

COVID Doctor Chatbot with Data Augmentation

Chenhao Wang, Yuxin Wang, Philip Cho

Abstract

We present a Seq2Seq based conversational agent (chatbot) that can answer queries from patients regarding topics related to COVID-19. The model uses patients' COVID related questions and previous replies of the certified doctors scraped from the internet as input. The model is trained with the attention based Transformer (Vaswani et al., 2017) architectures like GPT (Radford and Sutskever, 2018; Radford et al., 2019) and BART (Lewis et al., 2019), to ensure it pays attention to important part of the input sequence when decoding. Furthermore, we present data augmentation approach to efficiently counter the problems coming from shortage of data.

1 Introduction

As of January in 2020, the COVID-19 has killed over 1.56 millions people out of 68.1 millions confirmed cases around the world. An intelligent chatbot could serve the public to alleviate some of the problems coming from the extreme imbalance of available doctors compared to the number of patients and people who need more information regarding the pandemic. In this project, we present a Seq2Seq based conversational agent that can carry on conversation with the patients regarding topics related to COVID-19. In order to get the best results, we implement multiple techniques and evaluate the results with quantitative metrics and human evaluations. Our experiments including implementations of Transformer (Vaswani et al., 2017), GPT (Radford et al., 2019), BART (Lewis et al., 2019) and data augmentation. Finally, our chatbot is not only capable of understanding the intention behind questions asked by people, but also capable of carrying out a conversation that sounds as natural, and as doctor-like as possible, while correctly providing the answers to the questions asked by the patients.

2 Prior Work

In Natural Language Processing (NLP), detecting important relationship between two sequences of text and mapping the transition from one sequence to another, is referred to as sequence to sequence (Seq2Seq) problems (Yu, 2018). In the past decade, rapid development of Deep Neural Network has brought significant improvement to Seq2Seq tasks like Neural Machine Translation (NMT), but for many years, most approaches relied heavily on Recurrent Neural Network architectures that used LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014). Even the widely used applications like Siri, Cortana, Alexa, and Google voice assistant all used RNN as the backbone of their models. The problem with such approaches were that it applied neural networks as a component in a traditional statistical machine translation system, and retained the log-linear combination and only exchanged parts in the traditional architecture (Stahlberg, 2020). However, a major change in perspective took place in the NLP community starting 2015, when Bahdanau et al. (Bahdanau et al., 2016) first introduced a single large neural net that directly transforms the source sentence into the target sentence, using encoder decoder architecture augmented with attention mechanism. This has brought explosion of scientific publications related to NMT as well, and inspired some of the most recent state-of-the-art NMT models that will be discussed further in detail in the next section.

3 Approach

A traditional chatbot generates responses based on handcrafted rules laid out by human annotators, and it often stops functioning when it encounters questions outside of its expectations. Modern artificial conversational agent, on the other hand, is capable of continuing extremely human like re-

sponses intelligently based on NMT mechanism. But earlier chatbots that relied on Recurrent Neural Network (RNN) had limited success because the decoder was unable to correctly determine which input sequence to pay attention when it is generating a response. With the introduction of the attention mechanism (Bahdanau et al., 2016), some of the modern, highly effective dialogue generation models like Transformer (Vaswani et al., 2017), GPT (Radford and Sutskever, 2018), and BART (Lewis et al., 2019) were invented. In this section we will introduce some of these recent models that we have used for generating COVID-19 dialogues and compare the results using the metrics, such as perplexity, BLEU, NIST-4, etc. Furthermore, data augmentation will be presented as a way to counter problems coming from scarcity of data. Subsections down below will elaborate further in detail regarding each of the model's architecture and their implementation.

3.1 Transformer

Built on top of the first attention based neural machine translation model proposed by Bahdanau et al in 2015 (Bahdanau et al., 2016), Transformer (Vaswani et al., 2017) is one of the most well known encoder-decoder architectures for Seq2Seq (Sutskever et al., 2014) modeling, which has brought significant change to Deep Learning community overall. Like recurrent neural network (RNNs), Transformers are designed to handle sequential data, but Transformers do not require data be processed in order. Due to this aspect, the Transformer allows for much more parallelization than vanilla RNNs, and therefore reduces training times significantly. The Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. The idea behind Transformer is to handle the dependencies between input and output with attention and recurrence completely (attention is all you need).

