Exploiting Pairwise Mutual Information for Knowledge-Grounded Dialogue

Bo Zhang, Jian Wang, Hongfei Lin, Hui Ma, and Bo Xu

***Abstract*—External document knowledge is helpful for dialogue systems to generate high-quality responses. Although several knowledge-grounded dialogue models have been designed, external knowledge cannot be comprehensively exploited due to the complex relationships among dialogue context, knowledge, and responses. To this end, we propose a novel transformer-based model, named TransIKG, which incorporates external document knowledge for dialogue generation. TransIKG comprises a two-step integration mechanism, including correlation integration and overall integration. Correlation integration is designed to fully exploit the pairwise mutual information among dialogue context, knowledge, and responses, while overall integration adopts an integration gate to capture global information. Furthermore, we utilize the positional information of dialogue turns to better represent the dialogue context and enhance the generalization ability of our model on out-of-domain documents. Finally, we propose a novel knowledge-aware pointer network to generate knowledge-enhanced response tokens. Experimental results on two benchmark datasets demonstrate that our model outperforms state-of-the-art models on both open-domain and domain-specific dialogues.**

***Index Terms*—Dialogue position, discourse, knowledge-grounded dialogue, language generation, pairwise mutual information.**

I. Introduction

With the rapid development of end-to-end models, generation-based dialogue systems have received a great deal of attention from researchers [1], [2]. A key element for building a compelling dialogue system is the ability to generate fluent, logical and diverse responses based on dialogue contexts [3]. However, context-based end-to-end dialogue models suffer from generating short, generic, or repetitive responses [4]. One of the main reasons is that, unlike human-to-human dialogue, human-machine dialogue is often limited by external information outside the dialogue. Therefore, it is crucial to take full advantage of external knowledge information in developing effective dialogue models.

In recent years, several knowledge-grounded dialogue models have been proposed based on structured knowledge [5], [6] and unstructured knowledge [7]–[9]. Since abundant unstructured web documents contain rich and diverse information, recent studies have focused on incorporating document-based knowledge into dialogue systems [10]–[16]. Although the performance of dialogue systems is much improved, there are still some unresolved challenges.

The first challenge is the underconsidered complex relationships among context, knowledge, and responses. The complex relationships pertain to many topics, such as how to select knowledge based on dialogue context, how to generate responses based on dialogue context and knowledge, how to select new knowledge based on previously generated response information, and so on. In general, they consist of a mutual relationship between two of the three and the overall relationship of the three, which are defined as *pairwise mutual information*[1] in this paper. Previous models have mostly focused on choosing a single notion of gold knowledge [13]–[15], which focuses more on the relationship between context and knowledge and neglects the wealth of document-based knowledge information. To better integrate external knowledge, several dialogue models have sought to integrate multiple knowledge aspects for generating high-quality responses [10], [16]. Although previous works considered complex relationships more or less, how to fully use these relationships remains unsolved. We provide an example of a knowledge-grounded conversation in Fig. 1 to further explain this challenge. Fig. 1 illustrates that at the last turn of this conversation, the chatbot generated an utterance *"Studies suggest that all forms of walking have health benefits, and hiking is included"*. Although the first half of this utterance is knowledge-grounded, the second half ignores the dialogue context. Therefore, the relationship between knowledge, response, and context should be better modeled to generate a preferable response, such as *"..., and hiking is a good way to improve your health conditions"*.

The second challenge is that the attention of different dialogue turns is equally treated by most existing models. On the *Continuous Incrementality* view, messages can be continually prepared and updated in a human-to-human dialogue [17].

Manuscript received 29 September 2021; revised 24 January 2022 and 8 March 2022; accepted 8 March 2022. Date of publication 22 March 2022; date of current version 22 July 2022. This work was supported in part by the Natural Science Foundation of China under Grant 62006034, in part by the Natural Science Foundation of Liaoning Province under Grant 2021-BS-067, and in part by the State Key Laboratory of Novel Software Technology, Nanjing University under Grant KFKT2021B07. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jing Huang. *(Corresponding author: Jian Wang.)*

Bo Zhang, Jian Wang, Hongfei Lin, and Hui Ma are with the School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China (e-mail: zhangbo1998@mail.dlut.edu.cn; wangjian@dlut.edu.cn; hflin@dlut.edu.cn; huima@mail.dlut.edu.cn).

Bo Xu is with the School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China, and also with the Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: xubo@dlut.edu.cn).

Digital Object Identifier 10.1109/TASLP.2022.3161151

[1]Note that *pairwise mutual information* is different from the common notion of *mutual information* in information theory.

Fig. 1. Example of a knowledge-grounded conversation modified from the Wizard of Wikipedia dataset. The robot interacts with the person through the given document knowledge.

Therefore, interlocutors are mainly concerned with more recent utterances by others and partly omit other distant historical utterances, which can be regarded as information about the attention of different dialogue turns. Most early works have treated different dialogue turns in building dialogue generation models equally, which neglect the positional information of dialogue turns. Although recent models, such as the hierarchical recurrent encoder-decoder [18] and the incremental transformer encoder [10], have considered the positional information of the dialogue turns and achieved better performance, the gap between the relevant information and the point where it is needed has become larger due to their hierarchical architectures, which may exacerbate the problem of long-term dependencies [19]. Most importantly, none of them explicitly models the attention of different dialogue turns.

To tackle these two challenges, we propose a novel and effective **Trans**former-based architecture to **I**ncorporate **K**nowledge for dialogue **G**eneration, named TransIKG. **First**, for the initial challenge, a two-step integration mechanism is used in TransIKG to enhance the integration among context, knowledge, and responses. The first step, correlation integration, integrates the response at each decoding step with the fused context-knowledge through an integration gate. This allows TransIKG to use pairwise mutual information and keeps the response more relevant to the knowledge and context. The second step, overall integration, integrates the attentive context, attentive knowledge, and correlative response to predict response representation. This allows TransIKG to capture global information and utilize mutual information adequately. **Second**, for the next challenge, we introduce dialogue position, comprising dialogue position embedding (DPE) and dialogue position attention (DPA), to utilize the positional information of dialogue turns. Dialogue position embedding is used as an additional feature to represent the difference of different positions, and dialogue position attention is devised to represent the attention of different positions automatically. **Finally**, to further utilize knowledge, we leverage a novel pointer network upon multihead attention, which integrates transformer [2] and pointer generator networks [20]. The pointer network containing two pointers increases the probability of generating tokens from context and knowledge, which can be leveraged to further improve response quality.

