

Towards Improving Interpretability of Language Model Generation through a Structured Knowledge Discovery Approach

Shuqi Liu¹, Han Wu¹, Guanzhi Deng¹, Jianshu Chen³, Xiaoyang Wang³, Linqi Song^{1, 2,*}

¹City University of Hong Kong

²City University of Hong Kong Shenzhen Research Institute

³Tencent AI Lab

arXiv:2511.23335v1 [cs.CL] 28 Nov 2025

Abstract—Knowledge-enhanced text generation aims to enhance the quality of generated text by utilizing internal or external knowledge sources. While language models have demonstrated impressive capabilities in generating coherent and fluent text, the lack of interpretability presents a substantial obstacle. The limited interpretability of generated text significantly impacts its practical usability, particularly in knowledge-enhanced text generation tasks that necessitate reliability and explainability. Existing methods often employ domain-specific knowledge retrievers that are tailored to specific data characteristics, limiting their generalizability to diverse data types and tasks. To overcome this limitation, we directly leverage the two-tier architecture of structured knowledge, consisting of high-level entities and low-level knowledge triples, to design our task-agnostic structured knowledge hunter. Specifically, we employ a local-global interaction scheme for structured knowledge representation learning and a hierarchical transformer-based pointer network as the backbone for selecting relevant knowledge triples and entities. By combining the strong generative ability of language models with the high faithfulness of the knowledge hunter, our model achieves high interpretability, enabling users to comprehend the model’s output generation process. Furthermore, we empirically demonstrate the effectiveness of our model in both internal knowledge-enhanced table-to-text generation on the RotoWireFG dataset and external knowledge-enhanced dialogue response generation on the KdConv dataset. Our task-agnostic model outperforms state-of-the-art methods and corresponding language models, setting new standards on the benchmark.

Index Terms—structured knowledge, knowledge retrieval, language models, generation interpretability

I. INTRODUCTION

LANGUAGE models have proven to be highly effective in generating coherent and fluent text, such as GPT [1] and PALM [2], resulting in exceptional performance in various text generation tasks. Their utilization has also been subject to criticism due to their inherent “black box” nature, where the decision-making processes of the models are not transparent to human experts [3], [4]. This lack of interpretability can pose significant challenges, particularly in domains where model predictions need to be justifiable and explainable, such as in

knowledge-intensive tasks. The ability to provide explanations for the model’s decisions can increase transparency and accountability, improve trust, and enable better decision-making [5], [6]. Therefore, the development of interpretable language models is becoming increasingly important, and several approaches are being explored to enhance the transparency of these models [7], [8].

Recently, there has been significant progress in developing models that integrate language models with differentiable retrievers, such as REALM [9] and LAMBDA [10], which have demonstrated impressive results in unstructured document retrieval. While for structured knowledge retrieval, building upon the success of language models, current state-of-the-art approaches are centered around the integration of structured knowledge into language models, including the generation of descriptive text over multiple knowledge triples [11], the generation of question-answering over external knowledge bases [12], the incorporation of auxiliary supervised tasks around structured knowledge [13], reasoning knowledge paths over knowledge graphs [14], and developing highly specialized models [15], [16]. However, these existing methods face challenges when dealing with general large-scale structured inputs, leading to limitations in interpretability performance, such as knowledge tracing and factual consistency. Furthermore, many of these models are restricted to specific domains, tasks, or datasets. Ideally, models should possess the capability to provide interpretability across various tasks and structured knowledge types, including internal and external structured knowledge, tables, and knowledge graphs.

In this paper, we propose a novel approach to augment language models with a structured knowledge retriever, which leverages the organized hierarchy of structured knowledge, such as tables, knowledge graphs, and databases, to store large-scale world or domain knowledge. To capture structured knowledge in a more interpretable way, we augment the language model with a differentiable structure knowledge retriever, which allows the model to retrieve and attend over tables or knowledge graphs from a large corpus (shown in Fig. 1). Specifically, we propose a task-agnostic Structured Knowledge Hunter to hunt the most relevant knowledge triples for language model-based text generators, thus leveraging redundant structured knowledge under the language model’s

*Corresponding Author. This work was supported in part by the Research Grants Council of the Hong Kong SAR under Grant GRF 11217823 and Collaborative Research Fund C1042-23GF, the National Natural Science Foundation of China under Grant 62371411, InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies.

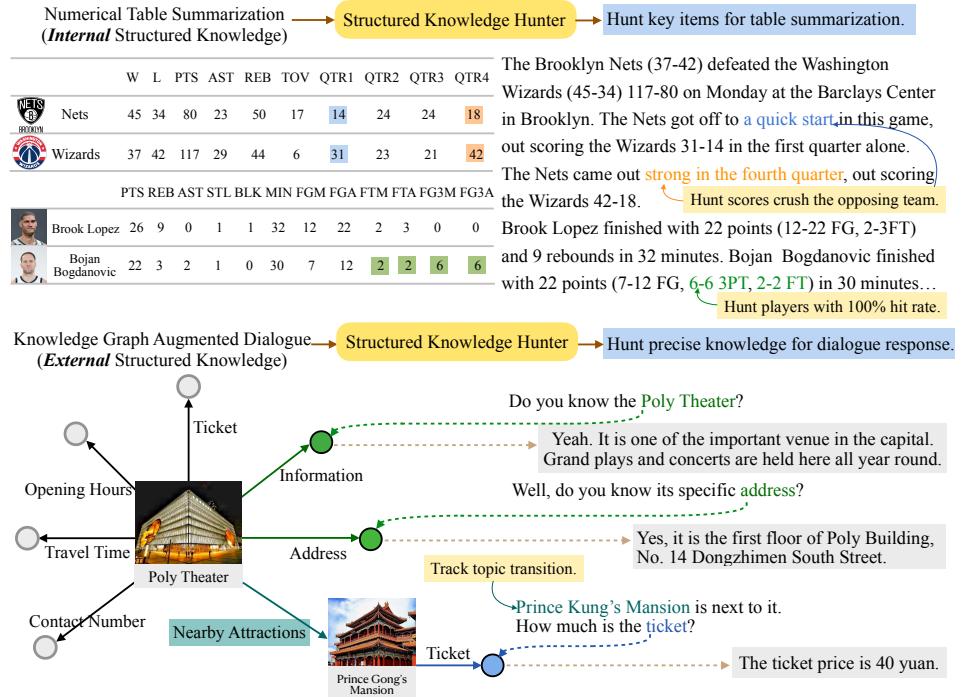


Fig. 1. Generation examples of Structured Knowledge Hunter for numerical table summarization (top) and knowledge graph augmented dialogue (down).

capacity. To avoid the task and dataset-specific design, we fully leverage the two-tier architecture of structured knowledge with the high-level entity and low-level knowledge triples. Specifically, Structured Knowledge Hunter follows a local-global interaction manner whereby the structured knowledge representation is learned by a hierarchical encoder that fuses local intra-entity representation and global inter-entity representation, and the factual guidance extraction is learned by a Hierarchical Knowledge Planner through multi-task learning. Hierarchical Knowledge Planner is designed inspired by the “macro-to-micro” selection behaviors of humans. “Micro” knowledge triples are hunted under the influence of selected “macro” entities, and multi-task learning encourages the matches between the hunted knowledge triple and the corresponding entity.

Our structured knowledge can be seen as a large external memory for language models to attend to. A key feature of our memory is that it is comprised of raw text rather than distributed representations, which makes the memory both (i) human-readable, lending a form of interpretability to our model, and (ii) human-writable, enabling us to dynamically update the model’s memory by editing the knowledge base. The task-agnostic feature of our Structured Knowledge Hunter enables it to acquire generic factual guidance from structured input pertaining to various tasks and flexible types of structured knowledge (e.g., tables or knowledge graphs) through implicit learning. We demonstrate this by both *internal* structured knowledge-augmented Rotowire-FG table-to-text generation task [17] and *external* knowledge graph-augmented KdConv dialogue response generation task [18]. Evaluation results indicate that our pipeline paradigm outperforms state-of-the-art methods and the corresponding language models,

thus setting new standards on the benchmark. Moreover, besides traditional text-level metrics, our Structured Knowledge Hunter also achieves higher knowledge selection accuracy relative to prior work with task-specific planning mechanisms.