The Figure 2 shows the architecture of Transformer model. In the Transformer's encoder section, each encoder layer processes its input to generate encodings, which contains information about the parts of the inputs that are relevant to each other. It passes its set of encodings to the next encoder layer as inputs. Each decoder consists of three major components: a self-attention mechanism,

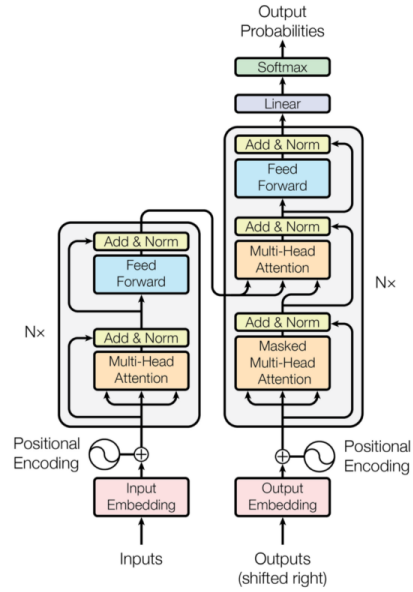


Figure 1: The Transformer - model architecture (Vaswani et al., 2017)

nism, an attention mechanism over the encodings, and a feed-forward neural network. The decoder functions in a similar fashion to the encoder, but an additional attention mechanism is inserted which instead draws relevant information from the encodings generated by the encoders.

3.2 GPT

Building on Transformer (Vaswani et al., 2017), the GPT (Radford and Sutskever, 2018) model is a language model that defines a marginal probability on a single sequence. In GPT, $p(x_i|x_1, \dots, x_{i-1})$ is defined using the Transformer decoder, which uses a stack of self-attention and feed-forward layers followed by layer normalization to encode x_1, \dots, x_{i-1} , then predicts x_i from these encodings. Extending from GPT, GPT-2 (Radford et al., 2019) proposed to move the layer normalization to the input of each sub-block and adding an additional layer normalization after the final self-attention block.

In this project, we want to study the effectiveness of GPT-2 model on dialogue generation. One popular GPT-2 model as a dialogue system is DialoGPT (Zhang et al., 2019), which is pretrained on 147M English Reddit dialogues extracting from comment chains on the platform over 10 years span. With the ground-truth pair of source sentence (dialogue history $S = x_1, x_2, \dots, x_m$) and target sentence (ground truth response $T = x_{m+1}, \dots, x_N$), DialoGPT works by maximize the probability

$$P(T|S) = \prod_{n=m+1}^N p(x_n|x_1, \dots, x_{n-1}).$$

3.3 BART

BART (Lewis et al., 2019) is one of the latest Seq2Seq pretraining model released by Facebook in 2019, which combines the idea of the popular BERT (Devlin et al., 2019) bidirectional Transformer model as well as GPT (Radford and Sutskever, 2018). In the pretraining process, the input text is intentionally corrupted with arbitrary noising function, and the Seq2Seq model learns to recover the original text by minimizing the negative log likelihood of the original document. The intentional noising function here is operations like randomly replacing tokens with [mask] element, randomly deleting tokens in the sequence, and randomly shuffling sequences of the tokens. The corrupted document becomes encoded with bidirectional encoder similar to BERT (Devlin et al., 2019), and the likelihood of the original document is calculated with an auto-regressive decoder seen in GPT models (Radford and Sutskever, 2018). In terms of fine-tuning with the COVID-19 dialogue, uncorrupted document is used as input to both encoder and decoder, using the representation from the final hidden state of the decoder (Lewis et al., 2019).

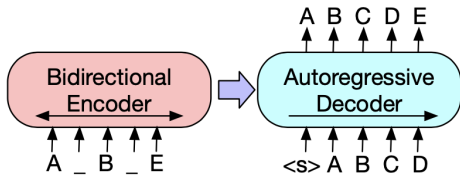


Figure 2: The basic idea architecture of BART(Lewis et al., 2019)

The key advantage to training with a noise is that it gives greater flexibility in terms of using the pretrained model to any sort of NLP tasks, not only NMT, but other tasks like span prediction, text generation, and etc.

3.4 Augmentation

Very few research papers address the need of data augmentation. We believe this is because text data is one of the most easily attainable source of data, but more specifically, very few models train with a specific purpose as that is more of a territory of Question and Answering. But Question and Answering models usually do not respond like humans, as focus is usually to provide accurate re-

sponse. Thus we are dealing with a special scenario in which we specifically need certified doctor's response (or at least COVID-19 related response) for fine-tuning, and finding the right data is inevitably difficult. Augmenting natural language, however, is challenging, as texts are discrete and therefore cannot be augmented by common image transformation approaches like random flip, crop, saturation, and etc.

