IMPROVING MEDICAL DIALOGUE GENERATION WITH ABSTRACT MEANING REPRESENTATIONS

Bohao Yang[1,2*], *Chen Tang*[3*], *Chenghua Lin*[1,2†]

[1]Department of Computer Science, The University of Manchester, UK
[2]Department of Computer Science, The University of Sheffield, UK
[3]Department of Computer Science, The University of Surrey, UK
bohaoyang217@gmail.com, chen.tang@surrey.ac.uk
chenghua.lin@manchester.ac.uk

ABSTRACT

Medical Dialogue Generation plays a critical role in telemedicine by facilitating the dissemination of medical expertise to patients. Existing studies focus on incorporating textual representations, which have limited their ability to represent text semantics, such as ignoring important medical entities. To enhance the model's understanding of textual semantics and medical knowledge including entities and relations, we introduce Abstract Meaning Representations (AMR) to construct graphical representations that delineate the roles of language constituents and medical entities within dialogues. In this paper, we propose a novel neural framework that models dialogues between patients and healthcare professionals using AMR graphs, where the framework incorporates both textual and graphical knowledge with a dual attention mechanism. Experimental results show that our framework outperforms robust baseline models in medical dialogue generation, demonstrating the effectiveness of AMR graphs in enhancing the representation of medical knowledge and logical relationships. Furthermore, to support future research in this domain, we provide the corresponding source code at **https://github.com/Bernard-Yang/MedDiaAMR**.

***Index Terms*—** Abstract Meaning Representation, Dialogue Generation, Language Model, Artificial Intelligence, AMR Graph

1. INTRODUCTION

The goal of telemedicine is to provide patients with digital access to medical information, particularly in situations where direct access to a medical professional may be limited [1, 2]. Prior research in this field has predominantly focused on incorporating medical knowledge by leveraging various types of additional annotations, including frequent items [3], named entities [4], entity relations [5], etc. However, unlike open-domain dialogue generation, the sequential features provided by text annotations struggle to comprehensively represent the intricate medical grounds and principles of diagnosis contained within medical dialogues. To address this limitation, and to facilitate the incorporation of medical entities and their relations, our approach aims to construct dialogue-level Abstract Meaning Representation (AMR) graphs [6], and exploit both textual and graph-based features to enhance the language model on medical dialogue generation.

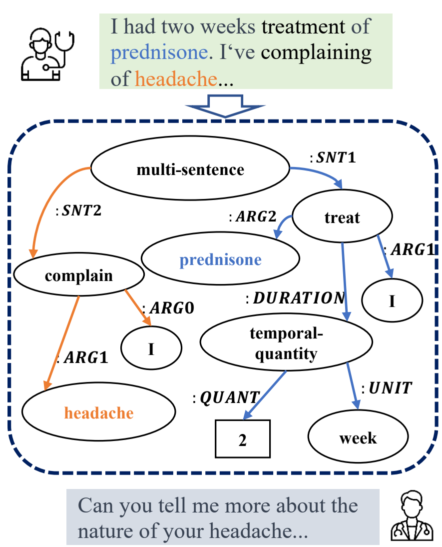

Fig. 1. An example showing how an AMR graph represents the dialogues of a patient. Terminologies from different sentences are shown in blue and orange colour.

As shown in Figure 1, AMR graphs provide a structured and semantically rich representation of language [6]. In the complex and critically important field of medicine, clear and precise communication is paramount. AMR graphs offer a standardised way to capture the relationships between words, entities, and their corresponding meanings, reducing ambiguity and potential misunderstandings in medical conversations. This enables medical professionals and patients to more easily interpret and trust the information conveyed, facilitating better decision-making, treatment adherence, and overall patient care. The incorporation of AMR graphs into the dialogue generation system enhances its capacity to comprehend intricate semantics and contextual nuances implicitly embedded within textual content. Consequently, this integration

*Equal contribution.
†Corresponding author.

empowers the system to generate context-aware medical dialogues. Our framework benefits vanilla language models by achieving more precise and naturally articulated medical dialogue generation through incorporating textual representations with the rich graph knowledge encapsulated by AMR graphs.