Our contributions to this paper are as follows:

- To the best of our knowledge, we are the first to directly use knowledge hierarchy to design a knowledge encoder, planner decoder, and optimization method for its task-agnostic ability to learn from large-scale structured input;
- Our model combines the strong generation ability of language models with the high faithfulness of extractive knowledge selectors. As a result, our model can process generation tasks with high interpretability, allowing users to understand how the model generates its outputs.
- We empirically demonstrate that our task-agnostic model can be adapted to both internal knowledge-enhanced table-to-text generation and external knowledge-enhanced dialogue response generation. It outperforms state-of-the-art methods and the corresponding language models, thus setting new standards on the benchmark.

II. RELATED WORK

a) Language Models: General-Purpose Architectures: The previous research on general-purpose architectures for NLP tasks has demonstrated remarkable success, even without the use of retrieval. Specifically, a single pre-trained language model has proven to yield robust performance on a variety of classification tasks in the GLUE benchmark after fine-tuning [19], [20]. GPT-2 [21] extended this approach by showing that a single, left-to-right pre-trained language model could excel in both discriminative and generative tasks. Furthermore, BART [22] and T5 [23] propose a pre-trained encoder-decoder

model that utilizes bi-directional attention to achieve even stronger performance on these tasks. In this paper, we aim to broaden the range of possible tasks by incorporating a retrieval module to enhance pre-trained generative language models' performance within a single, unified architecture.

b) Language Models in Knowledge-Augmented Text Generation:

Extensive research has been conducted on acquiring knowledge in information retrieval, particularly through the utilization of pre-trained neural language models (PLMs) [24], [25], which greatly inspires our study. One line is to directly insert the pre-computed knowledge representation as an auxiliary input to PLMs [26], [27]. However, the method of explicitly injecting knowledge representation into PLMs has been argued that the vector space of PLM's words is inconsistent with knowledge representation since they are learned from separate tasks [28]. Another line is to implicitly incorporate knowledge information into PLMs by performing knowledge-related tasks such as concept order recovering [29] and entity category prediction [30]. Others optimize the retrieval module to aid in a specific, downstream task using search [31], reinforcement learning [32], or a latent variable approach [33]. Moreover, RETRO [34] extended by scaling the retrieval memory to trillions of tokens and changing the model architecture to take retrieved documents as input. While previous successes have relied on various retrieval-based architectures and optimization techniques to achieve impressive performance on individual tasks, our study demonstrates that a single retrieval-based architecture achieves exceptional performance across a diverse range of tasks.

c) Table-to-Text Paragraph Generation: Unlike earlier table-to-text datasets, like WikiBio [35] for sentence-level generation, RotoWire corpus [36] is more challenging due to its paragraph-level target texts and more redundant contents in the table. The RotoWire corpus [36] requires the model to generate NBA game summaries from the box- and line-score tables. Different from earlier table-to-text datasets, RotoWire is more challenging due to its longer target texts and more redundant contents in the table. Some attempts find that content selection and planning are indispensable steps to good generation quality [37], [38]. To this end, they try to introduce auxiliary tasks to guide better content selection, and planning [13], [39], [40]. However, these methods are not generalized due to task-specific objectives. In contrast, Tableformer [41] focuses on directly learning better table representations by modeling the table structure-aware features. Expressly, a hierarchical encoder [15] represents the table by combining both the row- and column-level features to produce the target texts. The essential differences between our method and the previous models are 1) our model directly benefits from the knowledge hierarchy, thus being task-agnostic; 2) we use language models as text generators to produce more fluent and coherent texts.

d) Knowledge Grounded Conversation: Since the standard dialogue response generation models tend to produce dull and less informative responses, knowledge-grounded multi-turn dialogue generation [42], [43] has attracted more attention recently. Knowledge-grounded conversation frequently uses a knowledge graph (KG) for the semantics in linked entities and

relations [44], [45]. KGs are also essential in fact verification to ensure the generation of faithful and accurate information [46], [47]. KG-augmented dialogue generation encourages the model to generate diverse and meaningful content with the aid of retrieved knowledge. To address this task, one line of work learns KG-aware input text representations with graph neural networks [48], [49]. Another line performs reasoning over KGs via path-finding strategies to provide auxiliary guidance for the generation process [50], [51]. Others explored fine-tuning the language models directly on the knowledge graph triple to transfer knowledge into a language model [52], [53]. Instead, we use an explicit knowledge hunter and model more fine-grained features of the selected knowledge.

III. TASK DEFINITION

We define a general pipeline paradigm for different source knowledge-enhanced text generation tasks which can be customized for both internal and external knowledge-enhanced text generation tasks. We focus on developing a two-stage approach to generate coherent and faithful text using a structured knowledge input K . The input consists of n knowledge triples and p entities. The two-stage method is to first learn an extractive knowledge plan, denoted as P_m , which includes m knowledge triples from K , and then be used by a language model-based text generator to produce high-quality text. To achieve this, we propose a generation process that involves selecting a knowledge triple k_t at each t -th step based on a probability distribution $p(k_t|P_{t-1}, K; \theta)$, where P_{t-1} represents the previously selected plan and θ denotes the model parameters. The selected triple is then added to P_{t-1} to form P_t . We define the best plan \hat{P} as the one that maximizes the probability of generating coherent and faithful text, given K and P_m .

$$\hat{P} = \operatorname{argmax}_{P_m, \forall m} \prod_{t=1}^m p(k_t|P_{t-1}, K; \theta) \quad (1)$$

Suppose that we have a dataset $\mathcal{D} = \{X_i, K_i, y_i\}_{i=1}^N$, where $\forall i \in \{1, \dots, N\}$. In the context of dialogue response generation, the input X_i represents the dialogue context, while K_i is an external knowledge graph that contains relevant information associated with X_i . The objective is to generate a response y_i based on X_i and K_i . On the other hand, in the case of table-to-text generation, the input X_i is a sequence of attribute-value tuples, and the output y_i is a long descriptive text. In this case, K_i is equivalent to X_i , as both represent the internal knowledge required to generate the output text.

IV. MODEL

The proposed pipeline paradigm for steering language model-based text generator with a Structured Knowledge Hunter is depicted in Figure 2. In this paradigm, the Structured Knowledge Hunter module is responsible for processing structured knowledge and extracting associated knowledge triples that serve as factual guidance to the language model-based text generator. To make the extracted knowledge triples easily understandable, they are transformed into readable sentences using pre-defined templates. The transformed sentences are

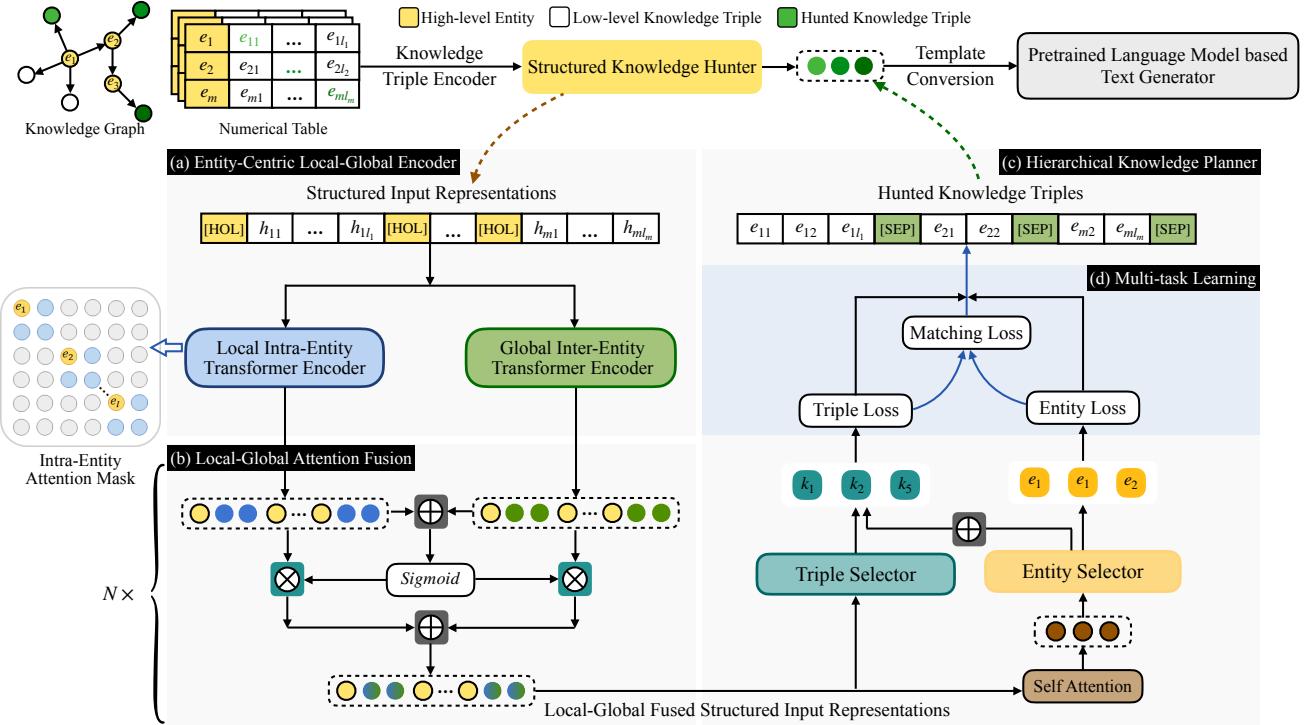


Fig. 2. Overall pipeline paradigm of steering language model-based text generator with a Structured Knowledge Hunter. The Knowledge Hunter is designed as a hierarchical encoder-decoder system that searches for relevant knowledge triples from input knowledge graphs or numerical tables. These knowledge triples are then processed through a Template Conversion stage that transforms them into a format that can be readily fed into the language model. Finally, the language model generates text based on the processed knowledge triples.

then used to fine-tune the language models to learn the factual knowledge guidance and improve text generation.

