# Incorporating Medical Knowledge to Transformer-based Language Models for Medical Dialogue Generation

**Usman Naseem[1], Ajay Bandi[2], Shaina Raza[3], Junaid Rashid[4], Bharathi Raja Chakravarthi[5]**
[1]School of Computer Science, University of Sydney, Australia
[2]School of Computer Science and Information Systems, Northwest Missouri State University Maryville, USA
[3]The Dalla Lana School of Public Health, University of Toronto Toronto, Canada
[4]Department of Computer Science and Engineering, Kongju National University, South Korea
[5] Data Science Institute, National University of Ireland Galway, Ireland

## Abstract

Medical dialogue systems have the potential to assist doctors in expanding access to medical care, improving the quality of patient experiences, and lowering medical expenses. The computational methods are still in their early stages and are not ready for widespread application despite their great potential. Existing transformer-based language models have shown promising results but lack domain-specific knowledge. However, to diagnose like doctors, an automatic medical diagnosis necessitates more stringent requirements for the rationality of the dialogue in the context of relevant knowledge. In this study, we propose a new method that addresses the challenges of medical dialogue generation by incorporating medical knowledge into transformer-based language models. We present a method that leverages an external medical knowledge graph and injects triples as domain knowledge into the utterances. Automatic and human evaluation on a publicly available dataset demonstrates that incorporating medical knowledge outperforms several state-of-the-art baseline methods.

## 1 Introduction

Medical dialogue systems, which have gained increasing attention, aim to communicate with patients to enquire about diseases beyond their self-reported and make an automatic diagnosis (Wei et al., 2018; Xu et al., 2019; Lin et al., 2019). It has the potential to substantially automate the diagnostic process while also lowering the cost of gathering information from patients (Kao et al., 2018). In addition, preliminary diagnosis findings that are generated by a medical dialogue system may help doctors make a diagnosis more quickly. Because of these advantages, researchers work on addressing sub-problems in a medical dialogue system, such as natural language understanding (Lin et al., 2019; Shi et al., 2020).

However, the dialogue system for medical diagnosis, on the other hand, has specific requirements for dialogue reasoning in the context of medical knowledge. The diagnosis elicited by the dialogue system should be associated with the underlying medical condition and coherent with medical knowledge. In the absence of medical knowledge, traditional generative dialogue models frequently use neural sequence modelling (Sutskever et al., 2014; Vaswani et al., 2017) and cannot be directly applied to the medical dialogue scenario.


Figure 1: An example of medical dialogue between a patient (left) and a doctor (right).

Recently, transformer-based language models (LMs) (Devlin et al., 2019; Radford et al., 2019; Song et al., 2019) are fine-tuned for medical dialogue tasks. Zeng et al. (2020) collected a MedDialog dataset and fine-tuned various transformer-based LMs which includes a vanilla transformer (Vaswani et al., 2017), GPT (Radford et al., 2019) and BERT-GPT (Wu et al., 2020; Lewis et al., 2020) for medical dialogue generation task. Yang et al. (2020), in another study, presented a CovidDialog dataset and then train dialogue generation models based on Transformer, GPT-based model, and BART (Lewis et al., 2020) and BERT-GPT for medical dialogue generation tasks. These LMs are trained on huge corpus but may not provide a good representation of specific domains (Müller et al., 2020) and need an adequate amount of task-specific data (Dou et al., 2019) in order to establish correlations between diseases and symptoms (see Figure 1). Instead of using publicly available models, we can pre-train a model that emphasizes domain-specificity. On the other hand, pre-training is time-intensive and computationally costly, making it unavailable for most users.