The proposed model performs extensive evaluations on two document-grounded conversations datasets, i.e., an open-domain Wizard of Wikipedia dataset [7] and a domain-specific Holl-E dataset [8]. Experimental results demonstrate that **(1)** our proposed TransIKG model achieves comparable results to the strong baseline MIKe [15] with a restricted number of parameters (26 M vs. 245 M) on the Wizard of Wikipedia dataset; **(2)** TransIKG pretrained on an additional corpus outperforms the state-of-the-art methods on both datasets.

Our main contributions are summarized below:

- We propose a novel architecture, TransIKG, which employs a two-step integration mechanism to incorporate knowledge for dialogue generation.
- We introduce the dialogue position to utilize the positional information of dialogue turns.
- The effectiveness of TransIKG is empirically validated on two benchmarks.

II. Related Work

We discuss two categories of related work: knowledge-grounded dialogue and multiturn dialogue modeling.

A. Knowledge-Grounded Dialogue

Knowledge-grounded dialogue in generative dialogue consists mainly of structured knowledge [21]–[23] and unstructured knowledge [7], [10], [12], [12], [14], [16], [24], [25]. The most relevant work to ours is the generative dialogue that incorporates unstructured knowledge such as documents. Dinan *et al.* [7] combine memory network architectures [26] to select knowledge and transformer architectures to generate responses. Based on this work, several works pay more attention to knowledge selection, such as using both prior and posterior distributions over knowledge [14] or introducing an unsupervised learning scheme [24]. Several other works focus on the integration of knowledge. Li *et al.* [10] designed a two-pass transformer decoder to improve context coherence and knowledge correctness. Zhao *et al.* [12] take into account the mutual information of responses with context and knowledge through a disentangled decoder. Lin *et al.* [16] incorporate appropriate knowledge by using a recurrent knowledge interaction among response decoding steps, which considers the mutual information between response and knowledge. However, these methods fail to fully use the pairwise mutual information and the overall information among context, knowledge, and responses. We provide a two-step integration decoder that sufficiently incorporates knowledge using pairwise mutual information and overall information.

B. Multi-Turn Dialogue Modeling

Existing work on the modeling of multi-turn dialogue turns can be categorized into two groups: *flat concatenation* and

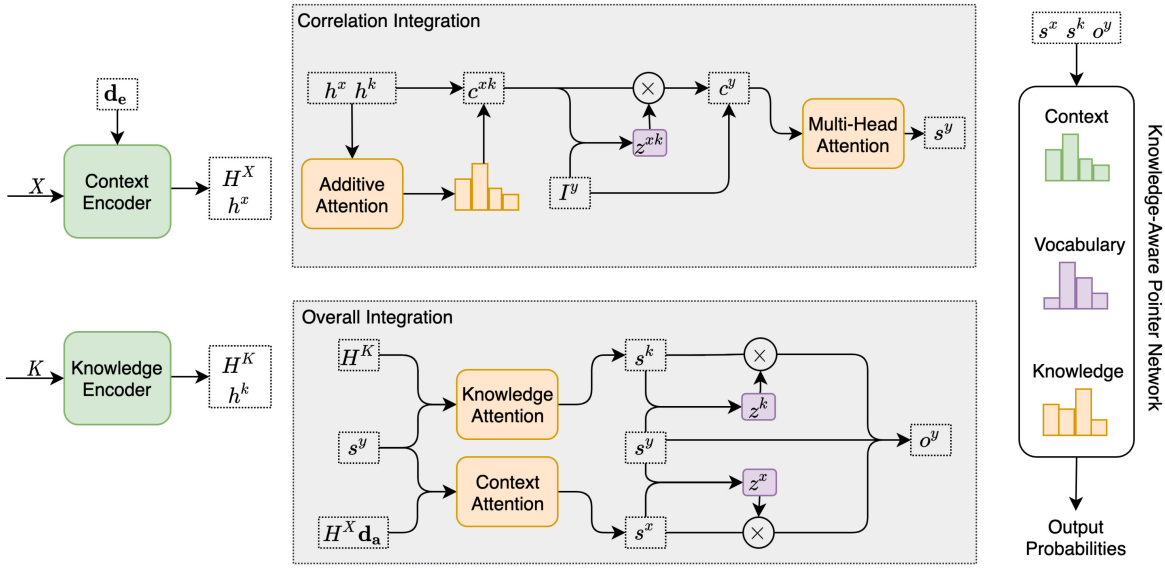

Fig. 2. Overview of the proposed architecture TransIKG.

hierarchical architectures. Methods in the first group, such as DialoGPT [27], concatenate the dialogue history as an input sequence. But these methods do not capture the positional information of the dialogue turns. Although DialoFlow [28] model the dynamic information flow across dialogue utterances, it does not model the attention of different dialogue positions explicitly. Methods in the second group commonly consider the positional information. Serban *et al.* [18] extend the hierarchical recurrent encoder-decoder architecture to the dialogue domain. Li *et al.* [10] then employ an incremental transformer encoder to encode multi-turn context incrementally. However, these models may exacerbate the problem of long-term dependencies. Our proposed Dialogue Position is inspired by both of these groups. Like the positional information of tokens in Transformer, we use an embedding to represent the positional information of dialogue turns and design a function to represent the attention of different positions.

III. Approach

A. Task Statement and Model Overview

Formally, given training data $\mathcal{D} = (X, K, y)$, where $X = (x_{1,1}, \ldots, x_{i,j}, \ldots, x_{l^x, n^x_{l^x}})$ is a dialogue context with $x_{i,j}$ being the j-th token of the i-th turn, $K = (k_{1,1}, \ldots, k_{l^k, n^k})$ represents a set of knowledge related to X containing l^k documents that have at most n^k tokens, and y is the response regarding X and K. The training goal is to learn a generation model $P(y|X, K; \theta)$, where θ denotes the model parameters, and the maximizing probability can be computed as:

$$\sum_{(X,K,y)\in\mathcal{D}} \frac{1}{|\mathcal{D}|} P(y|X,K;\theta) = \sum_{(X,K,y)\in\mathcal{D}} \frac{1}{|\mathcal{D}|} \prod_{t=1}^{l^y} P(y_t|X,K,y_{<t};\theta) \quad (1)$$

where y_t is the t-th token of the response; $y_{<t} = \{y_1, \ldots, y_{t-1}\}$. As illustrated in Fig. 2, the model TransIKG is based on the end-to-end model transformer. The major difference lies in the construction of **(1)** the context encoder that summarizes utterances with dialogue position embedding as contextual representations; **(2)** the integration decoder that employs a two-step integration mechanism; **(3)** dialogue position attention that is used in context attention to consider the attention of different positions; and **(4)** the knowledge-aware pointer network that can generate tokens from context, knowledge, and vocabulary.