In this study, we first construct AMR graphs by parsing the sentences within each patient's dialogue. Subsequently, these parsed AMR graphs are flattened and fed into a graph encoder, aligning with an independent sequence encoder for the text tokens in the input sentences. A module implemented by the dual-attention mechanism [6] is employed to incorporate the heterogeneous features originating from both the AMR graphs and input texts. This combined representation is then used for the subsequent response decoding in an autoregressive manner. Our experimental results demonstrate that our approach substantially enhances the performance of the original language model and achieves state-of-the-art performance, by capturing the additional structured knowledge during medical dialogue generation. Our contributions can be summarised as follows: (1) This is the first attempt to exploit AMR graphs for improving medical dialogue generation. (2) We propose a novel framework that incorporates both AMR graphs and text for medical dialogue generation and achieves state-of-the-art performance; (3) We conduct comprehensive experiments to illustrate the effectiveness of our approach and provide a thorough analysis of the main components.

2. METHODOLOGY

Our proposed framework is illustrated in Figure 2, which incorporates the heterogeneous features of the input text and parsed AMR graphs with two independent encoders. Subsequently, the decoder attends to the dual-attention features from the encoders to autoregressively predict response tokens.

2.1. Task Definition

We define the task as follows: The input sequence data is denoted as $X = x_1, x_2, ..., x_n$, which encompasses a medical inquiry, in addition to the historical dialogue between the healthcare provider and the patient.. In addition, the input AMR graphs of patient's inquiry is denoted as $G = g_1, g_2, ..., g_n$. The primary objective of this task is to generate a response $Y = y_1, y_2, ..., y_m$ by learning the conditional probability distribution $P(Y|X, G)$ of the dataset. This formulation encapsulates the essence of our research endeavor, which revolves around generating doctor-like responses in a medical dialogue context.

2.2. Sequence Encoding

The Sequence encoder employed in this study adheres to the conventional Transformer architecture [7], which is designed to take the input patient's inquiry, denoted as $\mathbf{S}_i = \{w_1, w_2, ..., w_\mathbf{S}\}$, and subsequently generates a corresponding sentence representation, denoted as $\mathbf{H}_S$. Formally, the sequence encoder is defined as follows:

$$\mathbf{H}_S = \mathrm{Transformer}(\mathbf{S}) \tag{1}$$

$$h_i = \sum_{j=1}^{\mathbf{S}} \alpha_{ij} \left(W^H h_j\right) \tag{2}$$

$$\alpha_{ij} = \mathrm{Attention}(h_i, h_j) \tag{3}$$

where $\mathbf{H}_S = \{h_1, h_2, ..., h_\mathbf{S}\}$, w_i represents the i-th tokens, $\mathbf{S}$ signifies the sequence lengths and W^H represents learnable parameters.

2.3. Graph Encoding

We employ a Graph Transformer [8] to encode AMR graphs. An AMR graph $\mathbf{G} = \{\mathbf{N}, \mathbf{R}\}$ consists of graph nodes denoted by $\mathbf{N}$ and graph edges denoted by $\mathbf{R}$. Each edge $\mathbf{e} \in \mathbf{E}$ comprises of a set of elements $\{\mathbf{n_i}, \mathbf{r_{ij}}, \mathbf{n_j}\}$, symbolising the relation $\mathbf{r_{ij}}$ between two graph nodes $\mathbf{n_i}$ and $\mathbf{n_j}$. The Graph encoder processes these nodes and relations as input and is formally defined as follows:

$$\mathbf{H}_G = \mathrm{GraphEncoder}(\mathbf{N}, \mathbf{E}) \tag{4}$$

$$h'_i = \sum_{j=1}^{M} \hat{\alpha}_{ij} \left(W^V h'_j + W^R \boldsymbol{r}_{ij}\right) \tag{5}$$

where $\mathbf{H}_G = \{h'_1, h'_2, ..., h'_M\}$, and W^V and W^R are learnable parameters.

