained Chinese embedding (Song et al., 2018) to initialize all word embedding matrix, a 100-dimensional pre-trained TransE embedding (Bordes et al., 2013) to initial the embedding of commonsense knowledge entities/relations. We use Adam as the optimizer. The mini-batch size is 32; the learning rate is 0.0001. If the loss on the validation set starts to increase after an epoch, the learning rate will be halved. The training will be automatically stopped if the loss on the validation set increases in two successive epochs. Consequently, our model costs about two days on an Nvidia RTX 3090 GPU. In the inference stage, we apply the beam-search decoding strategy, where the beam width is 10.

3.2 Automatic Metrics

For measuring the relevance between the ground-truth response and the generated responses. We use the sentence-level embedding-based Embed-A/G/X (Average / Greedy / Extreme) (Liu et al., 2016; Bai et al., 2021), the character-level uni-gram CharF1, the word-level BLEU-1/2/3/4 (Papineni et al., 2002), and the word-level Rouge-L (Lin, 2004). Following Zhang et al. (2020a), we use the uni/bi-gram DISTINCT (DIST-1/2) to evaluate the word-level diversity, and the 4-gram Ent-4 to evaluate the word-level informativeness.

3.3 Results

3.3.1 Automatic Evaluation

We report the results in Table 2. For MSKE and our KSAM, we evaluate their single-source ablated variants at the same time. For KSAM, we additionally evaluate some ablated/modified variants.

Single-Source Knowledge: Compared to traditional models, most single knowledge-enhanced
models have notable improvements, indicating the external knowledge is quite helpful in the open-domain dialogue generation. The recent GOKC, MSKE, and our KSAM are not limited to a specific type of knowledge, and such three models are almost the best three in each group. It implies that they do not improve flexibility at the expense of performance. Meanwhile, our KSAM is undoubtedly better: 1) KSAM has more the best results in every group; 2) The results among the knowledge sources are pretty stable and deliver similar trends; on the contrary, GOKC is not stable because it has quite different results with different knowledge. Consequently, we can say that every source-specific encoder and source-aware decoder in KSAM are well-designed and more efficient.

Multi-Source Knowledge: Only MSKE and KSAM can use all three knowledge sources at the same time. Two models have the best and the most balanced performance among all models. Comparing them, MSKE only achieves slight advantages in three metrics (Embed-G/X & ROUGE-L), but KSAM has more notable leaderships in the remaining metrics. In addition, the automatic evaluation can not fully reflect our advantages. Compared to MSKE, KSAM has better scalability and flexibility in using knowledge sources, due to the design of independent source-aware heads.

The Partial Degradation of KSAM: The full KSAM brings notable improvements except in DIST-1/2 (diversity) and Ent-4 (informativeness). Such performance degradation does not surprise us: 1) Copying words besides the fixed vocabulary is a crucial way to improve diversity and informativeness. In KSAM, the probability distribution used to copy is already fused in each decoder head; therefore, Source Fusion Network can not explicitly perceive all copy distributions when fusing single-source distributions to make the final prediction. This may impact the enthusiasm/chance of copying words when appending more decoder heads; 2) The adopted beam-search decoding algorithm can only consider one distribution; thus, we have no chance to leverage such source-aware distribu-

Table 1: Dataset Statistics. The coverage reports the ratio of how many dialogues can be matched from the three sets of knowledge. The last section evaluates the ablated/modified KSAM variants. * is collected from Wu et al. (2021a), - is not available or comparable (GOKC CSK outputs a abnormally large PPL), ↓ indicates lower is better, and PPL refers to perplexity. We use different colors to indicate the best performance in each group; and we use colored to indicate the best score among models except the ablated/modified KSAM.

Table 2: Automatic results. The last section evaluates the ablated/modified KSAM variants. * is collected from Wu et al. (2021a), - is not available or comparable (GOKC CSK outputs a abnormally large PPL), ↓ indicates lower is better, and PPL refers to perplexity. We use different colors to indicate the best performance in each group; and we use colored to indicate the best score among models except the ablated/modified KSAM.
tions. 3) DIST and Ent do not consider fluency and rationality, higher is not always better. For example, DIST/Ent will give high scores if we randomly generate some disordered sentences. We should comprehensively consider every dimension. We verified 1) and 3) in our model variant $P^*_V$, where three knowledge source-aware heads only output the copy probability without being fused with the vocab probability $P^*_V$. It can be seen that $-P^*_V$ increased diversity/informativeness, but decreased the relevance and fluency. We will continue to improve this in the future.

The Coupling among Heads: In KSAM, each decoder head $Dec_s$ is an independently and fully functional network. The internal state of a head can not communicate with each other. Does KSAM need to strengthen the coupling between heads? To verify this, we design a model variant $+Link$. Similar to the (Kim et al., 2020; Zhao et al., 2020), we use a GRU to manage a global sequential state $s_t$, which is updated with the memory readouts and the states of heads: $s_t = GRU(s_{t-1}, [y_t; \{z_{s,t}\}; \{a_{s,t}\}])$. Then, we replace $y_t$ by $[s_{t-1}; y_{t-1}]$ when operating each head, where $s_t$ can be regarded as a link to strengthen the coupling. As reported in Table 2, the performance has decreased. It indicates that there may be interference among different sources, and our decoupled design is helpful to alleviate this issue.