B. Encoder

1) Context Encoder: Given a dialogue context $X = (x_{1,1}, \ldots, x_{i,j}, \ldots, x_{l^x, n^x_{l^x}})$, the context encoder encodes all utterances into hidden representation $\mathbf{H}^X \in \mathbb{R}^{N^x \times d}$, where $N^x = \sum_{i=1}^{l^x} n^x_i$ and d is the dimension of the vector. The input of the context encoder ($\mathbf{I}^X$) is a sequence of embeddings.

$$\begin{aligned} \mathbf{I}^X_{i,j} &= \mathbf{e}^x_{i,j} + PE^x_{i,j} + DPE_i \\ PE^x_{i,j} &= Embedding(i * n^x_{l^x} + j) \\ DPE_i &= Embedding(l^x - i) \end{aligned} \quad (2)$$

where $\mathbf{e}^x_{i,j}$ is the word embedding of $x_{i,j}$, the token position embedding is represented as $PE^x_{i,j}$, and DPE_i denotes the dialogue position embedding of $x_{i,*}$. The embedding weights of both PE and DPE are learnable. Then, $\mathbf{I}^X$ is fed into a transformer encoder (TE) to represent the context as a sequence of hidden vectors by

$$\mathbf{H}^X = TE^X(\mathbf{I}^X) \quad (3)$$

Then, $\mathbf{H}^X$ is converted to a context vector ($\mathbf{h}^x \in \mathbb{R}^d$) by summing the representations at each token position and then normalized by context length.

$$\mathbf{h}^x = \frac{\sum \mathbf{H}^X}{N^x} \quad (4)$$

2) Knowledge Encoder: Unlike the context encoder, when given a set of documents K, the input of the knowledge encoder is one document at a time.

$$\mathbf{I}^K_{i,j} = \mathbf{e}^k_{i,j} + PE^k_j \tag{5}$$

where $i = 1, \ldots, l^k$, $\mathbf{e}^k$ is the same word embedding as $\mathbf{e}^x$ and PE^k is also the same as PE^x.

Similarly, the knowledge encoder exploits TE to represent documents as a sequence of hidden vectors $\mathbf{H}^K \in \mathbb{R}^{N^k \times d}$ and a document vector $\mathbf{h}^k \in \mathbb{R}^{l^k \times d}$, where $N^k = l^k \cdot n^k$. Note that there is sharing of parameters between the context encoder and the knowledge encoder, and DPE is not used in the knowledge encoder.

C. Integration Decoder

When given the first $t-1$ tokens in the response $y_1, \ldots, y_{t-1}$, the integration decoder incorporates the context and knowledge into the response through the two-step integration mechanism. The purpose is to predict the representation of the t-th token and transmit it to the pointer network, mentioned in the following sections, for generating the t-th token.

1) Correlation Integration: Correlation integration is the first step of integration, aiming to exploit the pairwise mutual information among context, knowledge, and responses. First, correlation integration uses additive attention [29] to detect the distribution of knowledge (α) that highly coincides with the dialogue context and combines context with context-related knowledge to obtain their comprehensive representation ($\mathbf{c}^{xk} \in \mathbb{R}^d$) by

$$\begin{aligned}
\mathbf{c}^{xk} &= MLP\left(\left[\mathbf{h}^x; \sum_i \alpha_i \mathbf{h}^k_i\right]\right) \\
\alpha_i &= \eta(f_a(\mathbf{h}^x, \mathbf{h}^k_i)) \\
f_a(\mathbf{h}^x, \mathbf{h}^k_i) &= \mathbf{v}^\top \varphi(\mathbf{W}^x_a \mathbf{h}^x + \mathbf{W}^k_{a_i} \mathbf{h}^k_i) + b_a)
\end{aligned} \tag{6}$$

where $MLP(\cdot)$ is a 2-layer multilayer perceptron activated by $gelu$ activation [30], η denotes $softmax$ activation, and φ is $tanh$ activation. $\mathbf{v}$, $\mathbf{W}^x_a$, $\mathbf{W}^k_a$, and b_a are trainable parameters. Second, motivated by GRU [31], the *integration* gate ($\mathbf{z}$) is designed to integrate the context and knowledge with the response. When the integration gate is close to 0, the fused representation ($\mathbf{c}^y \in \mathbb{R}^{(t-1)\times d}$) is forced to ignore the context-knowledge representation (c^{xk}), which means they are more irrelevant. This effectively enables the response at different decoding times to merge different levels of knowledge and context based on relevance. Finally, self-attention [2] is applied to capture the relationships between the fused representation to gain the correlative response ($\mathbf{s}^y$) by

$$\begin{aligned}
\mathbf{s}^y &= tMA(\mathbf{c}^y, \mathbf{c}^y, \mathbf{c}^y) \\
\mathbf{c}^y_j &= MLP([\mathbf{z}^{xk}_j \mathbf{c}^{xk}; \mathbf{I}^y_j]) \\
\mathbf{z}^{xk}_j &= \sigma(\mathbf{W}^{xk}_z [\mathbf{c}^{xk}; \mathbf{I}^y_j] + b^{xk}_z)
\end{aligned} \tag{7}$$

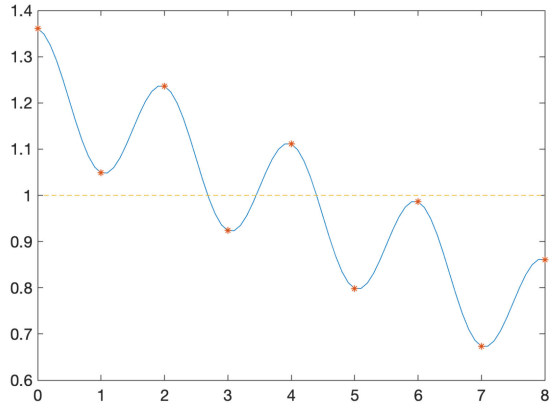

Fig. 3. Dialogue Position Attention with l^x as 8. The values of horizontal coordinates are given by $x = l^x - \mathrm{i}$ and the values of the vertical coordinates represent the attention weight of the i-th utterance.

where $MA(\cdot)$ signifies multihead attention with residual connection and layer normalization, $\mathbf{I}^y$ is the representation of the response similar to $\mathbf{I}^K$, and σ is a logistic $sigmoid$ function.

