Automated Medical Assistance: Attention Based Consultation System

Ambesh Shekhar
Birla Institute of Technology, Mesra
ambesh.sinha@gmail.com

Raj Ratn Pranesh
Birla Institute of Technology, Mesra
raj.ratn18@gmail.com

Sumit Kumar
Birla Institute of Technology, Mesra
sumit.atlancey@gmail.com

Abstract

With so many possibilities of disease and disorders, there had been a surge in demand for medical assistance. Although there are a number of well-practised physicians and doctors, the inability to reach has always been a concern to many diseased ones. To fill this gap, we propose a model based on deep learning methods, a conversational dialogue system that is able to provide better medication during the need of such and is able to answer any queries related to one's health. We utilized the human-generated medical assistance dataset collected from online platform containing professional assistance via a conversation system. We designed three transformers based encoder-decoder model, namely, BERT, GPT2, and BART and trained them on large the dialogue dataset for text generation. We performed a comparative study of the models and in our analysis, we found that the BART model generates a doctor-like response and contains clinically informative data. The overall generated results were very promising and show that through transfer learning pre-trained transformers are reliable for developing automated medical assistance system and doctor-like-treatments.

1 Introduction

For the past few years, advancements in the medical entity have been applied in the field of social media data and scientific literature. One of those entities is telemedicine, which refers to an operation of conveying patient care in remote ways. As the name suggests, the provider and patient are not physically present with each other during the consultation. The latest technologies and tools have made telemedicine possible and feasible for a majority of the population around the world. So, doctors can provide medical consultations to patients using HIPAA compliant video-conferencing tools such as Thera-Link, Zoom for Health Care, e.t.c.

Therefore, when telemedicine is compared with the orthodox face-to-face medicine practised physically in health centres and clinics, it complements the conventional methods of consultations in several ways. So, it has numerous advantages. First, it surges the access to care. For instance, people living in medically deprived sections of society (e.g., rural areas) that have a lack of proper facilities and specialists, telemedicine allows them to secure rapid and cheaper care contrast with long-distance travelling to see a specialist. The second one is, it decreases net treatment costs. Research by Jefferson health [1] showed that we can save more than 1500 per visits with telemedicine when distracted from visiting emergency wards. Third, telemedicine can raise the standards of care. The author [Pande et al., 2015] presented that telemedicine patients care less for anxiety, depression, and stress, also 38% decreased hospital admissions. Several other merits consist of improving patient engagement and

[1]https://www.healthleadersmedia.com/welcome-ad?toURL=/clinical-care/cost-savings-telemedicine-estimated-19-120-patient-visit

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

satisfaction, improved satisfaction of provider, etc. [Wootton et al., 2017] will provide an in-depth review of telemedicine.

Being optimistic, it has multiple restrictions. For instance, it adds a further burden to consultants. Additionally, physicians get highly occupied by both remote consultants and face to face meetings with patients, finally, it erupts the risk of burnouts. Also, remote patients are difficult to track and monitor, unlike hospital patients whose medical records can be easily examined through clinicians. Therefore, studies and research have been made in the field of artificial intelligence (AI) in order to cope up with the difficulties in implementing telemedicine effectively and efficiently. Generally, Natural Language Processing (NLP) based chatbots are made to develop to a server as a "virtual doctor". In short, "virtual doctor" interacts with the user via dialogues and provides clinical advice based on the medical conditions and history of the patients. Also, "virtual doctor" takes charge and does a sequence check to ask the progress of patients.

In order to make such dialogue systems, a vast dataset is required focused on conversations between doctors and patients in the form of training data. But, privacy concerns are one of the prime hindrances in obtaining the fruitful dataset for further experiments. However, existing datasets are restricted in size and datasets are particularly biased to certain diseases, which would not be enough to train dialogue-based systems that can attain intelligence near to medical specialists and consultants.

In this paper, we utilized a publicly available large-scale medical dialogue dataset MedDialog-EN [Chen et al., 2020] for designing and developing an automatic medical diagnosis system. We used various heavily pre-trained language models such as BERT [Devlin et al., 2018], GPT2 [Radford et al., 2019] and BART [Lewis et al., 2019] to build encoder-decoder models for the dialogue generation task. These models were fine-tuned on the MedDialog-EN dataset in a way such that, given a patient medical issue the models generate a possible diagnosis of the underlying medical problem. We also performed the performance analysis of various dialogue generating models using automatic evaluation metrics. The responses generated by the dialogue generation models were very promising and demonstrated the potential and application of natural language processing in the field of automated healthcare systems.