Specifically, (a) we introduce an Entity-Centric Local-Global Encoder that is designed to encode structured input representations, which have been embedded by a Knowledge Triple Encoder, in both local intra-entity and global inter-entity views. This allows us to capture both the details of individual entities as well as their relationships with other entities in the structured data; (b) we utilize the Local-Global Attention Fusion module to aggregate the local-global scale context for fusion output. This module is crucial in ensuring that important information from both the local and global views is captured and effectively integrated into the final output. (c) To generate high-quality knowledge triples, we introduce a Hierarchical Knowledge Planner that operates in a multi-task learning setting to jointly generate low-level hunted knowledge triples and corresponding high-level entities. This approach ensures the consistency of the hunted knowledge hierarchy and encourages matches between hunted knowledge triples and generated entities. By incorporating this hierarchical planner, we are able to generate more accurate and contextually relevant knowledge triples that can be used to enhance downstream tasks. Overall, our approach combines several innovative techniques to improve the accuracy and reliability of structured knowledge representation. By leveraging both local and global views of entities, utilizing attention-based fusion, and incorporating a hierarchical knowledge planner, we are able to generate more accurate and contextually relevant knowledge triples that can

be used to enhance various downstream tasks.

A. Knowledge Triple Representation

Knowledge triples are essential components of structured knowledge representations, including tables and knowledge graphs. Each triple in such representations captures either the relationships between different entities or the attribute-value pairs within a specific entity. For example, in a knowledge graph, a triple may represent a relationship between two entities, such as "Person A works for company B." Alternatively, a triple may represent an attribute-value pair within a single entity, such as "Person A has the job title of the software engineer." In this way, knowledge triples serve as the building blocks of structured knowledge representations, allowing for the efficient storage and retrieval of information.

Thus, we formulate each knowledge triple as $\langle \text{entity name}, \text{attribute name}, \text{attribute value} \rangle$, where each attribute value could be a number or a string of text for diverse knowledge sources.

a) *Numerical Table Embedding:* For numerically rich tables, we follow previous work [13] to learn triple representations from randomly initialized states. Considering each knowledge triple contains four embeddings: entity name n_{ij} , attribute name a_{ij} , attribute value v_{ij} , and attribute type t_{ij} . We encode each triple item into vector using concatenation $[;]$ and 1-layer MLP with trainable parameters W^e and b^e :

$$h_{ij}^{emb} = \text{Relu}(W^e[n_{ij}; a_{ij}; v_{ij}; t_{ij}] + b^e) \quad (2)$$

where i, j denotes the knowledge triple from i_{th} entity and j_{th} attribute.

b) Context-Aware Knowledge Embedding: Otherwise, for text-rich knowledge tuples in dialogue response generation, we choose BERT [54] as the backbone of our embedding module. Considering dialogue context $U = \{u_i\}_{i=1}^n$ of n utterances, where $u_i = (w_1^i, \dots, w_{l_u}^i)$ is a sequence of tokens of length l_u , and dialogue-related knowledge $K = \{k_i\}_{i=1}^m$ of m knowledge sentences from p entities, where $k_j = (w_1^j, \dots, w_{l_k}^j)$ is a l_k length token sequence, we define $S = (S_{11}, \dots, S_{pm})$ as the input to BERT encoder. S_{ij} concatenates the dialogue context and j_{th} knowledge from i_{th} entity: $S_{ij} = [\text{CLS}]w_1, \dots, w_{l_u}[\text{SEP}]w_1^j, \dots, w_{l_k}^j[\text{SEP}]$. Thus, each context and knowledge pair could be sufficiently interacted through multi-layer attention in BERT, leading to context-aware knowledge representation. Based on the hidden state of the first special token [CLS], each knowledge representation is finally given by $e_{ij}^{emb} = \text{CLS}(\text{BERT}(S_{ij}))$.

B. Entity-Centric Local-Global Encoder

Based on the transformer architecture, the Entity-Centric Local-Global Encoder is a novel model for encoding knowledge triples in structured data. This encoder is comprised of two main components: a Local Intra-Entity Transformer Encoder and a Global Inter-Entity Transformer Encoder. The former component is designed to encode knowledge triples within the same entity, whereas the latter component is responsible for encoding structured data based on all underlying knowledge triples.

To enable the local and global encoding of knowledge triples, we employ different mask matrices M . These matrices play a crucial role in controlling what context a given knowledge triple can attend to when computing its contextualized representation. By using distinct mask matrices, we are able to tailor the encoding process to the specific needs of each component, thereby ensuring that the resulting representations are optimized for their respective tasks. Overall, by leveraging the Entity-Centric Local-Global Encoder and carefully designed mask matrices, we can achieve highly effective encoding of knowledge triples in structured data. The mask matrix for different item i and j is formulated as:

$$M_{ij} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{otherwise no attending} \end{cases} \quad (3)$$

a) Local Intra-Entity Transformer Encoder: A local intra-entity transformer encoder enables the attention mechanism to focus only on the knowledge triples within the same entity, resulting in entity-aware knowledge representations. This is achieved through the use of a block diagonal matrix self-attention mask, which controls the attendance of each triple. The self-attention mask is composed of a main block diagonal part that is set to 0, and off-diagonal blocks that are set to $-\infty$ matrices. This is illustrated in Figure 2(a), where the self-attention mask is referred to as Intra-Entity Attention Mask. Each square matrix M_i in the mask corresponds to the attention mask for the i_{th} entity, which may contain varying numbers of knowledge triples. The use of this self-attention

mechanism in the local intra-entity transformer encoder allows for a more accurate and effective representation of entity-specific knowledge.

b) Global Inter-Entity Transformer Encoder: A global inter-entity transformer encoder enables all triples to attend to each other. In other words, the self-attention mechanism in this type of architecture allows for both inter-entity and intra-entity knowledge triples to be considered across all positions in the input sequence. This is achieved through the use of an all-zero matrix for the self-attention mask M , which effectively removes any limitations on the attention mechanism. As a result, the global inter-entity transformer encoder is able to capture more nuanced relationships and dependencies between entities, leading to improved performance in a variety of natural language processing tasks.

c) Attention Fusion Module: Attention Fusion Module aims to effectively combine the local and global contexts of a given input. This module accomplishes this task by allocating separate feature weights to each context, thereby allowing the model to selectively attend to relevant features. Given local view context L and global view context G , the fusion output H is aggregated as:

$$H = \sigma(L \oplus G) \otimes L + (1 - \sigma(L \oplus G)) \otimes G, \quad (4)$$

where σ is the Sigmoid function, \oplus denotes the element-wise summation and \otimes for the element-wise multiplication.

Local weights $\sigma(L \oplus G)$ and global weights $1 - \sigma(L \oplus G)$ are essential components of the Attention Fusion Module, which enable the network to perform a soft selection or weighted averaging between the local and global features. The local weights $\sigma(L \oplus G)$ and the global weights $1 - \sigma(L \oplus G)$ are real numbers between 0 and 1, which control the degree of emphasis given to the local and global features during the fusion process. This mechanism allows the network to balance the importance of local and global information, resulting in a more accurate and robust learning model. Additionally, the Attention Fusion Module can be executed multiple times by iteratively using the output of the fusion process as the initial structured input representations. By doing so, the network can refine the learned representations and further enhance the performance of the model. This iterative process can be repeated up to N times, where N is a hyperparameter that can be tuned based on the specific task and dataset. Therefore, the Attention Fusion Module is a powerful technique for integrating local and global features in a flexible and effective way, leading to improved performance in structured data encoding.