2) Overall Integration: Overall integration is the second step of integration, which aims to exploit general information. To enhance the decoder's understanding of context and knowledge, $\mathbf{s}^y$ is utilized to compute attention with context ($\mathbf{s}^x$) and knowledge ($\mathbf{s}^k$), respectively.

$$\begin{aligned}
\mathbf{s}^k &= MA(\mathbf{s}^y, \mathbf{H}^K, \mathbf{H}^K) \\
\mathbf{s}^x &= MA(\mathbf{s}^y, (\mathbf{d}^s + \mathbf{d}^b) \odot \mathbf{H}^X, \mathbf{H}^X) \\
\mathbf{d}^b &= \psi(\mathbf{W}_d \mathbf{H}^X + b_d)
\end{aligned} \tag{8}$$

$$\mathbf{d}^s_{i,*} = \frac{1}{8}(\cos(\pi \cdot (l^x - i) - \frac{4(l^x - i)}{l^x} + 7 + \frac{3l^x + 2}{l^x + 1}) \tag{9}$$

where $\mathbf{d}^s \in \mathbb{R}^{N^x}$ stands for the static part of dialogue position attention, as shown in Fig. 3, more details of which can be found in Section III-F. $\mathbf{d}^b$ is the dynamic bias of DPA. $\psi(x) = min(0.5, max(-0.5, x))$ is a $HardTanh$ function. $\odot$ denotes the face-splitting product.[2] Then, analogous to correlation integration, the predicted representation ($\mathbf{o}^y$) is obtained by employing the *integration* gate to integrate the attentive context, the attentive knowledge, and the correlative response by

$$\begin{aligned}
\mathbf{o}^y_j &= MLP([\mathbf{z}^x_j \mathbf{s}^x; \mathbf{z}^k_j \mathbf{s}^k; \mathbf{s}^y_j]) \\
\mathbf{z}^\star_j &= \sigma(\mathbf{W}^\star_z [\mathbf{s}^\star; \mathbf{s}^y_j] + b^\star_z)
\end{aligned} \tag{10}$$

where * is denoted as ^{x} or ^{k}.

D. Knowledge-Aware Pointer Network

The knowledge-aware pointer network in TransIKG is a pointer network based on multihead attention that exploits tokens copied from context and knowledge to improve the probability of generating corresponding tokens. Since there are

[2]The alternative concept of the matrix product, which uses rowwise splitting of matrices with a given quantity of rows, is also known as the transposed Khatri–Rao product.

multiple attention weights in the multihead attention, the final attention weights (β^x and β^k) can both be calculated by

$$\beta^\star = \eta(\mathbf{s}^\star \mathbf{W}_\beta^{\star\top}) \cdot f_\beta(\mathbf{s}^y, \mathbf{H}^\star) \tag{11}$$

where f_β means a function of obtaining multiple attentions in the multihead attention. $\mathbf{W}_\beta \in \mathbb{R}^{n^h \times d}$, where n^h is the number of heads.

1) Context Tokens: β^x from the last layer of the decoder can be directly regarded as the probability distribution of tokens copied from context, that is, $p^x = \beta^x \in \mathbb{R}^{N^x}$.

2) Knowledge Tokens: The distribution of knowledge that highly coincides with the context serves as a global distribution over all decoding steps. β^k from the last layer of the decoder is considered a local distribution at the current decoding step. Therefore, the distribution of tokens copied from knowledge is obtained by

$$\begin{aligned} p_i^k &= \lambda \cdot \beta_i^k + (1-\lambda) \cdot \alpha_j \\ \lambda &= \sigma(\mathbf{W}_\lambda \mathbf{s}^k + b_\lambda) \end{aligned} \tag{12}$$

where $j = \lfloor i \div l^k \rfloor$, $p^k \in \mathbb{R}^{N^k}$.

3) Vocabulary Tokens: The predicted representation from the integration decoder is used to generate a token from the vocabulary. The generation probability (p^y) is defined by

$$p^y = \eta(\mathbf{W}_v \mathbf{o}_j^y) \tag{13}$$

where the weight of $\mathbf{W}_v$ comes from the word embedding.

Then, the final probability to predict can be formulated as

$$\begin{aligned} P(y_t|X, K, y_{<t}) &= \gamma * [p^y; p^x; p^k] \\ \gamma &= \eta(\mathbf{W}_\gamma \mathbf{o}^y + b_\gamma) \in \mathbb{R}^3 \end{aligned} \tag{14}$$

E. Training Detail

1) Regularization: In addition to the dropout used as in [2], it is also used after attention softmax. Moreover, label smoothing [32] is used in training.

2) Loss Function.: The main training objective is to minimize the negative log-likelihood between the real response and the predicted response. In addition, knowledge loss [7] as an auxiliary loss is used to make the knowledge distribution α more closely resemble the gold knowledge distribution (α^g). Thus, TransIKG can be trained by minimizing the total objective:

$$\mathcal{L} = -\sum_t \log P(y_t|X, K, y_{<t}) - \sum_i \alpha_i \log \alpha_i^g \tag{15}$$

F. Dialogue Position Attention

According to the *Continuous Incrementality* view [17], in a long conversation, the respondent is mainly concerned with words spoken by the questioner and words spoken more recently. Therefore, we utilize linear decay to represent the decrease in attention over time and represent the difference between the responder and the questioner by initializing different attention weights. An explicit formula for the function that fits the above scenario can be written as

$$\begin{aligned} f(n) &= \begin{cases} -\dfrac{1}{2m}n + 1, & \text{if n is even} \\ -\dfrac{1}{2m}n + \dfrac{3}{4}, & \text{if n is odd} \end{cases} \\ &= \frac{1}{8}\left(\cos \pi n - \frac{4}{m}n + 7\right) \end{aligned} \tag{16}$$

where m is the maximum number of dialogue turns, n denotes the n-th closest to the present turns, and $0 \leqslant n \leqslant m$. Suppose m is even; then, the function for different m values is normalized by

$$\sum_{i=0}^{m} (f(i) + k) = m \tag{17}$$

where k is a constant related to m, and we obtain

$$k = \frac{3m+2}{8(m+1)} \tag{18}$$

Thus,

$$\mathbf{d}^s = \frac{1}{8}\left(\cos \pi n - \frac{4}{m}n + 7 + \frac{3m+2}{m+1}\right) \tag{19}$$

IV. Experiments

A. Dataset

We mainly evaluate our model on the open domain Wizard of Wikipedia dataset [7] to primarily verify its ability to incorporate knowledge. In addition, the specific domain Holl-E dataset [8] is used to primarily evaluate the ability to select knowledge.