The graph attention of the Graph Transformer module is formally represented as:

$$\begin{aligned} \hat{\alpha}_{ij} &= \frac{\exp(\hat{e}_{ij})}{\sum_{m=1}^{M} \exp(\hat{e}_{im})} \\ \hat{e}_{ij} &= \frac{\left(W^Q h'_i\right)^T \left(W^K h'_j + W^R \boldsymbol{r}_{ij}\right)}{\sqrt{d}} \end{aligned} \tag{6}$$

where W^Q and W^K are learnable parameters and d is hidden state size. The Graph Transformer effectively encodes structural information, represented by $\mathbf{r_{ij}}$, for all pairs of nodes within the AMR graphs. This incorporation of graph edge information enriches the node representations and enhances the overall encoding process.

2.4. Dual Attention Decoding

Once we obtained the sentence representation $\mathbf{H}_S$ and graph representation $\mathbf{H}_G$, we proceed to feed them into the decoder equipped with a dual attention mechanism. For each decoder hidden state d_t, the dual-attention mechanism produce a sequence context vector c_{tS} and a graph context vector c_{tG} at each time step t :

$$c_{ti} = \sum_{i=1}^{S} \hat{\alpha}_{ti} h_i \tag{7}$$

$$\hat{\alpha}_{ti} = \mathrm{Attention}(d_t, h_i) \tag{8}$$

$$c_{tj} = \sum_{j=1}^{M} \hat{\alpha}_{tj} h'_j \tag{9}$$

$$\hat{\alpha}_{tj} = \mathrm{Attention}\left(d_t, h'_j\right) \tag{10}$$

Subsequently, we concatenate the sequence context vector c_{tS} and the graph context vector c_{tG} to compose the ultimate context vector

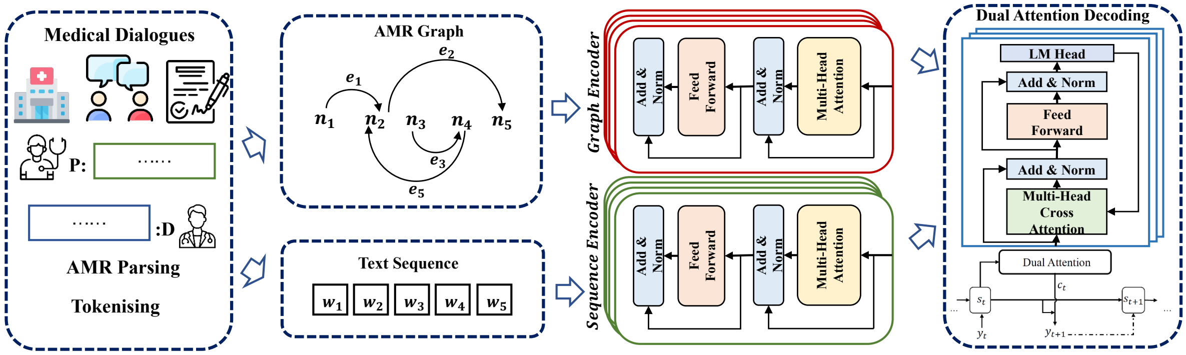

Fig. 2. The overview of our proposed framework.

$\hat{c}_t$. This combination is formally represented as:

$$c_t = W^C[c_{tS}; c_{tG}] + b \tag{11}$$

where W^C and b are learnable parameters.

2.5. Training and Inference

Finally, the fused sentence and graph features are autoregressively decoded to predict responses. The predicted tokens are forced to closely match the golden responses. We train the entire model using the following loss function:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\log P(Y \mid X, G) \tag{12}$$

where N denotes the size of the training data, and $\mathcal{L}$ is the cross entropy between the predicted response tokens and the golden responses.