C. Hierarchical Pointer Planner

We refer to the cognitive processes involved in planning sequential knowledge triples from structured inputs known as the “macro-to-micro” strategy. This strategy involves first identifying the entities mentioned in the input and then delving into each selected entity to determine the final required knowledge triples. This approach is frequently observed in human planning and reflects a top-down approach to information processing, where higher-level concepts are processed before lower-level details. By characterizing this selection behavior,

our research sheds light on the underlying mechanisms of human cognitive processing and may have implications for the design of a knowledge planner.

With the Local-Global Fused Structured Input Representations H as input, the Hierarchical Knowledge Planner generates a subset of relevant knowledge triples K and their corresponding entities E . To handle the variability of the size of K and E across different contexts, we treat the construction process as a sequence generation task. At each step t , the planners add one knowledge triple or entity to K or E , respectively, starting from an empty set and expanding it gradually. To ensure that each added knowledge triple or entity is selected directly from its structured input, we employ a transformer-based pointer network as the backbone of both the Triple Selector and the Entity Selector. This approach allows the model to effectively learn to select relevant knowledge triples and entities, enabling it to generate accurate and informative sequences of K and E for a given input H .

a) *Entity Selector*: To extract entity representations from H , we make use of the [HOL] special triple embeddings, which are gathered and passed through a self-attention module. This allows us to model entity-level embeddings $H_e = \{h_{e_i}\}_{i=1}^p$, where p is the number of entities present in H . To maintain a sequence of entity decoder hidden states $\{s_t^e\}_{t=0}^{T_e}$ and weight $\{e_i\}_{i=1}^p$, we employ an attention mechanism. Specifically, we compute the attention weights e_i for each entity based on its relevance to the current decoder hidden state s_t^e . These weights are then used to compute a weighted sum of the entity embeddings H_e , which is subsequently incorporated into the decoder hidden state. This attention mechanism allows us to selectively focus on the most relevant entities at each decoding step, resulting in more accurate and effective entity representations.

$$P(e_i|e_{j_{1:t-1}}) = \frac{\exp(g(h_{e_i}, s_t^e))}{\sum_i \exp(g(h_{e_i}, s_t^e))}, \quad (5)$$

$$g(h_{e_i}, s_t^e) = v_e^\top \tanh(W_{h_e} h_{e_i} + W_e s_t^e), \quad (6)$$

where W_{h_e} , W_e and v_e are trainable parameters. These parameters are used to calculate the likelihood of the next entity being e_i given the previous entities in the sequence. The index j_t is then determined by taking the argmax of this distribution over all possible entities $i \in \{1, \dots, p\}$, denoted in (7). Once the index j_t has been determined, the next entity e_{j_t} is inserted into the sequence E .

$$j_t = \operatorname{argmax}_{i \in \{1, \dots, p\}} P(e_i|e_{j_{1:t-1}}) \quad (7)$$

The decoder state s_{t+1}^e is then calculated based on this new entity and the previous decoder state s_t^e . This calculation involves a combination of the entity embedding H_e , the previous decoder state s_t^e , and the output entity embedding from the attention mechanism h_{j_t} .

$$s_{t+1}^e = \operatorname{TransformerDecoder}(h_{j_t}, s_t^e, H_e) \quad (8)$$

b) *Knowledge Selector*: Similarly, given structured knowledge representation $H = \{h_i\}_{i=1}^m$, we maintain a sequence of knowledge decoder hidden states $\{s_t^k\}_{t=0}^{T_k}$ and attention weight $\{k_i\}_{i=1}^m$ as

$$P(k_i|k_{j_{1:t-1}}) = \frac{\exp(g(h_i, s_t^k))}{\sum_i \exp(g(h_i, s_t^k))}, \quad (9)$$

$$g(h_i, s_t^k) = v_k^\top \tanh(W_h h_i + W_k s_t^k), \quad (10)$$

where W_h , W_k and v_k are trainable parameters. Next triple index j_t is likewise selected by the max probability. Then element-wise summation is operated to combine entity and knowledge states for the calculation of s_{t+1}^k ,

$$s_{t+1}^k = s_{t+1}^e \oplus \operatorname{TransformerDecoder}(h_{j_t}, s_t^k, H). \quad (11)$$

D. Multi-task Learning

To ensure the accuracy and consistency of the knowledge hierarchy, we develop a hybrid optimization approach for our Structured Knowledge Hunter. Our optimization approach simultaneously optimizes three loss functions: the low-level knowledge triple loss \mathcal{L}_k , the high-level entity loss \mathcal{L}_e , and the matching loss \mathcal{L}_m . The matching loss is introduced to encourage the predicted knowledge triples to match with the predicted entities, thereby maintaining the hierarchical knowledge structure. Both \mathcal{L}_k and \mathcal{L}_e are computed using cross-entropy loss. Specifically, \mathcal{L}_k measures the discrepancy between the predicted and ground-truth knowledge triples, while \mathcal{L}_e measures the discrepancy between the predicted and ground-truth entities. This hybrid optimization approach ensures that our model learns to predict accurate knowledge triples while maintaining the hierarchical structure of the knowledge hierarchy.

$$\mathcal{L}_k = - \sum_{t=1}^{|K|} \log P(k_i|k_{j_{1:t-1}}), \quad (12)$$

$$\mathcal{L}_e = - \sum_{t=1}^{|E|} \log P(e_i|e_{j_{1:t-1}}). \quad (13)$$

while matching loss first maps the predicted knowledge triples into its corresponding entity by $\operatorname{map}(\cdot)$ function, and then calculated the normalized distance between the mapped entity and the predicted entity by

$$\mathcal{L}_m = \frac{1}{p^2} \sum_{t=1}^{|K|} (\operatorname{map}(K_t) - E_t)^2 \quad (14)$$

The final loss we optimize is thus given by

$$\mathcal{L} = \mathcal{L}_k + \mathcal{L}_e + \mathcal{L}_m \quad (15)$$

V. EXPERIMENTAL SETUP

In this section, we conduct experiments on two representing knowledge-enhanced text generation tasks to show the effectiveness of our method. We experiment with both internal knowledge-enhanced table-to-text generation on the RotoWire-FG dataset and external knowledge-enhanced dialogue response generation on the KdConv dataset. We describe our datasets, baselines, and implementation details here.

A. Datasets

We experiment with representative datasets outlined below.

RotoWire-FG [17] is a commonly used dataset for testing data-to-text generation models. It involves generating NBA game reports from box scores, which contain numerical tables capturing the statistical performance of both teams and individual players. The dataset consists of 7476 pairs of box scores and summaries of NBA games from 2014 to 2019. The data is split into 5232 train, 1125 validation, and 1119 test examples. On average, there are 626 record triples in a box score per game. The average summary has 205.9 words and 8.6 sentences. By utilizing this dataset, we can evaluate our model’s ability to accurately extract relevant details from numerical tables and generate concise summaries.

KdConv [18] is a representative dataset for enhancing multi-turn dialogue response generation with the use of knowledge graphs. It is a multi-domain Chinese knowledge-driven conversation consisting of discussions on film, music, and travel. The dataset encompasses 4.5K conversations and 86K utterances, with an average of 19.0 turns and 2.3 topics per conversation. To utilize the pointer network-based knowledge planner effectively, we utilized the Longest Common Subsequence (LCS) algorithm to identify and label each attribute-value pair as an index within the dialogue knowledge. Through this dataset, we aim to demonstrate the effectiveness of our model in generating high-quality responses in complex, multi-domain, and knowledge-rich conversational contexts.

B. Baselines

For the RotoWire-FG dataset, we compare against the following six baselines.

- Template [36] method produces a summary by filling specified numbers into human-written templates.
- CC [36] is an LSTM-based encoder-decoder model with a conditional copy mechanism.
- ENT [55] extends the CC model with an entity-specific dynamic memory module and generates a summary with a hierarchical attention mechanism.
- NCP [39] model employs a pointer network to select a subset of table records as content plan and then generates a summary using the CC model.
- NCP + TR [17] extends the NCP model with a table reconstruction loss to improve fact grounding.
- AuxEncoder [13] is an entity-graph encoder with a reasoning module supervised by two auxiliary tasks of number ranking and importance ranking.

For the KdConv dataset, we compare against the following baselines.

