Automatic Generation of Electronic Medical Record Based on GPT2 Model

Junkun Peng*, Pin Ni*†, Jiayi Zhu*, Zhenjin Dai*, Yuming Li*†, Gangmin Li*‡ and Xuming Bai§
*Research Lab for Knowledge and Wisdom, Xi'an Jiaotong-Liverpool University, China
†Department of Computer Science, University of Liverpool, UK
§Department of Interventional Radiology, Second Affiliated Hospital of Soochow University, China
Email: ‡Gangmin.Li@xjtlu.edu.cn, §2005baixuming@163.com

***Abstract*—Writing Electronic Medical Records (EMR) as one of daily major tasks of doctors, consumes a lot of time and effort from doctors. This paper reports our efforts to generate electronic medical records using the language model. Through the training of massive real-world EMR data, the CMedGPT2 model provided by us can achieve the ideal Chinese electronic medical record generation. The experimental results prove that the generated electronic medical record text can be applied to the auxiliary medical record work to reduce the burden on the compose and provide a fast and accurate reference for composing work.**

***Index Terms*—Text Generation, GPT2, EMR**

I. Introduction

The EMR generation is one of the key steps to achieve the automatic recording of the medical record, which can greatly reduce the workload for doctors. This paper has carried out a preliminary test on the road of exploring how to make Chinese electronic medical records. Using the unsupervised characteristics of the GPT2 model, the data set composed of real patient medical records is trained and generated through the model, and the sample of the electronic medical record is obtained. The meaning of the project is about generating the content of the electronic medical record based on a keyword by the GPT2 model. Based on the fact that most similar symptoms are common in the same type of diseases, the EMR generation based on previous records allows the doctor to simply modify the automatic generated patient's medical record for specific situations without the heavy repetitive EMR writing work. In addition, it can be used in a similar way to the Next Word/Sentence Predictor (e.g. Google Smart Compose) in the search engine to assist in generating prompts for inputting specification terms and expressions during the EMR writing process. This conception has never been proposed and implemented before, and we hope to contribute to this aspect exploit through this work. Our model can be found here: https://github.com/knowis-org/CMed-GPT2.git.

II. Literature Review

Traditional supervised learning methods are sensitive to the distribution of data sets, therefore, these well-trained models are more like a versatile genius than a capable generalist [1]. This is not possible for data that is not labeled and whose type is nonuniform. The patient records in our project did not prioritize the sample as the sample set can only be classified according to the similarity between the samples (clustering) to try to minimize the intra-class gap and maximize the gap between the classes. This is the unsupervised learning method described below.

OpenAI-GPT proposes a semi-supervised approach to the task of language understanding using unsupervised pre-training and supervised fine-tuning. The setting of this model does not require the target task and the non-annotated data set to be in the same field. The model has two processes. the first step is learning a deep model using language model, and then the parameters are adjusted to the target task using the corresponding supervisory targets [2]. The GPT model inspired us to use its language model to learn Chinese and then predict and generate sentences. However, due to the limitations of the English version, the GPT model cannot be implemented. BERT as another method of pre-train language representations which uses the Deep Bidirectional Transformers model. Bi-directional prediction is the prediction of the part of the mask from the back to the back and from the back. But the author of BERT believes that bi-directional still cannot fully understand the semantics of the whole statement [3]. In this case, the GPT2 model is a better approach for our project since the key difference between them is that GPT2 could like the traditional language model, outputs a token at a time. Additionally, the GPT2 model currently has a version that can be applied to pre-training of Chinese.

III. Methodology

Our goal is to automatic generate electronic medical records to help reduce the workload for the doctor. The method adopted by the paper is to generate a sample of the electronic medical record by performing a unidirectional transformer training on the data set composed of the existing medical records through the GPT2 model. The core idea of GPT2 is that unsupervised pre-training models can be used to perform supervised tasks. The structure of the GPT2 model is the same as that of GPT, as Fig. 1 shows:

The GPT2 Language model as an autoregressive or autoencoding language training task through a large number of corpora, and judges the joint probability distribution of a sentence (a sentence is split into multiple tokens). The formula for prediction is as follows:

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

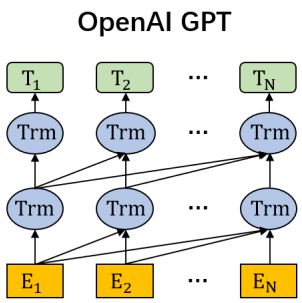

Fig. 1. The structure of GPT2.