Wizard of Wikipedia consists of open-domain conversations based on document knowledge retrieved from Wikipedia. Two participants engaged in chit-chat, the one as a wizard who could acquire knowledge sentences about a specific topic from Wikipedia and the other as an apprentice who was eager to discuss the topic in-depth with the wizard but could not gain external knowledge. It contains 18,430 dialogues for training, 1,948 dialogues for validation and 1,933 dialogues for testing. The test set is split into two categories: Test Seen and Test Unseen. Test Seen with 965 dialogues contains 533 overlapping topics with the training set. Test Unseen with 968 dialogues consists of 58 topics previously unseen in training or validation.

Holl-E contains movie conversations wherein each response is explicitly generated by copying or modifying sentences from external background knowledge. We use the version released by [14], which is suitable for knowledge selection. It contains 7,228 dialogues for training, 930 dialogues for validation, and 913 dialogues for testing. Similarly, two subsets of the test set are available: one with a single golden reference and the other with multiple golden references. Each dialogue turn has at least one golden reference, and multiple golden references are given in 4318 dialogue turns in total.

B. Baselines

The following models are employed as baselines for comparison with TransIKG:

- Seq2Seq [1] is a simple encoder-decoder model based on an RNN without access to external knowledge.
- Transformer [2] implements the state-of-the-art encoder-decoder framework based on multihead attention without access to external knowledge.
- TMemNet [7] is a knowledge-grounded generation model that uses a transformer memory network for knowledge selection and a transformer decoder for utterance prediction in an end-to-end manner.
- BART [33] is a pretrained sequence-to-sequence model widely used in natural language generation. The dialogue context and all knowledge are concatenated as input and fed into BART to generate responses.
- SKT [14] leverages a sequential latent variable model for knowledge selection. It uses BERT [34] to encode context and knowledge and generates responses through a copying mechanism.
- DukeNet [24] explicitly models knowledge tracking and knowledge shifting as dual tasks to promote knowledge selection. It has an encoder and decoder similar to SKT.
- MIKe [15] leverages a mixed-initiative knowledge selection method to improve performance by distinguishing between system-initiative and user-initiative knowledge selection. Furthermore, an initiative-aware self-supervised learning scheme is devised in MIKe to learn to discriminate the initiative type. It also has an encoder and decoder similar to SKT.
- KAT-TSLF [35] is a variant of the transformer with a decoupled decoder. It utilizes a three-stage learning framework based on weakly supervised learning that leverages large-scale ungrounded dialogues and an unstructured knowledge base for response generation and knowledge incorporation.

All models are implemented with the same method of data processing on the two datasets and additionally utilize previously unused data when training and testing on the Wizard of Wikipedia dataset.

C. Implementation Details

We use `ParlAI` [36] as the code framework to implement our models. To make the comparison fair, we implement three versions of TransIKG: TransIKG$_{base}$, TransIKG$_{bert}$ and TransIKG$_{kat}$. TransIKG$_{base}$ is described in Section III, TransIKG$_{bert}$ uses BERT as the context encoder and the knowledge encoder that share parameters, and TransIKG$_{kat}$ is pretrained on Reddit Conversation Corpus and Wikipedia dumps provided by [35].

For models without BERT, the word embedding size and hidden size are both set to 256. The text is represented by byte-pair encoding [37], and the embedding matrix is initialized with FastText [38]. Transformer-based models have 5 encoder layers and 5 decoder layers, with an FFN size of 512 and 4 attention heads. Models with BERT also have 5 decoder layers with an FFN size of 1024 and 4 attention heads. We optimize using Adam [39] and the inverse square root learning schedule with 5 k warmup updates. The initial learning rates are 0.0005 and 0.0001 for TransIKG$_{base}$ and TransIKG$_{bert}$, respectively. Gradient clipping is applied with a maximum gradient norm of 0.1, and dropout is set to 0.2. Label smoothing with a value of 0.1 is employed for response generation. We generate a beam search, setting the beam size to 2, and use 3-gram blocking. We turn the hyperparameters by Neural Network Intelligence [40]. It is worth noting that on the Holl-E dataset, the optimal model is selected directly based on performance on the test set since the validation set is not provided and used by [14].

D. Evaluation Metrics

1) Automatic Metrics: Following [7] and [15], metrics include unigram F1, BLEU-4, ROUGE-1, ROUGE-2 and ROUGE-L, which automatically measure the fluency, coherence, and relevance for response generation, and Recall@1 (R@1) for knowledge selection accuracy. Among the metrics, BLEU focuses on the precision rate, ROUGE focuses on the recall rate, and F1 is a combination of both rates. Because our approach incorporates knowledge via soft fusion rather than explicitly selecting knowledge, we use the knowledge that receives the most attention as selected knowledge to calculate Recall@1. We compute all scores by *parlai.core.metrics*.

2) Human Evaluation: We use the pairwise comparative evaluation `ACUTE` [41] to assess conversational quality between TransIKG and MIKe on Wizard of Wikipedia. First, 100 human-model conversations from each test set on Wizard of Wikipedia are selected at random. Then, the collected conversations are presented side by side, one between the human and TransIKG and the other between the human and MIKe. The A-B order is random with the masked model names. We ask three annotators the engagingness questions from [42], which measure from which model is more coherent, more knowledgeable, and more humanlike. We measured the percentage of time that one model was chosen over the other, taking the majority agreement among the annotators.

V. Results and Analysis

A. Quantitative Results

The automatic evaluation results on the Wizard of Wikipedia are reported in Table I. Compared with Seq2Seq and the transformer, which do not access knowledge, models with knowledge access are significantly improved by incorporating knowledge. However, the different approaches of incorporating knowledge also have a considerable impact on the results. TransIKG$_{base}$ with a restricted number of parameters (26 M vs. 245 M) significantly exceeds the performance of MIKe on the Test Unseen set. A smaller number of parameters means faster reasoning, which can be more easily applied in practice. There is no significant difference in performance between TransIKG$_{base}$ and MIKe on the Test Unseen set. However, TransIKG$_{bert}$ using BERT as an encoder outperforms MIKe in terms of all metrics on the Test Unseen set. Furthermore, it also has fewer parameters than MIKe. The performances of both TransIKG$_{kat}$ and KAT on the

TABLE I
QUANTITATIVE RESULTS ON THE WIZARD OF WIKIPEDIA DATASET.