3. EXPERIMENT

3.1. Experimental Setup

Data Preparation. We use the biomedical dialogue generation dataset from [1] as the raw dataset. In order to prepare the input data for the Graph Transformer, We adopt an open-source pre-trained AMR parser in transform utterances into corresponding AMR. Subsequently, we streamline the acquired AMR graphs using the AMR simplifier [9]. This simplification process allows us to extract pertinent concepts and relationships, which are pivotal components for our subsequent analytical endeavors.

Baselines. We comprehensively conduct a comparative analysis of our proposed framework against competitive language models used in several recent advances * [10, 11, 12, 13], as well as those relevant to our task [14, 15]. The models under consideration are as follows: BART [16]: A widely used language model renowned for text generation tasks. T5 [17]: A language model distinguished by its encoder-decoder architecture, successfully adapted for diverse generation tasks. GPT-2 [18]: A popular pre-trained language model widely applied in dialogue generation tasks. DialoGPT [19]: A dialogue-oriented pre-trained GPT-2 model known for its strong performance in dialogue generation tasks. Term-BART [19]: An advanced framework specifically tailored for medical dialogue generation, representing the state-of-the-art in the field.

Metrics. To assess the efficacy of our proposed framework, we employ a range of both referenced and unreferenced metrics in our experiments. BLEU (B-n) [20] and ROUGE (R-n) [21], including the longest common subsequence variant (R-L), gauge the quality of generated responses by assessing their n-gram overlaps with reference responses. Additionally, we employ the metric of response diversity, following the approach outlined in [22], by quantifying the n-gram distinction, denoted as Diversity (D-n), within generated responses, where n signifies the n-gram order. This comprehensive set of metrics ensures a rigorous evaluation of our framework's performance and its comparison to the baseline models.

3.2. Implementation Details

The pre-trained models for baselines employed in this study originate from publicly accessible checkpoints hosted on the Hugging Face platform†. All the models undergo a training process spanning a maximum of 10 epochs, executed on a Tesla A100 computational unit, over a duration of approximately one full day.

Our training configuration prescribes a batch size of 36, with a learning rate of $1e^{-4}$, and employs the Adam optimizer for the training procedure. These meticulous specifications underpin the rigorous methodology adopted in our academic inquiry.

3.3. Experimental Results

As shown in Table 1, our proposed framework outperforms all baseline models across both referenced and unreferenced evaluation metrics. This outcome underscores the positive effect of introducing AMR on enhancing the knowledge incorporating capabilities of language models. Specifically, the marked enhancement in referenced metrics, such as BLEU and ROUGE, indicates the model's capacity to generate responses that encapsulate medical knowledge grounds as the gold references do.

In addition, there is a substantial improvement on the unreferenced metrics, i.e., Dist-1 and Dist-4. This signifies that the diversity

*Our approach can also be applied to larger language models, e.g. ChatGPT, but to be fair, some extremely large language models (ChatGPT has around 50 times parameter size than our baselines.) are not included in comparisons.

†https://huggingface.co/models

Model	B-1↑	B-2↑	B-3↑	B-4↑	R-1↑	R-2↑	R-L↑	Dist-1↑	Dist-2↑	Dist-3↑	Dist-4↑
GPT-2	0.0725	0.0376	0.0267	0.0218	12.5431	4.3497	9.8125	0.0048	0.0245	0.0511	0.0725
DialoGPT	0.0599	0.0310	0.0225	0.0187	9.9041	3.5320	8.0158	0.0036	0.0169	0.0354	0.0531
T5-base	0.1255	0.0585	0.0319	0.0188	12.7913	2.1351	9.8115	0.0026	0.0102	0.0183	0.0265
T5-large	0.0952	0.0529	0.0388	0.0321	16.1452	5.8471	12.3637	0.0055	0.0282	0.0593	0.0883
BART-base	0.1205	0.0631	0.0423	0.0321	19.3620	5.1807	11.4057	0.0046	0.0374	0.1125	0.2087
BART-large	0.1142	0.0621	0.0420	0.0319	19.4735	5.3347	11.4324	0.0055	0.0435	0.1131	0.1933
Terms-BART	0.1547	0.0822	0.0555	0.0421	20.2111	5.4167	13.0137	0.0071	**0.0453**	**0.1462**	0.2899
Ours	**0.1705**	**0.1411**	**0.1613**	**0.1336**	**35.7853**	**14.5550**	**23.9057**	**0.0088**	0.0179	0.1376	**0.4321**
- w/o text	0.1222	0.0716	0.0620	0.0336	29.8947	9.1014	18.4940	0.0062	0.0082	0.0448	0.1626
- w/o AMR	0.0802	0.0432	0.0304	0.0221	24.1742	5.4295	12.9662	0.0042	0.0059	0.0376	0.0904