2 Dataset

In this section, we discuss the dataset we have used to fine-tune our transformer models. The MedDialog-EN dataset [Chen et al., 2020] consists of 257,454 English consultations between patients and doctors. The dataset contains multi-turn dialogues with 257,454 utterances from doctors and 257,454 utterances from patients. The dataset has been scraped from online medical consultation systems like icliniq[2] and Healthcaremagic.com[3] which is managed by professional doctors. The scraped dataset for each patient consultation consists of 2 parts: The description part where the patient describes the physical condition of the body and other Dialogue parts which consists of patient and doctor conversation with an informative response regarding the disease and its prevention. The dataset comprises 51 categories of health disorders like elderly-problems, pain management, diabetes, etc. and 96 specialities including andrology, cardiology, nephrology, pharmacology, etc. These consultations consist of data from 2008 to 2020.

3 Transformer based Dialogue Systems

In this section, we describe various language models which are capable of performing multi-turn dialogue generation with an informative response.

3.1 Bidirectional Encoder Representations from Transformers

BERT [Devlin et al., 2018] has been the revolution in the field of natural language processing since the research on Attention is all you need [Vaswani et al., 2017]. The model has solved numerous problems in natural language processing like classification, sentiment analysis, and question answering tasks using its Bi-directional training on contextual data. Since BERT [Devlin et al., 2018] is eligible

[2]https://www.icliniq.com/
[3]https://www.healthcaremagic.com/

in achieving great results in tasks like token mask modelling, with the help of the Bi-directional encoding method, we imply using this technique for conditional text generation. With the help of a self-attention mechanism, BERT is able to predict tokens depending on the likelihood of previous and upcoming tokens. We utilized **BERT-large-uncased** model with 24-layer, 1024-hidden, 16-heads and 340M parameters, available at Hugging face[4] for developing encoder-decoder model.

3.2 Generative Pretrained Transformer-2

GPT2 [Radford et al., 2019] is an autoregressive pretrained generative model, developed by OpenAI. The model is based on the transformer's decoder architecture. The model has been trained on a 40GB dataset called WebText, scraped from the internet by the researchers, and aims on estimating the next word in a sentence by calculating the probability for the next word depending on the tokens on the left. Since this model is built on the decoder block of the transformer, it uses the principle of mask attention for training the data for a specific task. For developing encoder-decoder model, we used **GPT2-medium** model available at Hugging[5]. The GPT2-medium model has 24-layer, 1024-hidden, 16-heads and 345M parameters.

3.3 BART-autoencoder

BART [Lewis et al., 2019] is a denoising autoencoder that tries to rebuild a corrupted document by performing masked token prediction with the help of bidirectional encoding methods and generates text regressively for natural language generation tasks using a masked attention mechanism. The mask attention mechanism enables the BART model to train on sequence from left to right, generating texts based on the left part of the sequence.

For this transformer-based dialogue system, we create a BART language model wrapper which includes the API of BART-large[6] model from hugging face-transformers. This pretrained model has 400M trainable parameters with 12 encoding and decoding layers in each block, 16 attention heads both at the encoding and decoding layer.

4 Experiment

In this section, we have elaborated on the data preparation and transformer models fine-tuning set-up and process in detail.

4.1 Dataset Preparation

We load the MedDialog-EN [Chen et al., 2020] dataset files on the local system. The files are in textfile format including *Ids*, *Description* and *Dialogues* for each consultation. After going through the dataset we considered the description section consists of the sequence which is a pretext for the query, therefore we consider taking the description text as an utterance by the patient. Therefore our first task is to convert these files to a more readable and easy to compute format. We consider storing each conversation in a list containing sub-lists which included the consultation id, number of turns, and text sequences for patient and doctor in dialogue format fashion. We save this format file in the JSON file. Next, we split the dataset to a ratio of 80:10:10 as training, validation, and test set respectively. We load these sets in our code with the help of de-serialization process during training and evaluation of the model. To mark the beginning of statement(BOS) and end of statement(EOS) we use special token. The formatted input sequence with special token is shown in the table 3. We have `[CLS]` and `[SEP]` special tokens to mark BOS and EOS in BERT [Devlin et al., 2018], similarly for GPT2 [Radford et al., 2019] we use `<|startoftext|>` and `<|endoftext|>` as BOS and EOS, and finally for BART [Lewis et al., 2019] we use `<s>` and `</s>` as BOS and EOS respectively.