Furthermore, while it is possible to inject



*Proceedings of the BioNLP 2022 workshop, Dublin, Ireland*, pages 110–115
May 26, 2022. 

domain-specific knowledge into LMs during pre-training, this method of acquiring knowledge can be expensive and inefficient. For instance, pre-training data must contain many occurrences of the words "Panadol" and "headache" occurring together for the model to learn that "Panadol" can treat headaches. What other options do we have to make the model an expert in its field besides this one? The knowledge graph (KG), also known as an ontology, was a good solution in the early stages of research. SNOMED-CT (Bodenreider, 2008), in the medical field, and HowNet (Dong et al., 2010), in the field of Chinese conception, are two examples of KGs developed as knowledge was distilled into a structured form. If KG can be incorporated into the LM, it will provide domain knowledge to the computational method, enhancing its effectiveness on domain-specific tasks while significantly lowering the expense of pre-training. To address the limitations mentioned above, this article describes a method for incorporating domain-specific external knowledge into transformer-based LMs for medical dialogue generation tasks. Our contributions are as follows:

- We presented a new method that incorporates medical knowledge to transformer-based language models;
- The proposed method first injects knowledge from a medical knowledge graph into an utterance. Next, the embedding layer transforms the utterance tree into an embedding that is fed to the masked self-attention of a transformer, followed by the decoder to generate the response.
- To evaluate the performance of the proposed method, we performed both automatic and human evaluations. Our results demonstrated that incorporating medical knowledge improves the performance compared to several state-of-the-art baselines on the MedDialog dataset.

## 2 Methodology

**Problem Definition**: Given a dialogue, we process a patient-doctor dialogue as a set of pairs $\{(s_i, t_i)\}$, where source $s_i$ is the dialogue from a patient and target $t_i$ is a doctor's response. A dialogue generation model generates *t* from *s*.

**Overview of Architecture**: As illustrated in Figure 2, the proposed method contains four modules, i.e., knowledge layer, embedding layer, masked transformer encoder, where we extend self-attention to mask-self attention, and transformer decoder. Our knowledge layer injects relevant triples into an input utterance (i.e., conversation) from a KG, converting it to a knowledge-rich utterance tree. Simultaneously, the utterance tree is fed into the embedding layer for token-level representation. The representation from an embedding layer is fed to the masked transformer encoder and decoder to generate a response. We will describe each of these modules in detail in the following discussion.

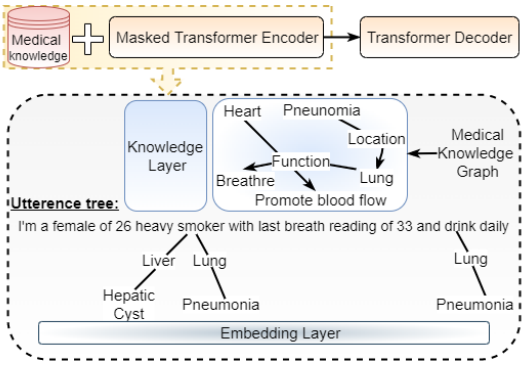

Figure 2: Overall architecture of proposed method

### 2.1 Knowledge layer

The knowledge layer incorporates domain-specific (medical) knowledge into utterances and transforms them into utterance trees. The knowledge layer generates an utterance tree given an input utterance (s) and a KG. This method involves two stages: query of medical knowledge, referred to as K-Query, and injection of knowledge, referred to as K-Inject. K-Query extracts all entity names from the utterance *s* and queries their correlating triples from knowledge *k*. K-Query can be expressed as follows:

$$E = K\_Query(s, KG), \quad (1)$$

Where *E* is a set of associating triples. K-Inject then injects the queried *E* into the utterance *s* by combining the triples in *E* to their corresponding positions, resulting in an utterance tree *t*. An utterance tree can have different branches; however, its depth is limited, indicating that entity names in triples will not iteratively derive branches. The formulation for K-Inject is as follows:

$$t = K\_Inject(s, E) \quad (2)$$

**Knowledge graph:** To generate knowledge, we use the medical knowledge graph released by Liu et al. (2021), which is centered on organs and related disorders. A set of 52.6K triplets (head, re-

lation, tail) containing medical information was retrieved. The head and tail represent entities such as organs or diseases. In contrast, the relation indicates the relationship between entities, such as function and treatment. In this study, we employed the English language vocabulary, which has 2,603 triples in total.

### 2.2 Embedding layer

The embedding layer aims to transform the utterance tree into embedded representations that can be forwarded to the transformer's encoder and then decoder to generate the dialogue. Our embedding layer consists of token, position, and segment embedding layers. However, it differs in that the proposed method's embedding layer receives an utterance tree rather than a token sequence as input. Below, we discuss a method adopted to transform an utterance tree into a sequence that retains its structural information.