Data augmentation is not the most common practice in Seq2Seq tasks(Xie et al., 2020), but studies such as Easy Data Augmentation (EDA) (Wei and Zou, 2019), demonstrates largely improved accuracy just from simple augmentation strategies like synonym replacement, random insertion, random swap, and random deletion. There are more sophisticated strategies for data augmentation like Unsupervised Data Augmentation (UDA) (Xie et al., 2020), and furthermore augmentation can also be performed in a character level as well as a entire sentence level. But for this research, we have decided to limit the scope of augmentation, and selected two of the most common approaches of augmentation in word level that makes meaningful difference without overly corrupting the data:

1. **Synonym replacement:** Replace randomly chosen n words from a sentence that are not stop words, with Wordnet Synonym
2. **Insertion based on word embedding:** Insert random words to a sentence from Fasttext word embedding space. Small degree of noise can actually improve generalization.

If you consider the context of synonym substitution, the idea is as simple as replacing "Hi doctor" to "Good day, doctor".

4 Experiment

This section will explain the dataset used for the project, the metrics we are using to measure the performance of the experiment, and details of each experiment conducted thus far.

4.1 Dataset

The dataset was scraped from online resources like healthcaremagic.com, covidquesitons.ca, and various other online articles. The information was manually screened and reshaped to match the structure of online chat conversation. For ease of the computation, each dialogue is initiated by the patient only, and it contains the exact same number of

questions and replies. Many dialogues have more than one pair of question and reply. An example of a typical conversation is like the following:

- **Patient:** Hello, I have a question for you. Is it safe to use public transit?
- **Doctor:** If you must use public transit, regularly and thoroughly wash your hands and avoid touching your face with unwashed hands. Continue to distance yourself 2 metres (6 feet) away from others when possible. If this cannot be achieved, you may choose to wear a non-medical mask or cloth mask. Also if you can, try to use public transit at off peak times.
- **Patient:** Okay thanks
- **Doctor:** No problem. Stay safe!

The Table 1 shows the number of dialogues and conversation pairs for our dataset. The dataset is randomly separated into train/validation/test set, with the ratio of 0.8/0.1/0.1

Dataset	Dialogues	Pairs
English	625	712

Table 1: Number of dialogues and pairs in dataset

4.2 Data Augmentation

We applied data augmentation on the source, which is the patient’s query, since applying augmentation on the responses may harm the validity of the doctor’s response. Our goal is to deliver stable replies back upon multiple different variations of the user’s query. The implementation for the augmentation has been sourced from Python library NLPAug (Ma, 2019). The augmentation of the dataset increases number of available pairs to 2127.

4.3 Transformer

We applied experiments on Transformer. We split dataset into a training, a validation, and a test set based on dialogues, with a the same ratio of 8:1:1. We used the Adam optimizer with linear learning rate scheduling, setting the initial learning rate as $4e-5$ and the batch size as 4. The objective is the cross entropy loss with label smoothing where the factor was set to 0.1. We trained Transformer model for 50 epochs and set a checkpoint at the end of each epoch. Finally, we take the model with

the lowest perplexity value on validate data as the final model.

4.4 GPT-2

DialoGPT is a pre-trained GPT-2 model, trained on 147M multi-turn dialogue from Reddit discussion thread. Using the pre-trained model from DialoGPT (Zhang et al., 2019), we fine-tuned the a GPT-2 model on our prepared English dataset for generating English COVID-19 dialogues for 4 epochs.

4.5 BART

The implementation of BART is based on Hugging Face, a popular library for NLP tasks. BART contains 24 layers and 406M parameters, and it’s pretrained on the weights provided by Facebook. Because this is an enormous model, the model was only trained for 5 epochs before running out of available RAM. Max sequence length was set to 50, in order to ensure that the sequence is not unnecessarily long.

5 Results and Analysis

5.1 Evaluation Metrics

In this project, we used both automatic metrics and human perception to evaluate the generated dialogue response.

We utilized perplexity metrics as it captures the degree of uncertainty a model demonstrate in predicting. A lower perplexity usually signifies that the model is able to assign a higher probability to a better results while training, thus the lower perplexity, the better a model we have. Since perplexity relies directly on cross entropy and the underlying probability distribution of the words in the response, we find it useful to evaluate the feasibility of the NLP model in some preliminary experiments.