Table 1. The automatic evaluation compares the performance of our model with various baseline systems on medical dialogue generation. The best-performing model for each metric is highlighted in bold.

of generated text is also improved via the incorporation of the graph representations. We posit that this outcome can be attributed to the improved flexibility inherent in the graph representations of the dialogue context. It is noticeable that the BART baselines have better Dist-1 and Dist-3 because they generate more repeated short phrases in dialogues, e.g., "I am"..

Furthermore, our experiments also investigate the impact of model size, ranging from base to large variant. While it is intuitive to expect performance improvements with larger models, empirical results on various metrics reveal that models with larger parameters, such as BART-large and T5-large, do not consistently outperform their smaller counterparts like BART-base and T5-base. This suggests that merely increasing model size does not guarantee performance enhancements. Existing language models primarily encoding the superficial structure of dialogues face substantial challenges in incorporating domain-specific knowledge without explicitly modeling medical concepts and relationships. Hence, our utilisation of AMR graphs serves as a crucial bridge to address this gap in medical concept representations.

In the ablation study, the results further demonstrates effectiveness of introducing AMR graphs as an additional feature to language modeling. our proposed model encodes heterogeneous features of both text and AMR in a graph format showing significant improvements across all evaluated metrics. Notably, several metrics exhibit over two-fold increments such as B-4 which experiences a substantial rise from approximately 0.0221 to 0.1336 and Dist-4 increases from approximately 0.09 to 0.43. Another finding is certain metrics (i.e. BLEU) indicate that our proposed model achieves scores surpassing mathematical sum of ablated models ("-w/o text" and "-w/o AMR"). This implies these two features encapsulate distinct semantic aspects of the dialogues. Therefore, integration of both graphical and textual features profoundly enhances process of medical dialogue generation.

For qualitative analysis, we present the generated dialogues from different models in Table 2. It can be observed that even though the previous state-of-the-art model, **Terms-BART**, uses more terminological knowledge when answering the question, dialogues generated by **ours** address the patient's queries more effectively. The terminologies in the response are more related to the context. This improvement can be attributed to our model's enhanced capability in capturing semantics and logical structures, which are attained from the additional AMR graphs. The performance improvement of our model is further demonstrated via a comparison with the response generated by a variant of the model that does not utilise AMR representations (**- w/o AMR**).

Context: Patient: I have a **sharp pain** at the front of my **hip** at the top of my **thigh** where it nears my **lower abdomen**. This **pain** only occurs when i bend over and lasts there after for a few seconds. It then goes away and does not recur until i bend over again. Any idea what this could be? maybe a hernia?
Reference: Hello, it is unlikely be an hernia. The **symptoms** are more likely suggestive of **musculo-skeletal causes** like **sprain** or contusion. As of now you can] use **analgesics/anti inflammatory combination** like **aceclofenac/seratiopeptdase** for **symptomatic relief**. You can go for an **ultrasound scan** to rule out **hernia**. Hope i have answered your query. Let me know if i can assist you further. Wishing you all the best. Regards, dr. shinas hussain, general & family physician.
Terms-BART: Hello and welcome to 'ask a doctor' service. I have reviewed your query and here is my advice. As per your case history of hip **pain**, it could be due to **hip muscle spasm**. For this reason, I would recommend consulting with your attending **physician** for a **physical exam** and some **tests**. Hope I have answered your query. Let me know if I can assist you further. Regards, Dr. Praveen Tayal, orthopaedic surgeon apollo hospitals.
Ours: Hello, it is unlikely be an **hernia**. You can use **analgesics/anti inflammatory** combination like **aceclofenac/seratiopeptdase** for symptomatic relief. Regards, dr. shinas hussain, general & family physician.
- w/o text: Hi, you can take drugs like **ibuprofen** for pain relief. You can take for a **mri abdomen**. Hope i have answered your query. Let me know if i can assist you further. dr. shinas hussain, & family physician.
- w/o AMR: Hi, thanks is a to due infection of pain. You have take **ibuprofen** for a relief. You can take for a **mri**. Hope i have answered your query. Let me know if i can assist you further. Regards, dr. shinas hussain, general & family physician.