- Language Model (LM) [56]: LM generates the next utterance through maximum log-likelihood estimation given a concatenated sequence of utterances from a dialogue.
- Seq2Seq [57] & KSeq2Seq [18]: Seq2Seq is an encoder-decoder model that incorporates an attention mechanism, allowing the model to focus on different parts of the input sequence while generating the output. Kseq2Seq extends the Seq2Seq model by incorporating a key-value memory module to leverage knowledge information. All

knowledge triples mentioned in a dialogue are treated as the knowledge information in the memory module. Key memory is the average word embeddings of the head entity and the relation, and the value memory is those of the tail entity.

- HRED [58] & KHRED [18]: HRED is a hierarchical recurrent encoder-decoder model. The past conversational utterances are integrated into a context state, which serves as the initial hidden state for the decoder. KHRED extends the HRED model by incorporating a key-value memory module, similar to Kseq2Seq.

C. Implementation Details

For the Rotowire-FG dataset, we adopt the settings from [13] with a word embedding size of 600. For the KdConv dataset, we follow the configurations outlined in [18] by utilizing the Jieba Chinese word segmenter for tokenization [59]. Each response includes two preceding history utterances, and the maximum knowledge length is limited to 256. All transformer-based modules, including the intra-entity transformer encoder, inter-entity transformer encoder, topic transformer pointer, and knowledge transformer pointer, are composed of two transformer encoder or decoder layers with two multi-attention heads. For the hierarchical knowledge planner, we employ the AdamW optimizer [60] with a learning rate of 1e-4. We set the dropout rate to 0.2 and use a beam size of 5 during inference. The knowledge planner is trained for 30 epochs, and the best checkpoint is selected based on the content selection F1 score on the validation set. After obtaining the knowledge tuple indexes using the Structured Knowledge Hunter, we convert the selected knowledge tuples into textual representations using templates. The transferred knowledge plans are then realized through the text generator. For the text generator, we employ the AdamW optimizer with a learning rate of 3e-5. During the decoding stage, for the RotoWire-FG dataset, we set the minimum decoding length to 20 and the maximum decoding length to 256. For the KdConv dataset, the maximum decoding length is set to 128. The text generator is fine-tuned for 20 epochs, and the best checkpoint is selected based on validation loss on the validation set.

VI. RESULTS

A. Table-to-Text generation

We now describe the results on the RotoWire-FG dataset and compare our methods and the baselines in detail. In the table-to-text generation task, we conduct evaluations considering both knowledge selection accuracy and text quality. Knowledge selection accuracy is crucial for evaluating the interpretability of knowledge-enhanced text generation tasks. The extracted knowledge plan, which consists of relevant and accurate knowledge, forms the backbone of the generated text. This enables users to comprehend and trust the generated text, thereby improving overall interpretability. Additionally, we evaluate the text quality to ensure that the generated outputs are of high quality and coherent.

TABLE I

THE KNOWLEDGE SELECTION ACCURACY RESULTS FOR THE ROTOWIRE-FG DATASET. THESE METRICS INCLUDE RELATION GENERATION (RG) PRECISION, CONTENT SELECTION (CS) PRECISION, RECALL, F1 SCORE, AND CONTENT ORDERING (CO) REPRESENTED BY THE NORMALIZED DAMERAU-LEVENSTEIN DISTANCE (DLD%). BOLDFACED NUMBERS ARE THE BEST RESULTS.

Model	Relation Generation		Content Selection			Content Order
	P	R	P	R	F1	DLD
Template	93.72	23.98	43.96	31.03		10.25
CC [36]	81.51	36.15	39.12	37.57		18.56
ENT [55]	98.89	39.04	49.29	43.57		17.5
NCP [39]	94.21	43.31	55.15	48.52		23.46
NCP + TR [17]	95.70	42.90	56.91	48.92		24.47
AuxEncoder [13]	94.75	42.72	57.56	49.04		25.23
T5	92.75	43.62	56.09	48.98		23.26
BART	94.71	44.58	57.24	50.12		24.35
HunterAug-T5	97.15	48.36	69.98	53.28		25.23
HunterAug-BART	97.53	49.65	71.40	55.78		26.74

a) Knowledge Selection Accuracy: We employ the knowledge selection metrics [17], such as Relation Generation (RG), Content Selection (CS), and Content Ordering (CO), to assess the model’s ability to capture representative and relevant information from structured data. RG measures the precision of extracted triples from generated text compared to the box score table. CS considers the precision and recall of extracted triples from generated text compared to those from the target text. CO quantifies the normalized Damerau-Levenshtein distance between the triples extracted from the generated text and the target text. To evaluate the accuracy and relevance of the knowledge included in the summaries, we train a knowledge extractor using ensemble models that combine convolutional and LSTM models. Table I presents the evaluation results for knowledge selection accuracy. Initially, we assessed the generation and summarization capabilities of two state-of-the-art pre-trained language models (PLMs), T5 and BART. Subsequently, we integrated our Structured Knowledge Hunter into T5 and BART to enhance their ability to capture precise and comprehensive information during the text generation and summarization tasks. Among the tested PLMs, BART exhibited superior performance, surpassing AuxEncoder in content selection F1 score (50.12 F1%).

Following that, we proceeded to assess the performance of models augmented with the Structured Knowledge Hunter, specifically HunterAug-T5 and HunterAug-BART. The results demonstrate that our Structured Knowledge Hunter significantly improves the accuracy of hunted knowledge triples. The relation generation metric indicates that vanilla pre-trained language models like T5 and BART are less effective in utilizing factual knowledge compared to the previous state-of-the-art method, AuxEncoder. The integration of our Structured Knowledge Hunter into these models leads to significant improvements in relation generation performance. HunterAug-T5 achieves a relative increase of 2.53% compared to AuxEncoder, while HunterAug-BART surpasses AuxEncoder with a relative increase of 2.93%. Regarding content selection, vanilla BART already surpasses AuxEncoder in terms of F1 scores. When augmented with our knowledge hunter, it achieves a sig-

nificant relative increase of 11.29% compared to vanilla BART and a relative increase of 13.74% compared to AuxEncoder. Similarly, equipping the T5 text generator with our knowledge hunter results in an 8.78% relative increase compared to vanilla T5 and an 8.64% relative increase compared to AuxEncoder in terms of content selection F1 score. In terms of content ordering, vanilla pre-trained models like T5 and BART exhibit lower performance compared to AuxEncoder. However, when enhanced with our Structured Knowledge Hunter, the content ordering results surpass the performance of AuxEncoder with a relative improvement of 5.98%. The consistent improvement in knowledge selection accuracy suggests the effectiveness of our model in developing a better understanding of the relationships and dependencies among the information.

TABLE II
TEXT GENERATION QUALITY RESULTS FOR THE ROTOWIRE-FG DATASET.

Model	BLEU	BERT Score		
		P	R	F1
AuxEncoder	24.52	76.37	73.17	75.07
T5	23.34	77.83	73.90	75.81
BART	24.51	79.17	73.63	76.30
HunterAug-T5	26.87	79.23	76.18	77.68
HunterAug-BART	28.69	78.62	78.12	78.37

b) Text generation Quality: Text quality metrics assess the similarity and appropriateness of generated text compared to the ground truth. We employ automatic metric BLEU score, which measures n-gram overlap between generated and reference text. We also utilize more recent metric BERTScore [61], which considers semantic similarity and contextual understanding for a comprehensive evaluation of text quality beyond simple n-gram matching. According to the evaluation results in Table II, pretrained models enhanced with our structured knowledge hunter demonstrate superior performance in summary generation compared to the previous SOTA baseline model AuxEncoder, as well as pretrained language models T5 and BART. The enhanced models achieve impressive BLEU scores and BERT scores, indicating a significant overlap with the ground truth summary and accurate semantic

TABLE III
AUTOMATIC METRICS FOR KdConv: BLEU SCORE AND DISTINCT SCORE. BOLDFACED NUMBERS ARE THE BEST RESULTS.