[4] https://huggingface.co/bert-large-uncased

[5] https://huggingface.co/gpt2

[6] https://huggingface.co/facebook/bart-large

4.2 Fine-tuning on the dataset

To fine-tune our language models on the MedDialog-ENChen et al. [2020] dataset, we follow the training procedure specified in respective papersDevlin et al. [2018]Radford et al. [2019]Lewis et al. [2019]. We fine-tuned each model for 10 epochs with a batch size of 64. We use the AdamKingma and Ba [2014] optimizer with the linear warm-up scheduler and an initial learning rate of 4e-5. We supplied the encoder with input tokens of length 400 with its respective attention mask tokens. Similarly, we tokenized the decoder input with the token length of 100 and feed decoder with its respective attention mask. In all of the models, the encoder and decoder architect were jointly trained. During the training, we calculated the cross-entropy loss with label smoothing(factor = 0.1). Based on validation score we save the model weights and use it on the test dataset evaluation

5 Results

In this section, we have presented a comparative evaluation of various language model for the medical dialogue generation task using automatic evaluation metrics, namely, Entropy Xia et al. [2020], DIST Li et al. [2015], BLEU Papineni et al. [2002],, METEOR Lavie and Agarwal [2007], NIST Doddington [2002] and perplexity score. In our study, we found that the pre-trained networks are capable of generating very promising medical intervention dialogues. Table 4 shows the response generated, in which BART's [Lewis et al., 2019] generated text contains more informative response compared to other trained models, whereas the GPT2 [Radford et al., 2019] model was able to generate semantically correct response text and the response is more aligned with the original dialogue. BERT [Devlin et al., 2018] was able to generate texts but there was not much of an informative response, the predicted words were poorly arranged and the generated text didn't hold much relevant information.

Table(1) shows that the BERT [Devlin et al., 2018] model's performed average as this model majorly aims at representing texts rather than the conditional generation tasks, GPT2 [Radford et al., 2019] performed great on text generating tasks, this is maybe due to the training on huge MNLI dataset [Williams et al., 2017], whereas the BART [Lewis et al., 2019] model performed outperformed other models as it has been trained on a much larger corpus of data, therefore the BART has more knowledge about informative responses and their semantics. The biggest advantage of BART was that it was trained on much bigger and diverse data in a way to reconstruct the texts from the corrupted documents, which therefore enhanced and increased BART capabilities as compared to other models. Since GPT2 [Radford et al., 2019] was trained on the MNLi dataset [Williams et al., 2017], therefore, it has greater NIST, DIST, BLEU, and METEOR scores compared to other models. This makes the GPT2 model generate favourable n-grams of text for a response. As for the Entropy and DIST metric, GPT2 and BART model reported almost comparable scores. BART achieved the lowest perplexity score of 24.14, whereas GPT2 and BERT achieved a score of 53.52 and 74.68 respectively. Since BART contains a relatively higher number of model parameters, therefore, this provided an additional advantage for BART and hence, improved it's learning capacity.

6 Conclusion

In this task, we adopted the application of transfer learning by using pre-trained transformers BERT, GPT2, and BART. With great performance on various language modelling tasks and easy fine-tuning methods, we develop a dialogue system using these pre-trained transformer models, which could be reliable for medical assistance and treatments. We fine-tune these transformer models layer on a large dataset of interactions between patients and doctors and consultation for different kinds of disorders. The result shows that that via transfer learning transformer models were able to generate clinically informative textual data with meaningful semantics. We believe that there exists a large room for improvement in the proposed models, for example, the training dataset is still limited and cover only small area of medical diagnosis, at the same time with upcoming state of the art models, this paper will be able to improve both on generating tasks and yielding great results. In future, we aim at developing and deploying efficient and reliable mHealth applications in order to provide remote health-related facilities to a larger population group.

References

S. Chen, Z. Ju, X. Dong, H. Fang, S. Wang, Y. Yang, J. Zeng, R. Zhang, R. Zhang, M. Zhou, P. Zhu, and P. Xie. Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*, 2020.

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.

G. Doddington. 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, 2002.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A. Lavie and A. Agarwal. 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, 2007.