3.3.2 Human Evaluation:
The comparison is pair-wise and we select 5 better baselines in the automatic evaluation. We employed 3 well-educated native speakers as volunteers to score 200 sampled cases (1,000 comparisons in total) from three criteria: 1) Fluency considers the fluency; 2) Rationality measures the relevance and rationality; 3) Informativeness measures the quality of the information offered in the generated response. Following (Wu et al., 2021a), we count the agreement among volunteers. The 2/3 agreements for three metrics are 98.7%, 93.7%, and 94.1%; the 3/3 agreements are 61.0%, 52.7%, 51.6%.

Table 3 reports the averaged human evaluation score. Notably, KSAM significantly outperforms baselines in all dimensions, demonstrating the same advantage as in the automatic evaluation. In terms of fluency, the results are less distinguishable than the other two metrics (except $GOKC_{CSK}$), indicating most models can already generate fluent responses in most cases. In terms of rationality and informativeness, the results are more distinguishable and can reflect the advantage of using external knowledge. $GOKC_{CSK}$ does not perform well in human evaluation because the generated responses are always disordered and unnatural.

### 3.4 Analyses and Discussions

| Metrics | Base | $Dec_H$ | $Dec_Kp$ | $Dec_{Kc}$ | $Dec_{Kp}$ | Full |
|---------|------|--------|--------|----------|----------|------|
| PPL     | 98.0 | 92.1   | 96.6   | 94.6     | 95.3     | 77.7 |
| Embed-A | 0.848| 0.848  | 0.852  | 0.850    | 0.851    | 0.856|
| BLEU-4  | 0.95 | 2.03   | 1.04   | 0.99     | 1.03     | 2.15 |
| DIST-2  | 8.56 | 36.8   | 10.4   | 12.5     | 9.9      | 30.2 |

Table 5: Head Ablation. Each is trained separately.

**Source Fusion Network (SFN):** We analyze it with Table 4: 1) The PPL of each source-aware head, even we select the best head case by case, is significantly worse than the fused one, which shows a) SFN can effectively and step-wisely select/fuse results from various heads; b) independent source-aware decoder heads can work together; 2) Although SFN has achieved promising results, the upper bound is still significantly better. It means fusing multiple complete distributions is quite challenging, there is still have great potential to study. We will continue to explore this in the future.

**Decoder Head Ablation:** Table 5 reports the performance if only using a single source-aware head.
**Case #1:**

- **History:** 吃什么可以瘦身？ How to lose weight through diet?
- **Text Know.:** 节食指限制吃的食物分量，通常是以减肥瘦身作为目的。Dieting refers to limiting the amount of food eaten, usually with the goal of losing weight.
- **Commonsense:** 多喝水 Drinking more water, MotivatedByGoal, 瘦身 Loosing weight

| Seq2Seq | ConKADI | MSKE | KSAM |
|---------|---------|------|------|
| 吃减肥药最容易发胖. Taking diet pills is the easiest way to gain weight. | 可以吃一些东西 喝就可以了. You can eat something and drink on it. | 少吃什么东西都可以瘦身. You can lose weight if you eat less. | 喝水、多吃蔬菜，少吃油腻的食物. Drink more water, eat more vegetables, and eat less greasy food |

**Case #2:**

- **History:** 您认为“吸烟有害健康”的宣传教育有效吗？ Do you think the publicity and education of “Smoking is harmful to health” is effective?
- **Table Attributes of Smoking:** (Category, 习惯 habit), (Category, 烟草 tobacco)

| Seq2Seq | ConKADI | MSKE | KSAM |
|---------|---------|------|------|
| 吸烟有害健康. Smoking is harmful to health. | 吸烟有害健康, 不知道[...]? Smoking is harmful to health. Don’t you know smoking is harmful to health? | 吸烟有害健康. Smoking is harmful to health. | 吸烟有害健康，或者戒酒！Smoking is harmful to health. Quit smoking and alcohol |

Table 6: Case Study. ‘[...]’=吸烟有害健康‘ in #2. Besides the history, we show the related available knowledge.

*Base* removes the usage of the history memory H from DecH, and we regard it as the baseline. 1) Compared to *Base*, DecK further adopts a single-source memory and achieves improvements. The dialogue history memory H is undoubtedly more crucial than the external knowledge; 2) Commonsense knowledge memory K ε brings more improvements than the other two knowledge memories; 3) Using all heads (Full) has the best performance, indicating the necessity of using multi-source knowledge. Meanwhile, the improvement of PPL is significantly more than other metrics, indicating the decoding algorithm (beam search) should be improved in the future.