After we solidified the experiment pipeline, we expand our evaluation metric with BLEU-n (Papineni et al., 2002) score, NIST-n (Cer et al., 2010), and Meteor, as they are common metrics for evaluating machine translation. BLEU, short for “Bilingual evaluation understudy, looks at n-grams overlap between the output and reference translations with a penalty for shorter outputs. NIST is a variant of BLEU, and it weights n-gram matches using information gain to penalize uninformative n-grams. BLEU-n and NIST-n compare the similarity between predicted responses and the ground-truth by

matching n-grams.

5.2 Quantitative Results

Table 2 summarizes the automatic evaluation results achieved by perplexity in validation. From this table, we can see that the Transformer has very high perplexity which is over 300. The high perplexity value means that the output from Transformer is quite different with reference response. For the other models, the GPT-2 and BART both get lower perplexity values after applying the data augmentation. The BART with data augmentation gets the lowest perplexity value while training. The reasons we did not try augmentation with Transformer model: Transformer needs to be trained from scratch since no pre-trained takes place, it needs long time for training and we had limited time. Moreover, the Transformer has poor performance compared to other two models both on quantitative and qualitative evaluations. So we focus more on GPT-2 and BART after the milestone.

Model	perplexity
Transformer	307.83
GPT-2	14.97
GPT-2 + Aug	8.23
BART	16.30
BART + Aug	1.99

Table 2: Perplexities of validate set.

Table 3 shows more metrics results by BLEU-n, NIST-n, and Meteor scores on our COVID-19 test dataset. Both tables are generated on the original dataset. From this table, we make the following observations. First, the GPT-2 and BERT have better performance than Transformer based on those metrics. Second, the BART in general gets better results than GPT-2. Third, considering the data augmentation, the GPT-2 could not get higher scores after applying data augmentation, but BART saw improvement in terms of Nist scores.

Model	BLEU-2	BLEU-4	Nist-2	Nist-2	Meteor
Transformer	0.17	0.07	0.90	0.93	0.23
GPT-2	0.19	0.10	0.85	0.89	0.25
GPT-2 + Aug	0.11	0.05	0.50	0.51	0.13
BART	0.34	0.20	1.94	2.01	0.35
BART + Aug	0.05	0.23	3.46	3.82	0.25

Table 3: Quantitative evaluation of test set.

User: Can I go for Coronavirus testing if my nose is blocked and I have traveled from a high risk country?

Reference: coronavirus testing. yes, get tested if symptomatic (early symptoms could be mild respiratory symptoms and nonspecific) or self quarantine given your high risk of exposure . .

Transformer: Hello the to your symptoms . If you like to video or text chat with me ?

GPT-2: Yes, you should definitely do this. The risk of corona is high. You should definitely take it to a local health department.

BART: Brief opinion: Covid testing You cannot get a test if your nose is blocked. It is not feasible. Would you like to video or text chat with me?

Table 4: Sample outputs from different models in the context of shorter user query

User: Can coronavirus symptoms be mild for some people versus severe ? for example, could it just involve being very fatigued, low grade fever for a few days and not the extreme symptoms? or is it always a full blown cold and struggle to breathe?

Reference: Brief opinion: symptoms vary. Some may have no symptoms at all . Some can be life threatening. Would you like to video or text chat with me?

Transformer: Thanks for your question on healthcare magic. I can understand your query. You can be a chest x ray on the and it is not have to video or text chat with me?

GPT-2: It depends on your severity, and type of cold. If you are having a mild fever and or dry cough, then should be treated. If is not having a cold, then you should be fine. Would you like to video chat with me?

BART: The symptoms of COVID-19 are similar to other respiratory illnesses including Influenza and the common cold. The most common symptoms include: Fever, chills cough or worsening cronic cough. You could also experience shortness of breath, sore

Table 5: Sample outputs from different models in the context of longer user query

5.3 Qualitative Results

While quantitative evaluation is simple to use, we want to use qualitative method to verify the reliability of the generated response. Thus, as part of the planned work, we performed human evaluation on the generated responses. Table 4 and Table 5 shows the sample output from each model with short query and long query. From those sample outputs, we make the following observations. First, pre-trained models GPT-2 and BART in general have better performance than un-pretrained Transformer. This observation demonstrates the effectiveness of transfer learning, which leverages external large-scale data to learn powerful representations of texts. Second, the Table 4 indicates that the response generated by GPT-2 is more relevant, informative, and human-like, compared with those generated by others. The BART's response said the user doesn't need test. While the GPT-2's response suggests the user to get a test for COVID-19 because the user said "I have traveled from a high risk country". This response gives correct and informative medical advice. Third, from the Table 5 we know that the response generated by BART performs better when the query is long than the short one. However, the GPT-2 in general sounds most like humans due to dialogues being used for pretraining.