Model	BLEU-1/2/3/4				Distinct-1/2/3/4		
	Travel						
LM [62]	27.51	17.79	12.85	9.86	3.18	8.49	13.99
Seq2Seq [57]	29.61	20.04	14.91	11.74	3.75	11.15	19.01
HRED [63]	30.92	20.97	15.61	12.30	4.15	12.01	27.16
KSeq2Seq [18]	37.04	27.28	22.16	18.94	4.25	13.64	24.18
KHRED [18]	36.87	26.68	21.31	17.96	3.98	13.31	24.06
HunterAug-BART	43.16	35.54	31.61	28.45	7.14	20.69	30.10
Music							
LM [62]	25.80	13.93	8.61	5.57	2.72	7.31	12.69
Seq2Seq [57]	28.89	16.56	10.63	7.13	2.52	7.02	12.69
HRED [63]	29.92	17.31	11.17	7.52	2.71	7.71	14.07
KSeq2Seq [18]	29.60	17.26	11.36	7.84	3.93	12.35	23.01
KHRED [18]	29.73	17.51	11.59	8.04	3.80	11.70	22.00
HunterAug-BART	28.63	19.97	15.93	13.04	4.69	15.39	25.55
Film							
LM [62]	24.22	12.40	7.71	4.27	2.32	6.13	10.88
Seq2Seq [57]	26.97	14.31	8.53	5.30	2.51	7.14	13.62
HRED [63]	27.03	14.07	8.30	5.07	2.55	7.35	14.12
KSeq2Seq [18]	27.45	14.51	8.66	5.32	2.85	7.98	15.09
KHRED [18]	27.94	14.69	8.73	5.40	2.86	8.08	15.81
HunterAug-BART	22.29	14.58	10.96	8.50	5.57	16.26	27.38
37.63							

representation of information. The BLEU score emphasizes the consistency of consecutive tokens, making it more sensitive to capturing factual information. On the contrary, BERTScore, being more semantic in nature, is less affected by factual information like numbers and names. Consequently, BERT scores tend to be higher when evaluating table summaries. Specifically, integrating our knowledge hunter model with pretrained models leads to an average BLEU improvement of 3.86 compared to vanilla pretrained models and a 3.26 improvement compared to AuxEncoder. Additionally, there is an average 1.97 BERTScore improvement compared to vanilla pretrained models and a 2.96 improvement compared to AuxEncoder.

B. Multi-turn Dialogue Response generation

We extend our model to handle text-rich, knowledge graph-driven tasks, specifically focusing on KdConv multi-turn dialogue response generation. We compare our methods with baselines, evaluating their performance in terms of knowledge selection accuracy and text quality. For evaluating KdConv multi-turn dialogue response generation, we use BLEU score and Distinct score as specific metrics. BLEU score measures the overlap between generated and reference text based on matching n-grams. Distinct score quantifies the diversity of the generated text by calculating the number of unique n-grams present. These metrics provide insights into the quality and diversity of the generated responses.

We present text generation quality results on the KdConv Dataset in Table III. The Chinese BART model [64] is utilized as the text generator for evaluating generation quality. Based on the results, the HunterAug-BART model demonstrates

superior performance compared to all other models in terms of BLEU-k and distinct-k metrics. Specifically, when compared to the state-of-the-art baseline model KHRED, HunterAug-BART achieves significant improvements in BLEU-4 and Distinct-4 scores in the travel, music, and film domains. The BLEU-4 score is raised by 10.49/5.00/3.10, and the Distinct-4 score is raised by 2.97/0.64/12.70, respectively. The increase in the distinct score highlights the enhanced generation capability of the pre-trained model utilized in HunterAug-BART. Moreover, the improvement in the BLEU score indicates the effective knowledge extraction ability of the hierarchical knowledge planner in selecting relevant knowledge for generating dialogue responses, as demonstrated by the knowledge selection score presented in Table IV. The Knowledge Selection (KS) score evaluates the factual consistency of the generated text by assessing the precision and recall of selected knowledge tuples from a knowledge graph compared to those in the target response. KS quantifies the alignment between the generated text and the underlying knowledge graph, which is crucial for maintaining factual consistency in KdConv systems.

TABLE IV
KNOWLEDGE SELECTION (KS) PRECISION (P%) AND RECALL (R%) FOR DIFFERENT DOMAINS IN KdConv DATASET.

Domain	KS-P%	KS-R%	KS-F1%
Travel	88.00	86.54	87.13
Music	85.48	82.39	83.65
Film	80.72	74.60	76.95

The difference between the BLEU score and the knowledge selection score can be explained by the limitations of the

KdConv dataset. The dataset only provides labeled responses that are relevant to knowledge, allowing the knowledge planner to accurately predict the absence of knowledge. However, the BART text generator may struggle to initiate a conversation when faced with incomplete or empty knowledge, which can impact the quality of the generated response and subsequently affect the BLEU score. Our study also observed significant performance variations across different domains in terms of BLEU score and knowledge selection score. These variations can be attributed to statistical differences in the domain datasets. The film dataset presented greater challenges due to its higher knowledge diversity, consisting of 1837 entities and 318 relations, as well as a higher average utterance count per dialogue (24.4), compared to the travel dataset, which had a more limited knowledge base of 699 entities and seven relations, and an average of 16.1 utterances per dialogue.

C. Ablation Study

To validate the effectiveness of our Structured Knowledge Hunter, we conducted ablation studies to assess the impact of different model components on overall performance. Specifically, we focused on evaluating the hierarchical knowledge encoder (Table V) and the combination of loss functions (Table VI). We focus on the Content Selection score as it is crucial for understanding the model’s decision-making process, explaining its performance and results, and guiding model improvements. It provides insights into how the model selects relevant information from structured knowledge and helps analyze its behavior during the knowledge extraction and selection process.

TABLE V
ABLATION STUDY ON HIERARCHICAL ENCODER

Encoder	Content Selection		
	P%	R%	F1%
Plain Transformer Encoder	43.68	61.41	51.11
+ Entity-Aware	45.38	62.95	52.71
+ Entity-Aware + 1 LG fusion	46.53	64.10	54.07
+ Entity-Aware + 2 LG fusion	46.88	65.07	54.49

In the ablation study (Table V), we evaluated the effectiveness of the hierarchical encoder by gradually incorporating the entity-aware design and local-global fusion modules into a plain transformer encoder optimized solely using knowledge loss. The results showed that the plain transformer encoder combined with a transformer pointer decoder achieved satisfactory content selection results with an F1 score of 51.11%. Introducing the entity-aware module, which considers structured knowledge hierarchically in an entity-centric manner, improved the F1 score by 1.60%. This highlights the importance of modeling structured knowledge through an entity-centric approach for performance enhancement. Additionally, we assessed the effectiveness of the local-global attention fusion (LG fusion) module by incorporating one and two LG fusion layers. The results demonstrated a 1.36% improvement in the F1 score with the addition of one LG fusion layer. However, introducing an additional LG fusion layer did not

yield a significant performance gain, suggesting that a single LG fusion layer is sufficient for optimal performance.

TABLE VI
ABLATION STUDY ON LOSS FUNCTIONS.

Loss	Content Selection		
	P%	R%	F1%
Only Triple Loss	46.88	65.07	54.49
+ Entity Loss	49.22	68.24	57.03
+ Entity Loss + Matching Loss	50.16	71.61	57.25

Table VI reports the effectiveness of our multi-task learning approach using different loss combinations in the proposed Structured Knowledge Hunter architecture. Our experiments revealed that maintaining consistency within the knowledge hierarchy was crucial for accurate prediction of triples belonging to the correct entity. When using only the triple loss, we achieved an F1 score of 51.49%, which improved by 2.54% when the entity loss was added. Furthermore, conducting additional tests with different random seeds showed a mean F1 score of 57.25 with a variance of 0.13 when using the matching loss. These results indicate that while the increase in F1 score may be marginal, the inclusion of the matching loss provides additional guidance and constraints that contribute to the overall performance of the model in content selection. These findings offer valuable insights into the effectiveness of our proposed architecture and the impact of different loss function combinations on the model’s overall performance.

D. Human Evaluation

To assess the agreement between human judgments and improvements in automatic evaluation metrics, we conducted a human evaluation. Three proficient English-speaking graduate students were recruited for this task. Specifically, we randomly selected 30 games from the RotoWire-FG test set. Each game was independently rated by three evaluators on a scale of 0 to 5 across four criteria: Coherence (do the sentences, in summary, follow a coherent discourse?), Fluency (is the summary fluent and grammatical?), Information Coverage (does the summary covers more supported facts?), and Information Accuracy (does the information in the summary is factual and correct?). The human evaluation result is presented in Table VII.

TABLE VII
HUMAN EVALUATION ON THE ROTOWIRE-FG DATASET.

Model	Coherence	Fluency	InfoCov	IncoAcc
AuxEncoder	2.97	3.67	3.30	2.93
T5	3.20	4.20	3.47	3.17
BART	3.63	4.27	3.87	3.37
HunterAug-T5	3.40	3.57	4.10	3.97
HunterAug-BART	3.57	3.73	4.13	4.03

As can be seen, texts generated by pretrained language model BART exhibits advantages in terms of coherence and fluency. On the other hand, PLMs equipped with our Structured Knowledge Hunter achieve superior information coverage and accuracy. While coherence and fluency are important,

TABLE VIII

A HYPOTHETICAL EXAMPLE FROM THE ROTOWIRE-FG DATASET FOR AN NBA GAME OF 5 POSSIBLE KNOWLEDGE PLANS AND THE CORRESPONDING SENTENCE REALIZATIONS. THE SUMMARIZATION OF THE PERFORMANCE IS HIGHLIGHTED IN BOLDFACED.