Methods	Test Seen (%)						Test Unseen (%)						Parameters
	F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	R@1	F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	R@1	
Seq2Seq	14.02	0.29	14.09	2.61	12.34	–	11.82	0.11	11.95	1.47	10.51	–	24 M
Transformer	13.88	0.37	15.19	2.53	13.24	–	11.93	0.09	11.99	1.25	11.02	–	14 M
TMemNet	16.52	0.87	16.19	4.00	14.44	21.01	13.04	0.18	13.34	1.89	11.65	11.87	16 M
BART	19.38	1.51	18.70	5.89	17.55	–	17.36	1.28	17.21	4.39	15.57	–	139 M
SKT	18.41	1.94	19.03	6.01	16.94	25.01	15.98	0.88	16.14	3.91	14.50	17.93	174 M
DukeNet	19.33	2.16	19.73	6.46	17.59	26.13	17.08	1.40	17.33	4.78	15.33	19.39	184 M
MIKe	19.69	2.39	20.43	6.97	18.26	28.43	17.11	1.49	17.65	4.94	15.73	21.22	245 M
KAT†	20.46	2.08	21.74	6.72	19.01	–	18.82	1.67	19.75	5.56	17.06	–	198 M
TransIKG$_{base}$	19.91	2.67	21.37	6.99	18.51	26.46	16.90	1.45	17.79	4.59	15.71	17.74	26 M
TransIKG$_{bert}$	20.66	2.79	21.82	7.32	18.73	27.71	18.15	1.51	18.77	5.49	16.35	20.14	194 M
TransIKG$_{kat}$†	**21.31**⋆	**3.19**⋆	**22.09**	**8.20**⋆	**19.52**⋆	**29.09**	**19.40**⋆	**2.01**⋆	**20.09**	**6.92**⋆	**18.22**⋆	**23.75**	194 M

Methods Pretrained on an Additional Corpus are Marked by †. Significant Improvements Over the Best Baseline are Marked by ⋆ (T-Test, $p < 0.05$)
Bold font indicates the best performance in a column.

TABLE II
QUANTITATIVE RESULTS ON THE HOLL-E DATASET

Methods	Single Reference (%)						Multi Reference (%)					
	F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	R@1	F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	R@1
Seq2Seq	17.64	4.74	17.09	9.03	16.41	–	21.64	6.46	21.61	11.72	20.83	–
Transformer	18.28	4.88	17.46	9.11	16.84	–	22.42	6.34	21.97	11.37	21.25	–
TMemNet	24.07	13.65	25.32	16.24	23.89	24.20	30.61	19.86	25.32	16.24	23.89	33.61
BART	30.26	17.89	32.25	22.68	30.96	–	37.27	25.08	39.10	39.71	38.06	–
SKT	30.46	19.55	30.52	23.07	29.52	28.78	37.16	26.89	37.60	29.76	36.51	39.17
DukeNet	30.61	19.33	31.59	23.43	30.50	30.16	37.75	26.86	38.94	30.21	37.70	40.18
MIKe	32.11	21.10	32.76	25.05	31.69	32.55	38.37	28.09	39.30	31.17	38.12	41.31
TransIKG$_{base}$	29.52	19.47	31.02	22.81	39.04	27.46	37.03	26.51	38.86	29.44	37.71	38.99
TransIKG$_{bert}$	33.69	22.77	34.07	26.45	33.12	31.83	39.15	28.53	39.82	31.70	38.76	40.78
TransIKG$_{kat}$†	**34.83**⋆	**23.91**⋆	**35.14**⋆	**27.62**⋆	**34.15**⋆	**33.73**	**39.75**⋆	**28.99**	**40.40**⋆	**32.27**	**39.31**⋆	**42.42**

Bold font indicates the best performance in a column.

TABLE III
ABLATION STUDY ON THE WIZARD OF WIKIPEDIA DATASET

Methods	Test Seen (%)					Test Unseen (%)					Parameters
	F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	
TransIKG$_{bert}$	**20.66**	**2.79**	**21.82**	**7.32**	**18.73**	**18.15**	**1.51**	**18.77**	**5.49**	**16.35**	194,156,852
-BERT	19.91	2.67	21.37	6.99	18.51	16.90	1.45	17.79	4.59	15.71	25,565,457
-DPE	19.59	2.64	20.75	6.83	18.17	16.58	1.23	17.13	4.24	15.01	25,564,172
-DP	19.11	2.57	20.04	6.73	17.65	15.73	1.11	16.75	3.78	14.64	25,562,124
-DP, IG	18.81	2.05	19.90	6.31	17.33	15.55	0.99	16.68	3.71	14.05	23,595,537
-DP, CI	18.69	2.23	18.96	6.11	16.74	15.55	0.90	15.60	3.79	13.70	22,933,004
-DP, CI, OI	18.03	1.93	18.79	6.01	16.45	15.14	0.61	15.58	3.68	13.66	22,605,324
-DP, CI, OI, Copy	16.79	0.79	16.13	3.78	14.20	13.67	0.22	13.15	2.04	11.45	19,973,120

Bold font indicates the best performance in a column.

Test Unseen set significantly outperform the other models. This is because pretraining can substantially improve the generalization ability on out-of-domain knowledge [12].

Table II reports evaluation results on Holl-E. TransIKG$_{base}$ has lower scores than MIKe. This is because responses in Holl-E are explicitly copied or modified from gold knowledge, which means that the importance of selecting knowledge is more significant than incorporating knowledge on Holl-E. SKT, DukeNet and MIKe focus more on knowledge selection than TransIKG, and BERT can also enhance the performance of knowledge selection through greater representational capabilities. Nevertheless, TransIKG$_{bert}$ still outperforms MIKe in terms of overall results. TransIKG$_{kat}$ pretrained on an additional corpus significantly outperforms MIKE. It is worth noting that although our model is more focused on incorporating knowledge, its ability to select knowledge is equally powerful.