$$p(x) = \prod_{i=1}^{n} p(s_n|s_1, ..., s_{n-1}) \quad (1)$$

GPT-2 can handle 1024 tokens. Each token passes through all the decoder structures along its own path. The model has only one input token, which has processed sequentially through all layers so the path will be the only active path. And then a vector is generated along the path. The vector can be scored according to the vocabulary of the model, after that the next word is selected by the score. The role of the Self-Attention layer interprets the context of a word before processing the word and passes it to the neural network layer. GPT2 can use roam to training (generating unconditional samples), or given a hint to generate text about a topic (i.e. generate an interactive conditional sample), this paper uses a hint to give medical records related words to generate interactive samples. The work was trained based on the GPT2-Chinese model [4].

IV. EXPERIMENT

A. Experiment environment

Our experimental environment is as follows: CPU Intel Xeon E5-2678 v3, RAM: Dual 2.50GHz, GPU: Dual Nvidia GeForce GTX 1080 Ti. The specic parameters of the training process of our models are as follows: batch size=16, epoch=50, LSTM units=256, GRU units=256

B. Dataset and parameters description

. The dataset used in this paper consists of two parts, one for the dataset given from 2019 China Conference on Knowledge Graph and Semantic Computing (CCKS 2019) and the other for the real patients' medical records of the Second Afliated Hospital of Soochow University (SAHSU). These data are cleaned and converted to json files for model training. Next, the relevant parameter table for the model operation is given by TABLE I:

V. RESULTS AND ANALYSIS

The first step is to pre-train the data through the GPT2-Chinese train model. After the training, we get the loss curve of the training set through the visualization function of tensorboard, as shown in the Fig. 2:

TABLE I
PARAMETERS DESCRIPTION

Parameter Name	value
initializer_range	0.02
layer_norm_epsilon	1e-0.5
n_ctx	1024
n_embd	768
n_head	12
n_layer	10
n_positions	1024
vocab_size	13317

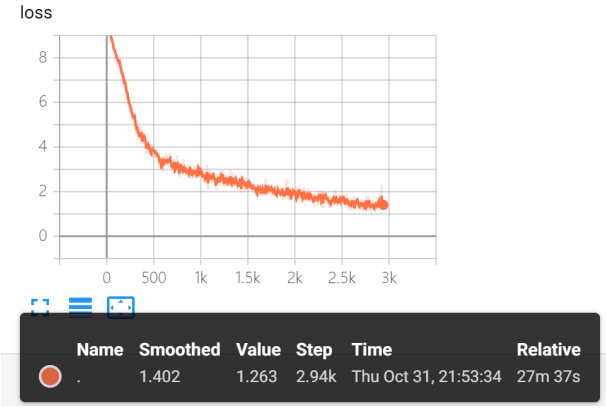

Fig. 2. The loss curve of training

Through the loss curve, it can be found that the loss of the training has become decreased significantly, and finally approximately stabilizes at 1.263. Through the trained data to the given keywords to generate an interactive medical record sample. Table II gives the samples generated after the two types of keywords, respectively. The setting keywords are "Patient" and "Physical Examination".

We have also tried keywords such as "Hepatic cyst", "Cancer", etc., but the generated samples are not shown here since the sample will take up a large section. More samples could be found in the link of GitHub mentioned in the Introduction. After analysis, it can be concluded that the two keywords of the "Patient" and the "Physical examination" can obtain a highly similar patient medical record on account of the keyword frequently appearing in the real medical record, in the meantime, the GPT2-Chinese generation model can generate better models under keyword-based prefix conditions. However, after we perform the [MASK] treatment of the prefix, the current flaw of the model will exposure when the result of the model output is unstable. The reason is that the data set of the training is not enough, and the available medical record data in Chinese is limited.

VI. CONCLUSION

In conclusion, this paper based on the GPT2 model can be used to generate electronic medical records through the training set to generate related samples. The results show that the samples generated under certain keywords have a general