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. *arXiv preprint arXiv:1910.13461*, 2019.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

R. L. Pande, M. Morris, A. Peters, C. M. Spettell, R. Feifer, and W. Gillis. Leveraging remote behavioral health interventions to improve medical outcomes and reduce costs. *Am J Manag Care*, 21(2):e000–e000, 2015.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, 2002.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

R. Wootton, J. Craig, and V. Patterson. *Introduction to telemedicine*. CRC Press, 2017.

Y. Xia, J. Zhou, Z. Shi, C. Lu, and H. Huang. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1062–1069, 2020.

7 Appendix

7.1 Manual Human Evaluation

Liu et al. [2016] point out the unreliability in the automatic evaluation of the generated response. Hence, we also perform the human evaluation of these responses. We employed five undergraduate (in final year) medical students and four computer science students (two graduate and two undergraduate) for the task of manually reviewing and rating the automatic generated responses, from 1 to 5; 5

Metrics	BERT-base	GPT2-small	BART-large
Perplexity	74.68	53.52	**24.14**
NIST4	1.06	**1.19**	1.18
BLEU2(in %)	4.08	**7.48**	6.41
BLEU4(in %)	0.13	**0.65**	0.38
METEOR(in %)	15.61	**16.46**	16.26
Entropy4(in %)	11.07	**9.50**	9.16
Dist1(in %)	9.81	11.18	**19.51**
Dist2(in %)	34.76	25.60	**48.88**

Table 1: Evaluation Scores and generated text's average length

being the best. They were to rate each of the two aspects: (1) Relevancy: based on how relevant the response was to the conversation history; (2) Human-like: how far close the response sounded like a real human. The responses were de-identified to keep the response generation method anonymous to the annotators. Thus, the ground-truth replies from the doctor were also given ratings (in an anonymous way). The ratings from different annotators were finally averaged.

	BERT	GPT2	BART	Groundtruth
Relevancy	2.34	2.93	3.26	3.80
Human-like	2.18	2.82	3.12	3.65

Table 2: Human evaluation results of various models

Table 2 shows human evaluation results. From this table, we can observe the following. Firstly, pre-trained models BART and GPT2 perform better than the Transformers. These further illustrate the effectiveness of large and diverse pretraining. Secondly, although GPT2 achieves better scores on machine translation metrics, in manual human screening, we found that the BART model performed better than GPT2. It mirrors the issues raised in Liu et al. [2016], pointing out the fact that machine translation metrics are not appropriate for evaluating dialogue generation. Third, BART achieves a human-like score that is very close to the ground-truth. It indicates that the auto-generated responses have high language quality. The relevancy rating of BART which is higher than 3 indicates a good level of relevance between the generated responses and conversation histories.

Models	Sequence
BERT	[CLS] This is input for BERT. [SEP]
GPT2	$<$ |startoftext| $>$ This is input for GPT2 $<$ |endoftext| $>$
BART	<s>This is input for BART </s>

Table 3: Input Sequence Format

Models	Text
Patient	What causes spotting of blood during urination? Hello i am a teenager male and i'm worried when i urinate i have no problems but me and my dad go out on his boat alot so i have to urinate in a bottle whenever i urinate in a bottle a tiny bit off blood comes out when i'm finished urinating and it always bleed for like a week after not constant just a drop or two please help me.
Doctor	Hello, There are many reasons for hematuria at your age. But i don't think You are having more serious problem since you don't have problem in urinating. Inflammation of the kidney, urethra, bladder or Inherited diseases, like poly-cystic kidney disease causes blood in urine. I suggest you to consult a urologist to rule out above mentioned conditions. Take care. Hope I have answered your question. Let me know if I can assist you further.
BERT-output	Welcome to healthcaremagic. Please consult your doctor for further evaluation. You may require antibiotics after urination. It may be due to prostate inflammation without consulting a doctor. Kindly let me know if you are you not taking any antibiotics as they will be alright. take tablet after one month. Still have a query. Thanks.
GPT2-output	Hi, You should be investigated for urinary tract infection and blood in urine to rule out urinary tract infection. There are many reasons for the blood during urination. It can be due to urinary tract infection, urethritis, prostate inflammation, urinary tract infection or urinary stones. Hope I have answered your query. Let me know if I can assist you further. Dear.
BART-output	Hi, You may have hematuria. It will be on the grounds that aggravation of kidney, acquired diseases, urethritis kidney ailment to cause blood in pee. Take a stab at counseling urologist and specialist. Expectation I have responded to your inquiry. Tell me I can help further.

Table 4: Generated responses from various models on a test dialogue