6 Conclusion

Dataset Volume and Augmentation We have tested augmentation methods on both GPT-2 model (Radford et al., 2019) and BART (Lewis et al., 2019) and discovered that augmentation leads to significant improvement in terms of perplexity. However, this could be due to simple fact that the available data for fine-tuning just became larger as augmentation did not necessarily lead to better performance for GPT-2 model. This is just a hypothesis, but this result could be due to the fact that BART Model is specifically trained on the noisy data, so when the data gets augmented, which is similar to noise being inserted, it reacts much more flexibly than GPT-2. Furthermore, we cannot ignore the fact that only two types of augmentation policy was used for testing. We suspect that the result may be improved upon testing out different augmentation policies. Some of the reinforcement based augmentation policy searcher like AutoAugment (Cubuk et al., 2019) or the previously mentioned UDA (Xie et al., 2020) may be used to yield

better performance, but they are outside the scope of this paper.

Transfer Learning We find that transfer learning could have large impact on small datasets. In transfer learning, pretraining a language model on a very large corpus of text enables the model to generate long stretches of contiguous coherent text. With the pre-trained model, we can focus on fine tune it in adaptation of our task at hand - a Covid chatbot.

In our experiments, we witnessed huge improvement of using pre-trained models (GPT2, BART) to fine-tune. GPT works well on shorter query sequences. BART works better on longer query sequences. Both of them prove to be more effective and efficient (fine-tune for less than 5 epochs) than Transformer, which is time-consuming (train for around 50 epochs) to train and yield less qualified results.

7 Future Work

Human evaluation In this research, we have evaluated our models using objective quantitative metrics. However, there are no standard metrics for Seq2Seq models like sentiment analysis, so we can be using many other metrics for evaluation as well. Furthermore for chatbots, it is very important to be able to measure how natural the conversation is, as the purpose here is not to create simple Question and Answering model, but to create a doctor chatbot that can provide information while sounding like a human. We believe that manual human evaluation can also be helpful to determine if the output is doctor like or not. We have previously discussed possibilities of using external sources like Amazon Mechanical Turk (AMT) as well, but this was not fulfilled due to time and resource constraints. In the future studies, we would like to include human evaluation as well.

Multiple Languages As covid-19 is a global pandemic, similar efforts in different languages, such as Chinese, French, Spanish and so on would be valuable for an international audience.

Experiment XLNet XLNet (Yang et al., 2020) is one of the latest and most promising autoregressive language model. As an extension to this project, we thinking exploring a neural dialogue generator model with a pre-trained XLNet, and fine-tune the model on the English dataset we have obtained would be a meaningful direction.

Experiment MBART In 2020, Facebook research team has announced an improved version of BART (Lewis et al., 2019) called MBART (Liu et al., 2020), which expands the original work to multilingual context by using 24 pairs of publicly available corpora. For future work, we could also explore the option of MBART to create a multilingual chatbot.

8 Contribution

We worked collaboratively on the report and the project notebook.

Chenhao Wang Implemented Transformer training and evaluation.

Yuxin (Azmarie) Wang Implemented GPT2 fine-tuning and evaluation.

Philip Cho Generated dataset, applied augmentation, created BART model.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. [Neural machine translation by jointly learning to align and translate](#).
- Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The best lexical metric for phrase-based statistical mt system optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 555–563, USA. Association for Computational Linguistics.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. [Empirical evaluation of gated recurrent neural networks on sequence modeling](#).
- Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. 2019. [Autoaugment: Learning augmentation policies from data](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#).
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#).
- Edward Ma. 2019. [Nlp augmentation](#). <https://github.com/makcedward/nlpaug>.
- 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 and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *arxiv*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Felix Stahlberg. 2020. [Neural machine translation: A review and survey](#).
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to sequence learning with neural networks](#). *CoRR*, abs/1409.3215.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#).
- Jason Wei and Kai Zou. 2019. [Eda: Easy data augmentation techniques for boosting performance on text classification tasks](#).
- Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. [Unsupervised data augmentation for consistency training](#).
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. [Xlnet: Generalized autoregressive pretraining for language understanding](#).
- Lei Yu. 2018. [Tackling sequence to sequence mapping problems with neural networks](#).
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.