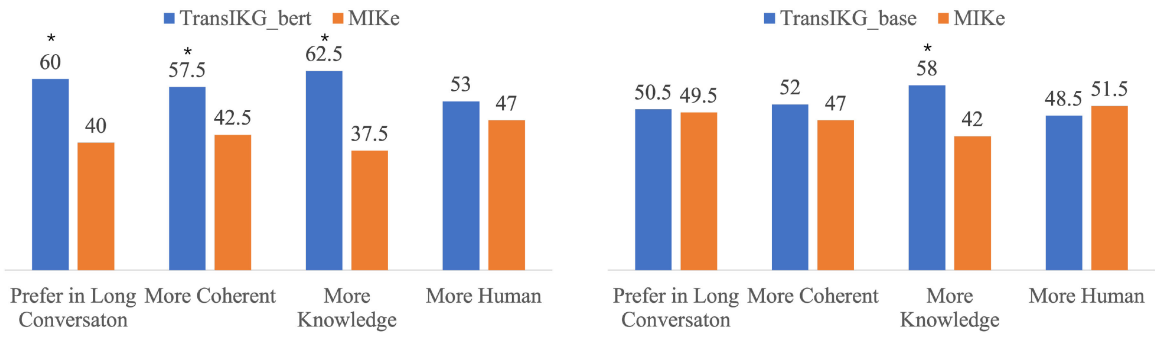

Fig. 4. Human Evaluation on the Wizard of Wikipedia dataset. Significant improvements are marked with * (binomial test, $p < 0.05$).

Context	A: The UFC is mixed martial arts based in Las Vegas. I love it! B: What do the initials UFC mean? A: The Ultimate Fighting Championship. B: What styles are in mixed martial arts? A: I am not sure but I do know Dana White is the president of the UFC. B: How old is Dana White? A man or woman? A: He is a man who has been president since 2001. I am not sure of his age. B: OK. That's a long time. Do you know when the UFC was started?	**Knowledge**	The Ultimate Fighting Championship (UFC) is an American mixed martial arts organization based in Las Vegas, Nevada, that is owned and operated by parent company WME–IMG. It is the largest MMA promotion in the world and features the top-ranked fighters of the sport. As of 2017, the UFC has held over 400 events. Dana White serves as the president of the UFC. He has held that position since 2001; while under the leadership of Dana White the UFC has grown into a globally popular multibillion-dollar enterprise. The first event was held in 1993 at the McNichols Sports Arena in Denver, Colorado.
Response	Human	I do not but I know it's the world largest MMA promotion.	
	SKT	It was started in Las Vegas, Nevada, and it was founded in Las Vegas, Nevada.	
	DukeNet	I' m not sure but i know that the ultimate fighting championship is owned by Las Vegas .	
	MIKe	I ' m not sure but i know they are owned by Las Vegas.	
	TransIKG	The first event was held in 1993 at the McNics Sports Arena in Denver, and the UFC has held over 400 events as of 2017.	

Fig. 5. Case Study from the Test Seen set on the Wizard of Wikipedia dataset.

B. Qualitative Results.

Human evaluation results are shown in Fig. 4. The blue and orange bars represent the percentage of scores corresponding to each of the two models. We find that TransIKG$_{bert}$ achieves a large improvement over MIKe, which is statistically significant for the evaluation questions (in Fig. 4, left). Our model scores well on human evaluations that ask how coherent or knowledgeable a dialogue sounds; however, it does not significantly surpass MIKe in sounding more human. One possible reason is that our model directs more attention to incorporating knowledge. Nevertheless, it still avoids making several mistakes that other generative models often make, such as repetition. The qualitative results between TransIKG$_{base}$ and MIKe are similar to the quantitative results (in Fig. 4, right). TransIKG$_{base}$ is comparable to MIKe overall, only stronger than MIKe in knowledge integration.

C. Ablation Study

To demonstrate the validity of our proposed individual modules, we perform an ablation study on the Wizard of Wikipedia. There are four factors: dialogue position embedding and dialogue position attention (DP), correlation integration (CI), overall integration (OI), and knowledge-aware pointer network (Copy). When OI is not used, the fully connected layer is used instead. To analyze the effectiveness of the integration gate (IG), we also perform ablation experiments by dropping them. As shown in Table III, the performance declines gradually as the number of components decreases, which means that each key component of the TransIKG model plays a critical role. The parameters increase slightly with the addition of each component except for BERT, which indicates that TransIKG has a good cost performance. Additionally, we employ DP in the transformer without external knowledge to verify its effectiveness on the basic multiturn dialogue. As shown in Table IV, all metrics are greatly improved after using DP. In particular, the performance decreases significantly without

TABLE IV
RESULTS OF THE TRANSFORMER MODEL WITH OR WITHOUT DIALOGUE POSITION ON THE BASIC WIZARD OF WIKIPEDIA DATASET

Methods	Test Seen (%)		Test Unseen (%)	
	F1	ROUGE-L	F1	ROUGE-L
Transformer	13.88	13.24	11.93	11.02
Transformer + DP	**14.98**	**14.04**	**13.65**	**11.80**

Bold font indicates the best performance in a column.

using DP on the Test Unseen set, implying that DP allows the model to exhibit good generalization ability on out-of-domain knowledge. The most prominent reason is that DP helps to better model the dialogue content. According to [12], pretraining is crucial to make the proposed model generalize well, whereas we accomplish this more directly through DP.

D. Case Study

As shown in Fig. 5, we show the responses generated by the baseline models and our proposed model TransIKG$_{base}$ and visualize the knowledge sources of the responses with colors. The knowledge highlighted in green is employed by TransIKG to generate the response. This knowledge is different from gold knowledge (highlighted with golden color) but provides a better fit to the context. The knowledge chosen by SKT, Duke and MIKe (highlighted with red color) is also different from the gold knowledge, yet it does not fit the context at all. The possible reason is that SKT, Duke, and MIKe do not explicitly distinguish the positional information of the dialogue turns, causing them to be confused by other unimportant turns of dialogue. SKT, Duke, and MIKe have various deficiencies in incorporating knowledge, which leads them to generate responses that conflict with the knowledge sources, whereas TransIKG integrates knowledge well. TransIKG can generate responses from two kinds of knowledge, which is difficult for other baseline models. In this case, although MIKe scores more than TransIKG on the automatic metrics, it is clear that TransIKG produces a higher-quality response.