TABLE II
OUTPUT RESULT

Serial number	Output
1	患者为行进一步治疗来我院，门诊拟“左肺结核”收住入院。病程中患者 无畏寒等，无头昏迷及抽搐，无发热，食纳睡眠可，大小便尚可，体重下降。初步诊断：肝转移癌。 肝癌术后1月余，乏力1年，进行性加重1月余。患者2013年5月24日 无明显诱因下出现中上腹胀痛，无恶心呕吐，无腹部胀腹胀，无呕吐，遂至我院行mri检查 Translation：The patient came to our hospital for further treatment, and the outpatient clinic planned to treat left tuberculosis. During the course of the disease, the patient has no chills, no head coma, and convulsions. He also has no fever and has nice eating, sleeping, and urine. Besides, the patient loses weight. Initial diagnosis: liver metastasis. After 1 month of liver cancer surgery, the fatigue was 1 year, and the progressive exacerbation was more than 1 month. On May 24, 2013, there was no obvious incentive for mid-abdominal pain, no nausea, and vomiting, no abdominal distension, no vomiting. Then the patient came to the hospital to have an MRI examination.
2	体检查提示：肝左叶占位，查afp：1111ng/ml，考虑肝癌，后患者为行介入治疗，至我院就诊，门诊拟以“肝恶性肿瘤”收住我科。 入院时患者一般情况尚可，食欲睡眠可，大小便正常，近期体重未见明显变化。初步诊断：肝癌；肝硬化；胆囊结石；脾肿。 肝癌末次介入术后2月余。患者2015年6月因上腹部隐痛不适至苏大学 Translation: Physical examination suggestion: the left lobe of the liver is occupied, check AFP: 1111 ng/ml, consider liver cancer, after the patient for interventional treatment, to our hospital, the clinic is intended to ”hepatic malignant tumor” to receive our department. At the time of admission, the general condition of the patient was acceptable. The appetite was sleepy, the bowel movements were normal, and the weight did not change significantly in the near future. Initial diagnosis: liver cancer; cirrhosis; gallstones; splenomegaly. The last intervention of liver cancer was more than 2 months after surgery. The patient was unwell to the Soochow University in June 2015 due to pain in the upper abdomen.
3	体检查提示：肝左叶占位，查afp：1111ng/ml，考虑肝癌，后患者为行介入治疗，至我院就诊，门诊拟以“肝恶性肿瘤”收住我科。 入院时患者一般情况尚可，食欲睡眠可，大小便正常，近期体重未见明显变化。初步诊断：肝癌；肝硬化；胆囊结石；脾肿。 肝癌末次介入术后2月余。患者2015年6月因上腹部隐痛不适至苏州大学 Translation: Physical examination suggestion: the left lobe of the liver is occupied, check AFP: 1111 ng/ml, consider liver cancer, after the patient for interventional treatment, to our hospital, the clinic is intended to ” hepatic malignant tumor” to receive our department. At the time of admission, the general condition of the patient was acceptable. The appetite was sleepy, the bowel movements were normal, and the weight did not change signicantly in the near future. Initial diagnosis: liver cancer; cirrhosis; gallstones; splenomegaly. The last intervention of liver cancer was more than 2 months after surgery. The patient was unwell to Soochow University in June 2015 due to pain in the upper abdomen.
4	体检查提示肝转移癌，2019-402-03-05于我科行肝肿瘤消融术，术后好转出院。今患者为行进一步介入治疗入， 门诊拟“肝转移癌”收入我科，病程中患者饮食、睡眠可，大小便尚可，体重无明显变化。 初步诊断：肝转移癌；结肠癌术后；直肠癌术后。右乳腺癌术后一年余。患者2015-01-06无明显诱因出现右下肢 Translation: Physical examination showed liver metastasis, 2019-402-03-05 in our department of liver tumor ablation, postoperatively improved and discharged. Today’s patients are further involved in the treatment, the outpatients intend to ” liver metastasis cancer” income in our department, the patient’s diet, sleep the bowel movements are acceptable, bodyweight has no signicant changes. Preliminary diagnosis: liver metastases; postoperative colon cancer; postoperative rectal cancer. More than one year after surgery for right breast cancer. In 2015-01-06 the patient had no obvious cause of right lower extremity.

format and content of formal medical records.
In the future works, we will continue to follow up on the development of this project, and improve the diversity and rationality of the generated electronic medical records.

VII. ACKNOWLEDGEMENT

This work is partially supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02) and KSF-A-17. And it is also partially supported by Suzhou Science and Technology Programme Key Industrial Technology Innovation programme with project code SYG201840. We appreciate their support and guidance.

REFERENCES

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[2] M. Riebe, H. Häffner, C. Roos, W. Hänsel, J. Benhelm, G. Lancaster, T. Körber, C. Becher, F. Schmidt-Kaler, D. James *et al.*, “Deterministic quantum teleportation with atoms,” *Nature*, vol. 429, no. 6993, p. 734, 2004.

[3] 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.

[4] Z. Du, “Gpt2-chinese: Tools for training gpt2 model in chinese language,” https://github.com/Morizeyao/GPT2-Chinese, 2019.