**Token embedding**: In our study, the token embedding, including the vocabulary used, is consistent with the original transformer-based LM (see section 3.3). Each token in the expression tree is transformed into a *H* dimensional embedding vector by a trainable lookup table. Token embeddings made using the proposed method differ from those made using the original LMs. The utterance tree tokens must first be rearranged before embedding can occur. After incorporating tokens in the branch, we reverse the order of the tokens in the following nodes. Even though this process is simple, it makes the utterance hard to read and loses important structural information that can be solved using soft-position.

**Soft-position embedding**: Without position embedding, encoders within a transformer will behave similarly to a bag-of-words (BoWs) method, leading to a loss of structural information (i.e., the order of tokens). The position embedding contains all of the structural information in the encoder's input sentence, allowing us to reconstruct the unreadable rearranged utterance. As an alternative to using the transformer encoder's self-attention score for words that appear to be connected but are not, we used masked self-attention (see section 2.3).

**Segment embedding**: Like the transformer encoder, the proposed method uses segmentation embedding to detect utterances when multiple utterances are included. For instance, when two utterances are fed, `[SEP]` is used to incorporate them. A sequence of segment tags is used to denote the combined utterance.

### 2.3 Transformer Encoder with Masked-Self Attention

We present a mask-self-attention to avoid false semantic changes, which is a self-attention extension. Mask-self-attention is defined as follow:

$$Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q, h^i W_k, h^i W_v \quad (3)$$

$$S^{i+1} = softmax(\frac{Q^{i+1} K^{i+1}}{\sqrt{d_k}}) \quad (4)$$

$$h^{i+1} = S^{i+1} V^{i+1} \quad (5)$$

where $Wq$, $Wk$, and $Wv$ are model parameters that can be trained. The hidden state of the $i-th$ mask-self-attention blocks is $hi$. The scaling factor is $dk$. This process improves the representation but does not affect the original utterance's meaning.

### 2.4 Transformer Decoder

The knowledge enriched representation from the transformer encoder is fed to the decoder of an original LM to generate a response. The working process of the decoder layers is similar to that of the vanilla transformer decoder layers.

## 3 Experiments

### 3.1 Datasets

In this study, we used the English version of MedDialog (Zeng et al., 2020) dataset. Table 1 presents statistics of the MedDialog dataset.

Table 1: Dataset Statistics

| Dataset | MedDialog-EN |
|---|---|
| # dialogues | 257,332 |
| # utterances | 514,664 |
| # tokens | 44,527,872 |
| #diseases | 172 |
| Avg. # of utterances | 2 |
| Max # of utterances | 2 |
| Min # of utterances | 2 |
| Avg. # of tokens | 87 |
| Max # of tokens | 3,672 |
| Min # tokens | 1 |

### 3.2 Experimental Settings

We used five different LMs, and all configuration and pre-training settings are consistent with the original LMs used (see section3.3). Adam (Kingma and Ba, 2014) optimizer is used to train our model at 1e-6 initial learning rate. We used a batch size

Table 2: Results: Automatic (BLEU$_2$, $BLEU_4$, $METEOR$, $NIST-4$) and Human (5-point scale) evaluation

