Retrieve, Reason, and Refine: Appendix of Generating Accurate and Faithful Patient Instructions

Fenglin Liu[1]
[1]Department of Engineering Science, University of Oxford
fenglin.liu@eng.ox.ac.uk

A Medical Knowledge Graph

In our work, we construct an off-the-shelf medical knowledge graph $\mathcal{G} = (V, E)$ ($V = \{v_i\}_{i=1:N_{\text{KG}}} \in \mathbb{R}^{N_{\text{KG}} \times d}$ is a set of nodes and $E = \{e_{i,j}\}_{i,j=1:N_{\text{KG}}}$ is a set of edges), which models the domain-specific knowledge structure, to explore the medical knowledge [11, 18]. In implementation, we consider all clinical codes (including diagnose codes, medication codes, and procedure codes) during hospitalization as nodes, i.e., each clinical code corresponds to a node in the graph. The edge weights are calculated by the normalized co-occurrence of different nodes computed from training corpus. Figure 1 gives an illustration of the constructed medical knowledge graph. It is worth noting that more complex graph structures could be constructed by using more large-scale external medical textbooks. Therefore, our approach is not limited to the currently constructed graph and could provide a good basis for the future research of Patient Instruction generation.

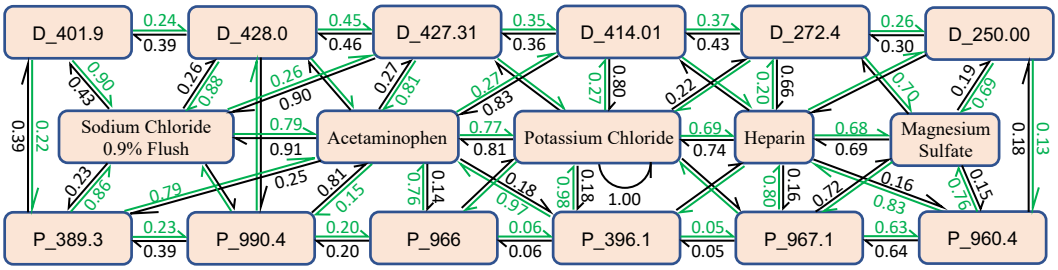

Figure 1: The constructed medical knowledge graph. Each clinical code corresponds to a node in the graph. We present the most frequent 6 diagnose nodes (the first row), 5 medication nodes (the second row), and 6 procedure nodes (the third row), and parts of their edge weights. Please refer to Table 1 for the exact meanings of these diagnose and procedure nodes.

Table 1: The exact meanings of the most frequent diagnose and procedure nodes in Figure 1.

#	Diagnose Nodes	Procedure Nodes
1	**D_401.9**: Unspecified essential hypertension	**P_389.3**: Venous catheterization, not elsewhere classified
2	**D_428.0**: Congestive heart failure, unspecified	**P_990.4**: Transfusion of packed cells
3	**D_427.31**: Atrial fibrillation	**P_966**: Enteral infusion of concentrated nutritional substances
4	**D_414.01**: Coronary atherosclerosis of native coronary artery	**P_396.1**: Extracorporeal circulation auxiliary to open heart surgery
5	**D_272.4**: Other and unspecified hyperlipidemia	**P_967.1**: Continuous invasive mechanical ventilation for less than 96 consecutive hours
6	**D_250.00**: Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled	**P_960.4**: Insertion of endotracheal tube

For the constructed knowledge graph, we use randomly initialized embeddings $H^{(0)} = \{v_1, v_2, \ldots, v_{N_{\text{KG}}}\} \in \mathbb{R}^{N_{\text{KG}} \times d}$ to represent all node features. To obtain the final Medical Knowledge

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Graph $G_{\text{Pr}} = \{v'_1, v'_2, \ldots, v'_{N_{\text{KG}}}\} \in \mathbb{R}^{N_{\text{KG}} \times d}$, we adopt graph convolution layers [15, 4, 3] to encode the graph $\mathcal{G} = (V, E)$, which is defined as follows:

$$H^{(l+1)} = \text{ReLU}(\hat{A}\hat{D}^{-1}H^{(l)}W^{(l)} + b^{(l)}), \quad l \in [0, L-1] \tag{1}$$

where ReLU denotes the ReLU activation function, $\hat{A} = A+I$ is the adjacency matrix $A \in \mathbb{R}^{N_{\text{KG}} \times N_{\text{KG}}}$ of the graph $\mathcal{G}$ with added self-connections, $I \in \mathbb{R}^{N_{\text{KG}} \times N_{\text{KG}}}$ is the identity matrix, $\hat{D} \in \mathbb{R}^{N_{\text{KG}} \times N_{\text{KG}}}$ is the out-degree matrix where $D_{ii} = \sum_j A_{ij}$, $W^{(l)} \in \mathbb{R}^{d \times d}$ and $b^{(l)} \in \mathbb{R}^d$ are trainable parameters, and L is the number of layers. We empirically set $L = 1$ and regard $H^{(1)} = \{v'_1, v'_2, \ldots, v'_{N_{\text{KG}}}\} \in \mathbb{R}^{N_{\text{KG}} \times d}$ as the medical knowledge $G_{\text{Pr}} \in \mathbb{R}^{N_{\text{KG}} \times d}$ in our Re^{3}Writer.