Input Table									
Team	Name	City	Wins	Losses	Points	Arena	1/4 Points	FG_PCT	FG3_PCT
Brooklyn Nets	Net	Brooklyn	37	42	117	Barclays_Center	31	51	60
Washington Wizards	Wizard	Washington	45	34	80	N/A	14	39	33
Player	Points	FGM	FGA	FTM	FTA	Reb	FGM	FGA	Minutes
Brook_Lopez	26	12	22	2	3	9			32
Bojan_Bogdanovic	12	7	12	2	2		6	6	30
Knowledge Plan									
S1: [Nets, city, Brooklyn], [Nets, name, Nets], [Nets, wins , 37], [Nets, losses , 42], [Wizards, city, Washington], [Wizards, name, Wizards], [Wizards, wins , 45], [Wizards, losses , 34], [Nets, points , 117], [Wizards, points , 80], [Nets, arena , Barclays_Center], [Nets, city, Brooklyn].									
S2: [Nets, name, Nets], [Nets, first_quarter_points , 31], [Wizards, first_quarter_points , 14].									
S3: [Nets, name, Nets], [Nets, FG_PCT , 51], [Nets, FG3_PCT , 60], [Wizards, name, Wizards], [Wizards, FG_PCT , 39], [Wizards, FG3_PCT , 33].									
S4: [Brook_Lopez, name, Brook_Lopez], [Brook_Lopez, points , 26], [Brook_Lopez, FGM , 12], [Brook_Lopez, FGA , 12], [Brook_Lopez, FTM , 2], [Brook_Lopez, FTA , 3], [Brook_Lopez, rebound , 9], [Brook_Lopez, minutes , 32]									
S5: [Bojan_Bogdanovic, name, Bojan_Bogdanovic], [Bojan_Bogdanovic, points , 12], [Bojan_Bogdanovic, FGM , 7], [Bojan_Bogdanovic, FGA , 12], [Bojan_Bogdanovic, FG3M , 6], [Bojan_Bogdanovic, FG3A , 6], [Bojan_Bogdanovic, FTM , 2], [Bojan_Bogdanovic, FTA , 2], [Bojan_Bogdanovic, minutes , 30]									
Realization									
S1: The Brooklyn Nets (37-42) defeated the Washington Wizards (45 - 34) 117-80 on Monday at the Barclays Center in Brooklyn.									
S2: The Nets <i>got off to a quick start</i> in this game, outscoring the Wizards 31-14 in the first quarter alone.									
S3: The Nets <i>defeated</i> the Wizards with a <i>strong shooting performance</i> , led by a 51% field goal percentage and a 60% from beyond the arc. The Wizards <i>struggled to keep up</i> , shooting 39% from the field and 33% from three-point range.									
S4: Brook_Lopez finished with 26 points (12 - 22 FG, 2 - 3 FT) and 9 rebounds in 32 minutes.									
S5: While Bojan_Bogdanovic finished with 22 points (7 - 12 FG, 6 - 6 3PT, 2 - 2 FT) in 30 minutes.									

providing accurate and comprehensive information is crucial in tasks prioritizing factual correctness, such as knowledge-grounded text generation or fact verification. By emphasizing information coverage and accuracy, our method ensures that the generated text is reliable and trustworthy.

E. Error Analysis

Having established that our Structured Knowledge Hunter can improve performance in many settings, we now move on to the research question: How do the failure modes of our method compare to other methods for knowledge-enhanced text generation?

1. The structured knowledge hunter may encounter difficulties in generating informative responses when the dialogue does not explicitly demand specific knowledge. This can result in an empty knowledge set being extracted, which is less satisfactory compared to utilizing the complete knowledge graph for end-to-end generation. Consequently, the dialogue response may be less informative.

2. The structured knowledge hunter may face challenges in accurately extracting knowledge tuples when the context does not explicitly indicate specific entities and relationships. As a result, the model may encounter difficulties in extracting the relevant knowledge that aligns with the intended meaning of the dialogue.

3. The structured knowledge hunter prioritizes factualness in text generation, it may rely heavily on structured knowledge sources to ensure the accuracy and reliability of the generated content. While this emphasis on factualness is beneficial for generating informative and trustworthy text, it can also lead to reduced diversity, especially when generating longer texts.

F. Discussion

Our structured knowledge hunter, with its hierarchical encoder and planner architecture, is a task-agnostic model capable of effectively handling both tables and knowledge graphs. Our model surpasses previous approaches by achieving superior knowledge selection accuracy and text generation quality while maintaining a reasonable increase in model size. In domains with incomplete, inaccurate, or poorly structured data, our hierarchical planner enhances the model's reliability as it has an additional level of assurance to select knowledge tuples from the queried entity.

G. Case Study

Table VIII presents a case study from the Rotowire-FG test dataset, which involves summarizing a table with redundant information. Our structured knowledge hunter first extracts representative knowledge tuples from the table to create a knowledge plan. Then, based on this plan, PLM generates the surface realization of the summary. Figure 3 presents a sample conversation from the KdConv test dataset in the travel domain between user A and HunterAug-BART. We observe that HunterAug-BART effectively selects the most relevant knowledge for user A's questions and generates responses that are both faithful and fluent.

VII. CONCLUSION

In this paper, we propose a novel structured knowledge hunter that enhances interpretability in knowledge-enhanced text generation. Our method directly leverages the two-tier

Travel Conversation	Hunted Knowledge Triples		
	Entity Name	Attribute	Value
A: 知道保利剧院吗? A: Do you know the Poly Theater?			
HuntAug-BART: 知道呀，首都重要的演出场所之一；这里常年会上演重量级的话剧和交响音乐会 HuntAug-BART: You know, one of the important performance venues in the capital; heavyweight plays and symphony concerts are held here all year round	保利剧院 Poly Theater	Information	北京保利剧院是首都重要的演出场所之一，这里常年会上演重量级的话剧和交响音乐会。感兴趣的朋友可以提前到官网看看近期有没有合适的演出，官网上还能选座。 Beijing Poly Theater is one of the most important performance venues in the capital, where heavyweight plays and symphony concerts are held all year round. Interested friends can go to the official website in advance to see if there is a suitable performance in the near future, and you can also choose a seat on the official website.
A: 嗯，那它的具体地址你知道吗? A: Well, do you know its specific address?			
HuntAug-BART: 知道，北京东城区东直门南大街14号保利大厦一层。 HuntAug-BART: Yes, it is the first floor of Poly Building, No. 14 Dongzhimen South Street, Dongcheng District, Beijing.	保利剧院 Poly Theater	地址 Address	北京东城区东直门南大街14号保利大厦一层。 The first floor of Poly Building, No. 14 Dongzhimen South Street, Dongcheng District, Beijing.
A: 啊，那它周边的景点有恭王府，一个很不错的景点 A: Ah, then the attractions around it include Prince Kung's Mansion, a very good attraction.			
A: 那它的门票是怎么收费的? A: So how are the tickets charged?			
HuntAug-BART: 门票价格是40元。	恭王府 Prince Gong's Mansion	门票 Ticket	40元 40 yuan

Fig. 3. A hypothetical example from the KdConv dataset and responses generated by our model HuntAug-BART for a travel conversation.

architecture of structured knowledge, consisting of high-level entities and low-level knowledge triples, to design our task-agnostic structured knowledge hunter. By combining the strong generative capabilities of language models with the faithful knowledge-hunting process, our model achieves high interpretability, enabling users to understand the model’s output generation process. We empirically demonstrate the effectiveness of our model in both internal knowledge-enhanced table-to-text generation and external knowledge-enhanced dialogue response generation tasks. Our task-agnostic model outperforms state-of-the-art methods and corresponding language models, establishing new benchmarks while maintaining a high level of interpretability.