VI. Conclusion

This paper presents a novel knowledge-grounded model, TransIKG, which effectively exploits the pairwise mutual information and the general information among context, knowledge, and responses. Motivated by human perception in the real world, we devise dialogue position attention to represent the attention of different dialogue turns. Experimental results demonstrate that our model can incorporate knowledge to create more high-quality dialogues.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 27th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

[2] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[3] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 21:1-21:32, 2020.

[4] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *Proc. Annu. Meeting Assoc. Comput. Linguistics, Tut. Abstr.*, 2018, pp. 2–7. [Online]. Available: https://www.aclweb.org/anthology/P18-5002

[5] S. Moon, P. Shah, A. Kumar, and R. Subba, "OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 845–854. [Online]. Available: https://www.aclweb.org/anthology/P19-1081

[6] W. Wu *et al.*, "Proactive human-machine conversation with explicit conversation goal," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3794–3804. [Online]. Available: https://www.aclweb.org/anthology/P19-1369

[7] E. Dinan *et al.*, "Wizard of wikipedia: Knowledge-powered conversational agents," in *Proc. Int. Conf. Learn. Representation*, 2019. [Online]. Available: https://openreview.net/forum?id=r1l73iRqKm

[8] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra, "Towards exploiting background knowledge for building conversation systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2322–2332. [Online]. Available: https://www.aclweb.org/anthology/D18-1255

[9] K. Zhou, S. Prabhumoye, and A. W. Black, "A dataset for document grounded conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 708–713. [Online]. Available: https://www.aclweb.org/anthology/D18-1076

[10] Z. Li *et al.*, "Incremental transformer with deliberation decoder for document grounded conversations," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 12–21. [Online]. Available: https://www.aclweb.org/anthology/P19-1002

[11] X. Chen *et al.*, "Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 3426–3437. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.275

[12] X. Zhao *et al.*, "Low-resource knowledge-grounded dialogue generation," in *Proc. Int. Conf. Learn. Representation, 2020*. [Online]. Available: https://openreview.net/forum?id=rJeIcTNtvS

[13] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu, "Learning to select knowledge for response generation in dialog systems," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 5081–5087.

[14] B. Kim, J. Ahn, and G. Kim, "Sequential latent knowledge selection for knowledge-grounded dialogue," in *Proc. Int. Conf. Learn. Representation, 2020*. [Online]. Available: https://openreview.net/forum?id=Hke0K1HKwr

[15] C. Meng *et al.*, "Initiative-aware self-supervised learning for knowledge-grounded conversations," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2021*, pp. 522–532.

[16] X. Lin, W. Jian, J. He, T. Wang, and W. Chu, "Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 41–52. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.6

[17] S. Brown-Schmidt and A. E. Konopka, "Processes of incremental message planning during conversation," *Psychon. Bull. Rev.*, vol. 22, no. 3, pp. 833–843, 2015.

[18] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3776–3784. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957

[19] Y. Bengio, P. Frasconi, and P. Simard, "The problem of learning long-term dependencies in recurrent networks," in *Proc. Int. Conf. Neural Netw.*, 1993, pp. 1183–1188.

[20] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Annu. Meeting Assoc. Comput. Linguistics.*, 2017, pp. 1073–1083. [Online]. Available: https://www.aclweb.org/anthology/P17-1099

[21] J. Xu, H. Wang, Z. Niu, H. Wu, and W. Che, "Knowledge graph grounded goal planning for open-domain conversation generation," in *Proc. AAAI Conf. Artif. Intell., 2020*, pp. 9338–9345. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6474

[22] S. Wu, Y. Li, D. Zhang, Y. Zhou, and Z. Wu, "Diverse and informative dialogue generation with context-specific commonsense knowledge awareness," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5811–5820. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.515

[23] J. Xu *et al.*, "Conversational graph grounded policy learning for open-domain conversation generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1835–1845. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.166

[24] C. Meng *et al.*, "DukeNet: A dual knowledge interaction network for knowledge-grounded conversation," in *Proc. Int. ACM SIGIR Conf. Res. Devlop. Inf. Retrieval, 2020*, pp. 1151–1160.

[25] H. Rashkin, D. Reitter, G. Singh Tomar, and D. Das, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 704–718. [Online]. Available: https://aclanthology.org/2021.acl-long.58

[26] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2015, pp. 2440–2448. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html

[27] Y. Zhang *et al.*, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2020, pp. 270–278. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-demos.30

[28] Z. Li, J. Zhang, Z. Fei, and J. Zhou, "Conversations are not flat: Modeling the dynamic information flow across dialogue utterances," in *Proc. Annu. Meet. Assoc. Comput. Linguistics. Int. J. Conf. Nat. Lang. Process.*, 2021, pp. 128–138. [Online]. Available: https://aclanthology.org/2021.acl-long.11

[29] D. Bahdanau and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representation*, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[30] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with Gaussian error linear units," CoRR, vol. abs/1606.08415, 2016. [Online]. Available: http://arxiv.org/abs/1606.08415

[31] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empir. Methods Nat. Lang. Process*, 2014, pp. 1724–1734. [Online]. Available: https://www.aclweb.org/anthology/D14-1179

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308

[33] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2020, pp. 7871–7880. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.703

[34] J. Devlin, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol*, 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[35] S. Liu *et al.*, "A three-stage learning framework for low-resource knowledge-grounded dialogue generation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2021, pp. 2262–2272. [Online]. Available: https://aclanthology.org/2021.emnlp-main.173

[36] F. Miller *et al.*, "ParlAI: A dialog research software platform," in *Proc. Conf. Empirical Methods Nat. Lang. Process., Syst. Demonstrations*, 2017, pp. 79–84. [Online]. Available: https://www.aclweb.org/anthology/D17-2014

[37] R. Sennrich and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. Annu. Meet. Assoc. Comput. Linguist*, 2016, pp. 1715–1725. [Online]. Available: https://www.aclweb.org/anthology/P16-1162

[38] P. Bojanowski, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://www.aclweb.org/anthology/Q17-1010

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[40] Microsoft, "Neural network intelligence," 2021. [Online]. Available: https://github.com/microsoft/nni

[41] M. Li and S. Roller, "ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons," CoRR, vol. abs/1909.03087, 2019. [Online]. Available: http://arxiv.org/abs/1909.03087

[42] A. Fan, C. B. Gardent, and A. Bordes, "Augmenting transformers with KNN-based composite memory for dialog," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 82–99, 2021. [Online]. Available: https://www.aclweb.org/anthology/2021.tacl-1.6