B Effect of the Number of Retrieved Instructions

Table 2 shows that all variants with different number of retrieved instructions N_{P} can consistently outperform the baseline model, which proves the effectiveness of our approach in retrieving the working experience to boost the Patient Instruction generation. In particular, when the number of retrieved instructions N_{P} is 20, the model gets the highest performance, explaining the reason why the value of N_{P} is set to 20 in our Re^{3}Writer. For other variants, we speculate that when N_{P} is set to small values, the model will suffer from the inadequacy of information. When N_{P} is set to large values, retrieving more patient instructions will bring more irrelevant noise to the model, impairing the performance.

Table 2: Effect of the number of retrieved instructions N_{P} in our Retrieve module when retrieving the working experience.

N_{P}	Dataset: Patient Instruction (PI)							
	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baseline	19.9	39.0	20.3	37.1	41.6	32.5	27.9	25.1
5	20.6	40.5	**21.9**	38.4	41.7	33.2	28.9	26.3
10	20.7	40.6	**21.9**	38.5	42.4	33.6	29.3	26.5
20	**20.9**	**40.8**	**21.9**	**38.6**	**43.2**	**34.2**	**29.7**	**26.8**
30	20.5	40.4	21.8	38.3	42.0	33.4	29.0	26.3
50	20.3	40.1	21.5	37.9	41.8	33.1	28.8	26.0

C Multi-Head Attention and Feed-Forward Network

In recent years, Transformer [17, 6], which includes the Multi-Head Attention (MHA) and the Feed-Forward Network (FFN), achieves the state-of-the-art performances on multiple natural language generation tasks. The MHA consists of n parallel heads and each head is defined as a scaled dot-product attention:

$$\begin{aligned} \text{Att}_i(X, Y) &= \text{softmax}\left(\frac{X\text{W}_i^{\text{Q}}(Y\text{W}_i^{\text{K}})^T}{\sqrt{d_n}}\right) Y\text{W}_i^{\text{V}} \\ \text{MHA}(X, Y) &= [\text{Att}_1(X, Y); \ldots; \text{Att}_n(X, Y)]\text{W}^{\text{O}} \end{aligned} \tag{2}$$

where $X \in \mathbb{R}^{l_x \times d}$ and $Y \in \mathbb{R}^{l_y \times d}$ denote the Query matrix and the Key/Value matrix, respectively; $\text{W}_i^{\text{Q}}, \text{W}_i^{\text{K}}, \text{W}_i^{\text{V}} \in \mathbb{R}^{d \times d_n}$ and $\text{W}^{\text{O}} \in \mathbb{R}^{d \times d}$ are learnable parameters, where $d_n = d/n$; $[\cdot;\cdot]$ stands for concatenation operation.

Following the MHA is the FFN, defined as follows:

$$\text{FFN}(x) = \max(0, x\text{W}_{\text{f}} + \text{b}_{\text{f}})\text{W}_{\text{ff}} + \text{b}_{\text{ff}} \tag{3}$$

where $\max(0, *)$ represents the ReLU activation function; $\text{W}_{\text{f}} \in \mathbb{R}^{d \times 4d}$ and $\text{W}_{\text{ff}} \in \mathbb{R}^{4d \times d}$ denote learnable matrices for linear transformation; b_{f} and b_{ff} represent the bias terms. It is worth noting that both the MHA and FFN are followed by an operation sequence of dropout [16], residual connection [2, 8], and layer normalization [1].

References

[1] L. J. Ba, R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[3] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[4] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel. Gated graph sequence neural networks. In *ICLR*, 2016.

[5] F. Liu, X. Ren, Y. Liu, H. Wang, and X. Sun. simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In *EMNLP*, 2018.

[6] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun. Aligning visual regions and textual concepts for semantic-grounded image representations. In *NeurIPS*, 2019.

[7] F. Liu, X. Ren, X. Wu, S. Ge, W. Fan, Y. Zou, and X. Sun. Prophet attention: Predicting attention with future attention. In *NeurIPS*, 2020.

[8] F. Liu, X. Ren, Z. Zhang, X. Sun, and Y. Zou. Rethinking skip connection with layer normalization. In *COLING*, 2020.

[9] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou. Federated learning for vision-and-language grounding problems. In *AAAI*, 2020.

[10] F. Liu, S. Ge, and X. Wu. Competence-based multimodal curriculum learning for medical report generation. In *ACL/IJCNLP*, 2021.

[11] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, 2021.

[12] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun. Contrastive attention for automatic chest x-ray report generation. In *ACL (Findings)*, 2021.

[13] F. Liu, C. You, X. Wu, S. Ge, S. Wang, and X. Sun. Auto-encoding knowledge graph for unsupervised medical report generation. In *NeurIPS*, 2021.

[14] F. Liu, X. Wu, C. You, S. Ge, Y. Zou, and X. Sun. Aligning source visual and target language domains for unpaired video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[15] D. Marcheggiani and I. Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *EMNLP*, pages 1506–1515, 2017.

[16] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, pages 1929–1958, 2014.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[18] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. L. Yuille, and D. Xu. When radiology report generation meets knowledge graph. In *AAAI*, 2020.