**Case Study:** Table 6 provides two cases for four better models in human evaluation. As a whole, we can find the *Indefinite Alignment* issue appears, where case #1 is aligned to both plain text knowledge and commonsense knowledge, and #2 is aligned to table knowledge. In addition, we can also notice the *Knowledge Diversity*, where such three knowledge sources have different characteristics. In case #1, *Seq2Seq* and *ConKADI* generated irrational responses. The response generated by *KSAM* is more informative than *MSKE* while both two responses are acceptable. In case #2, the provided knowledge is not straightforward; all baseline repeated the question. *KSAM* provided new information by reasoning on the table knowledge.

## 4 Related Work

**Dialogue Systems:** Dialogue systems have achieved promising results (Vinyals and Le, 2015; Chen et al., 2017). However, traditional models tend to generate safe but meaningless responses (Li et al., 2016). To this end, massive efforts are devoted to diversity the generated dialogues: leveraging the large-scale pretrained model (Zhang et al., 2020b; Gu et al., 2021), incorporating visual features (Das et al., 2017; Wang et al., 2021), employing topics (Xu et al., 2021; Zhong et al., 2021), and many others (Zhao et al., 2021).

**Knowledge-Enhanced Methods:** Recently, researchers noticed that a crucial reason that results in meaningless responses is the insufficient knowledge carried by the dialogue history (Ghazvininejad et al., 2018; Yu et al., 2020). Thus, infusing external knowledge into the dialogue generation has become a trend. Knowledge sources are diverse. The text knowledge can be easily collected and can provide rich information (Dinan et al., 2019; Ren et al., 2020; Meng et al., 2020). The commonsense knowledge includes the every knowledge (Speer et al., 2017; Zhou et al., 2018; Zhang et al., 2020a; Wang et al., 2020a). The table knowledge (Wu et al., 2019, 2021b) provides the entity-centric information. To improve the knowledge coverage and combine the advantages of different sources. (Liu et al., 2019) uses both text+commonsense knowledge; (Liang et al., 2021) uses different emotional sources; (Bai et al., 2021) treats goal knowledge as an additional source. (Wu et al., 2021a) does not limit the number/type of knowledge in theory; however, it ignored the *Knowledge Diversity / Indefinite Alignment* issue. In addition, the proposed multi-head decoding is different from the multi-processor decoding (Zhao et al., 2020): 1) our head is a fully functional decoder rather than a partially functional module; 2) we do not use a sequential state to strengthen the decoupling of heads; 3) our approach is not a single-source method.
5 Conclusion & Future Work

This paper studies the multi-source knowledge-enhanced dialogue generation. We find three challenging problems, i.e., 1) Knowledge Diversity, 2) Indefinite Alignment, and 3) Insufficient Flexibility and Scalability. Consequently, this paper proposes a novel Knowledge Source Aware Multi-Head Decoding approach, KSAM, which employs multiple source-aware decoder heads to handle each knowledge source more efficiently. In the future, we will continue to improve the applicability and the performance of multi-source knowledge-enhanced dialogue generation. For example, improving the fusing the predictions of heads.

Ethical Considerations: We did not propose a new dataset or use any private dataset. In addition, this work did not involve any sensitive topic. Thus, we believe no ethical concern in this paper.

References

Jiaqi Bai, Ze Yang, Xinnian Liang, Wei Wang, and Zhoujun Li. 2021. Learning to copy coherent knowledge for response 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, pages 12535–12543. AAAI Press.

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, Xiaojia 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.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 1080–1089.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.

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.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAIAI 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, pages 12911–12919. AAAI Press.

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 NAAACL, pages 110–119.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAIAI 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, pages 13343–13352. AAAI Press.
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.

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.

Shuman Liu, Hongshen Chen, Zhaocun Ren, Yang Feng, Qin Liu, and Dawei Yin. 2018a. Knowledge diffusion for neural dialogue generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 1489–1498.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Su. 2018b. 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 Hai Feng 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. 2020. Refnet: A reference-aware network 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 8496–8503.

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.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking locally, 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.

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

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

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. Computer Science.

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.

Shuhe Wang, Yuxian Meng, Xiaofei Sun, Fei Wu, Rongbin Ouyang, Rui Yan, Tianwei Zhang, and Jiwei Li. 2021. Modeling text-visual mutual dependency for multi-modal dialog generation. CoRR, abs/2105.14445.

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, Minghui Wang, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2021a. More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 2286–2300. Association for Computational Linguistics.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. 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, Minghui Wang, Ying Li, Dawei Zhang, and Zhonghai Wu. 2022. Improving the applicability of knowledge-enhanced dialogue generation systems by using heterogeneous knowledge from multiple sources. In WSDM ’22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022, pages 1149–1157. ACM.

Sixing Wu, Minghui Wang, Dawei Zhang, Yang Zhou, Ying Li, and Zhonghai Wu. 2021b. 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.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 3794–3804.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. 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, pages 14540–14548. AAAI Press.

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

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.