Table 2. A Sample collected for the case study, where terminologies are highlighed in **bold**.

4. CONCLUSION

In conclusion, our approach is the first to integrates Abstract Meaning Representations (AMR) graphs to capture the semantics and medical terminologies embedded within dialogues, offers a novel and effective framework for modeling patient-doctor dialogues. By incorporating the textual and graphical knowledge into a unified language model, our proposed framework achieves state-of-the-art performance.

5. REFERENCES

[1] Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin, "Terminology-aware medical dialogue generation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] Noor Fazilla Abd Yusof, Chenghua Lin, and Frank Guerin, "Analysing the causes of depressed mood from depression vulnerable individuals," in *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, 2017, pp. 9–17.

[3] Azadeh Givchi, Reza Ramezani, and Ahmad Baraani-Dastjerdi, "Graph-based abstractive biomedical text summarization," *Journal of Biomedical Informatics*, vol. 132, pp. 104099, 2022.

[4] Keqin Peng, Chuantao Yin, Wenge Rong, Chenghua Lin, Deyu Zhou, and Zhang Xiong, "Named entity aware transfer learning for biomedical factoid question answering," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 2365–2376, 2021.

[5] Yue Shang, Yanpeng Li, Hongfei Lin, and Zhihao Yang, "Enhancing biomedical text summarization using semantic relation extraction," *PLoS one*, vol. 6, no. 8, pp. e23862, 2011.

[6] Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang, "Semantic representation for dialogue modeling," 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)*, Online, Aug. 2021, pp. 4430–4445.

[7] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[8] Jiehan Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou, "Modeling graph structure in transformer for better amr-to-text generation," in *Conference on Empirical Methods in Natural Language Processing*, 2019.

[9] Ioannis Konstas, Srini Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer, "Neural amr: Sequence-to-sequence models for parsing and generation," in *Annual Meeting of the Association for Computational Linguistics*, 2017.

[10] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie, "MedDialog: Large-scale medical dialogue datasets," in *Proceedings of EMNLP*, Nov. 2020.

[11] Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric Xing, and Pengtao Xie, "On the generation of medical dialogs for COVID-19," in *Proceedings of ACL (Volume 2: Short Papers)*, Aug. 2021.

[12] Dingmin Wang, Chenghua Lin, Qi Liu, and Kam-Fai Wong, "Fast and scalable dialogue state tracking with explicit modular decomposition," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021, pp. 289–295.

[13] Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin, "Enhancing dialogue generation via dynamic graph knowledge aggregation," in *Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, July 2023, pp. 4604–4616.

[14] Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui, "Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information," *ArXiv*, vol. abs/2305.16967, 2023.

[15] Chen Tang, Chenghua Lin, Henglin Huang, Frank Guerin, and Zhihao Zhang, "EtriCA: Event-triggered context-aware story generation augmented by cross attention," in *Findings of EMNLP 2022*, Dec. 2022.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter 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.

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[19] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," *arXiv preprint arXiv:1911.00536*, 2019.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318.

[21] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "A diversity-promoting objective function for neural conversation models," *arXiv:1510.03055*, 2015.