| | Automated Evaluation for MedDialog-EN | | | | Human Evaluation |
|---|---|---|---|---|---|
| Model | BLEU$_2$ | BLEU$_4$ | METEOR | NIST-4 | Avg. Score |
| BERT-GPT (Wu et al., 2020) | 5.72 | 4.82 | 0.28 | 0.42 | 3.70 |
| BERT-GPT+Knowledge (Ours) | 9.38 | 6.07 | 17.62 | 0.61 | 4.00 |
| Performance Increase | 3.66↑ | 1.25↑ | 17.34↑ | 0.19↑ | 0.30↑ |
| Transformer (Vaswani et al., 2017) | 2.13 | 2.28 | 11.57 | 0.03 | 2.70 |
| Transformer+Knowledge (Ours) | 2.48 | 2.46 | 12.32 | 0.31 | 3.00 |
| Performance Increase | 0.35↑ | 0.18↑ | 0.75↑ | 0.28↑ | 0.30↑ |
| mT5 (Xue et al., 2020) | 2.59 | 0.84 | 0.20 | 0.41 | 2.70 |
| mT5+Knowledge (Ours) | 7.32 | 3.63 | 1.11 | 0.94 | 3.00 |
| Performance Increase | 4.73↑ | 2.79↑ | 0.91↑ | 0.53↑ | 0.80↑ |
| BART (Lewis et al., 2020) | 15.92 | 9.72 | 0.70 | 2.03 | 3.90 |
| BART+Knowledge (Ours) | 17.25 | 11.07 | 1.73 | 2.07 | 4.15 |
| Performance Increase | 1.33↑ | 1.35↑ | 1.03↑ | 0.04↑ | 0.250↑ |
| T5 (Raffel et al., 2019) | 7.05 | 1.79 | 0.95 | 1.05 | 3.50 |
| T5+Knowledge (Ours) | 15.20 | 8.96 | 1.73 | 1.78 | 4.00 |
| Performance Increase | 8.15↑ | 7.17↑ | 0.78↑ | 0.73↑ | 0.50↑ |

of 64 for 50 epochs. We used grid-search optimization to derive the optimal parameters. We divided all datasets into training, validation, and test sets, with an 80:10:10 ratio for all experiments. The number of heads in multi-head attention is set to 12. The trained models were evaluated using automatic metrics such as NIST-4 (Doddington, 2002), $BLEU_2$, $BLEU_4$ (Papineni et al., 2002), and METEOR (Lavie and Agarwal, 2007).

### 3.3 Baselines

We compared our results with state-of-the-art LMs that are used in previous studies for medical dialogue generation tasks. To be precise, we used BERT-GPT (Wu et al., 2020), Transformer (Vaswani et al., 2017), mT5 (Xue et al., 2020), BART (Lewis et al., 2020), and T5 (Raffel et al., 2019) to compare the performance.

### 3.4 Results

**Automated Evaluation:** Table 2 demonstrates the automatic evaluation results achieved by different LMs, with and without knowledge. The results show that adding medical knowledge to LMs improves the performance across all evaluation metrics. For the MedDialogue-EN, we observed an increase in $BLEU_2$ score ranging from 0.35% to 8.15%, for $BLEU_4$, the improvement range is 0.18% to 7.17%, For METEOR, the increase is from 0.91% to 17.34%, and finally, for NIST-4, the increase in performance is in the range of 0.04% to 0.73%. From the results in Table 2, we can conclude that adding medical knowledge to LMs is beneficial and increases the performance of medical dialogue generation tasks.

**Human Evaluation:** We randomly selected 100 dialog examples for human evaluation. Five medical doctors were asked to rate the generated responses independently on a scale of 1 to 5. The greater the score, the better. The final results are obtained by averaging the ratings provided by various experts. From the human evaluation scores (right column) in Table 2, we deduce that incorporating medical knowledge into LMs generates a more accurate, clinically informative, and human-like response.

## 4 Conclusion

We present a method for enabling LMs with KGs to achieve domain knowledge like doctors. The proposed method transforms an utterance into a knowledge-enriched utterance tree by injecting medical knowledge from KG. The embedding layer converts the utterance tree into an embedding fed to the masked self-attention of a transformer, followed by the decoder to generate the response using medical dialogue history. Experimental results demonstrated that our method outperforms state-of-the-art LMs trained on general data. Further, through human evaluation, we conclude that generated responses are informative and doctor-like. In future, we aim to expand this work to other tasks and datasets.

## References


Olivier Bodenreider. 2008. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, 17(01):67–79.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. HowNet and its computation of meaning. In *Coling 2010: Demonstrations*, pages 53–56, Beijing, China. Coling 2010 Organizing Committee.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. 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)*, pages 5033–5042.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8838–8845.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Qingyang Wu, Lei Li, Hao Zhou, Ying Zeng, and Zhou Yu. 2020. Importance-aware learning for neural headline editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9282–9289.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, et al. 2020. On the generation of medical dialogues for covid-19. *arXiv preprint arXiv:2005.05442*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.