REFERENCES

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, “Palm: Scaling language modeling with pathways,” *arXiv preprint arXiv:2204.02311*, 2022.
- [3] T. Sun, Y. Shao, H. Qian, X. Huang, and X. Qiu, “Black-box tuning for language-model-as-a-service,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 841–20 855.
- [4] S. Diao, Z. Huang, R. Xu, X. Li, L. Yong, X. Zhou, and T. Zhang, “Black-box prompt learning for pre-trained language models,” *Transactions on Machine Learning Research*, 2022.
- [5] D. Yin, L. Dong, H. Cheng, X. Liu, K.-W. Chang, F. Wei, and J. Gao, “A survey of knowledge-intensive nlp with pre-trained language models,” *arXiv preprint arXiv:2202.08772*, 2022.
- [6] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz *et al.*, “Augmented language models: a survey,” *arXiv preprint arXiv:2302.07842*, 2023.
- [7] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, “Qa-gnn: Reasoning with language models and knowledge graphs for question answering,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 535–546.
- [8] A. Creswell, M. Shanahan, and I. Higgins, “Selection-inference: Exploiting large language models for interpretable logical reasoning,” *arXiv preprint arXiv:2205.09712*, 2022.
- [9] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval augmented language model pre-training,” in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.
- [10] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, “Lamda: Language models for dialog applications,” *arXiv preprint arXiv:2201.08239*, 2022.
- [11] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C.-S. Wu, M. Zhong, P. Yin, S. I. Wang *et al.*, “Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 602–631.
- [12] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [13] L. Li, C. Ma, Y. Yue, and D. Hu, “Improving encoder by auxiliary supervision tasks for table-to-text generation,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5979–5989.
- [14] J. Ni, V. Pandelea, T. Young, H. Zhou, and E. Cambria, “Hitkg: Towards goal-oriented conversations via multi-hierarchy learning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 10, 2022, pp. 11 112–11 120.
- [15] H. Gong, X. Feng, B. Qin, and T. Liu, “Table-to-text generation with effective hierarchical encoder on three dimensions (row, column and time),” 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, pp. 3143–3152.
- [16] Q. Zhong, L. Ding, J. Liu, B. Du, H. Jin, and D. Tao, “Knowledge graph augmented network towards multiview representation learning for

- aspect-based sentiment analysis,” *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- [17] H. Wang, “Revisiting challenges in data-to-text generation with fact grounding,” in *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 311–322.
- [18] H. Zhou, C. Zheng, K. Huang, M. Huang, and X. Zhu, “Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7098–7108.
- [19] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, “Superglue: A stickier benchmark for general-purpose language understanding systems,” *Advances in neural information processing systems*, vol. 32, 2019.
- [20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.
- [23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [24] R. Nogueira and K. Cho, “Passage re-ranking with bert,” *arXiv preprint arXiv:1901.04085*, 2019.
- [25] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense passage retrieval for open-domain question answering,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.550>
- [26] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, “Ernie: Enhanced representation through knowledge integration,” *arXiv preprint arXiv:1904.09223*, 2019.
- [27] Y. Liu, Y. Wan, L. He, H. Peng, and S. Y. Philip, “Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 6418–6425.
- [28] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, “K-bert: Enabling language representation with knowledge graph,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2901–2908.
- [29] W. Zhou, D.-H. Lee, R. K. Selvam, S. Lee, and X. Ren, “Pre-training text-to-text transformers for concept-centric common sense,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=3k20LAiHYL2>
- [30] D. Yu, C. Zhu, Y. Yang, and M. Zeng, “Jaket: Joint pre-training of knowledge graph and language understanding.”
- [31] J. Menick, M. Trebach, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glæse, S. Young, L. Campbell-Gillingham, G. Irving et al., “Teaching language models to support answers with verified quotes,” *arXiv preprint arXiv:2203.11147*, 2022.
- [32] J. Liu, S. Hallinan, X. Lu, P. He, S. Welleck, H. Hajishirzi, and Y. Choi, “Rainier: Reinforced knowledge introspector for commonsense question answering,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 8938–8958.
- [33] J. Zhang, X. Zhang, J. Yu, J. Tang, J. Tang, C. Li, and H. Chen, “Subgraph retrieval enhanced model for multi-hop knowledge base question answering,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5773–5784.
- [34] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Milligan, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark et al., “Improving language models by retrieving from trillions of tokens,” in *International conference on machine learning*. PMLR, 2022, pp. 2206–2240.
- [35] R. Lebret, D. Grangier, and M. Auli, “Neural text generation from structured data with application to the biography domain,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1203–1213.
- [36] S. Wiseman, S. M. Shieber, and A. M. Rush, “Challenges in data-to-document generation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2253–2263.
- [37] R. Puduppully and M. Lapata, “Data-to-text generation with macro planning,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 510–527, 2021.
- [38] R. Puduppully, Y. Fu, and M. Lapata, “Data-to-text generation with variational sequential planning,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 697–715, 2022.
- [39] R. Puduppully, L. Dong, and M. Lapata, “Data-to-text generation with content selection and planning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6908–6915.
- [40] L. Li, R. Geng, C. Fang, B. Li, C. Ma, B. Li, and Y. Li, “Plan-then-seam: Towards efficient table-to-text generation,” in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 205–219.
- [41] J. Yang, A. Gupta, S. Upadhyay, L. He, R. Goel, and S. Paul, “TableFormer: Robust transformer modeling for table-text encoding,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 528–537. [Online]. Available: <https://aclanthology.org/2022.acl-long.40>
- [42] X. Zhao, T. Fu, C. Tao, and R. Yan, “There is no standard answer: Knowledge-grounded dialogue generation with adversarial activated multi-reference learning,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1878–1891. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.123>
- [43] K. Shuster, M. Komeili, L. Adolphs, S. Roller, A. Szlam, and J. Weston, “Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 373–393. [Online]. Available: <https://aclanthology.org/2022.findings-emnlp.27>
- [44] K. Wang, Z. Li, J. Wang, J. Qu, Y. He, A. Liu, and L. Zhao, “Rt-kgd: Relation transition aware knowledge-grounded dialogue generation,” in *The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings*. Springer, 2022, pp. 319–335.
- [45] R. Sarkar, M. Arcan, and J. P. McCrae, “Kg-cruse: Recurrent walks over knowledge graph for explainable conversation reasoning using semantic embeddings,” in *Proceedings of the 4th Workshop on NLP for Conversational AI*, 2022, pp. 98–107.
- [46] Z. Yuan and A. Vlachos, “Zero-shot fact-checking with semantic triples and knowledge graphs,” *arXiv preprint arXiv:2312.11785*, 2023.
- [47] R. Aly, M. Strong, and A. Vlachos, “Qa-natver: Question answering for natural logic-based fact verification,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 8376–8391.
- [48] K. Xu, L. Wu, Z. Wang, Y. Feng, M. Witbrock, and V. Sheinin, “Graph2seq: Graph to sequence learning with attention-based neural networks,” *arXiv e-prints*, pp. arXiv–1804, 2018.
- [49] M. Kang, J. M. Kwak, J. Baek, and S. J. Hwang, “Knowledge-consistent dialogue generation with knowledge graphs,” in *ICML 2022 Workshop on Knowledge Retrieval and Language Models*, 2022. [Online]. Available: <https://openreview.net/forum?id=McHtKDi5h9>
- [50] S. Moon, P. Shah, A. Kumar, and R. Subba, “Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 845–854.
- [51] Y.-L. Tuan, S. Beygi, M. Fazel-Zarandi, Q. Gao, A. Cervone, and W. Y. Wang, “Towards large-scale interpretable knowledge graph reasoning for dialogue systems,” in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 383–395. [Online]. Available: <https://aclanthology.org/2022.findings-acl.33>
- [52] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang, “A knowledge-enhanced pretraining model for commonsense story generation,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 93–108, 2020.
- [53] D. Yu, C. Zhu, Y. Yang, and M. Zeng, “Jaket: Joint pre-training of knowledge graph and language understanding,” in *Proceedings of the*

- AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 630–11 638.
- [54] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
- [55] R. Puduppully, L. Dong, and M. Lapata, “Data-to-text generation with entity modeling,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2023–2035. [Online]. Available: <https://aclanthology.org/P19-1195>
- [56] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf
- [57] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Advances in neural information processing systems*, vol. 27, 2014.
- [58] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 3776–3784. [Online]. Available: <https://doi.org/10.1609/aaai.v30i1.9883>
- [59] J. Sun. (2012) Jieba chinese word segmentation tool. [Online]. Available: <https://github.com/fxsjy/jieba>
- [60] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: <https://openreview.net/forum?id=Bkg6RiCqY7>
- [61] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: <https://openreview.net/forum?id=SkeHuCVFDr>
- [62] Y. Bengio, R. Ducharme, and P. Vincent, “A Neural Probabilistic Language Model,” vol. 13, 2000.
- [63] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.
- [64] Y. Shao, Z. Geng, Y. Liu, J. Dai, F. Yang, L. Zhe, H. Bao, and X. Qiu, “Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation,” *arXiv preprint arXiv:2109.05729*, 2021.