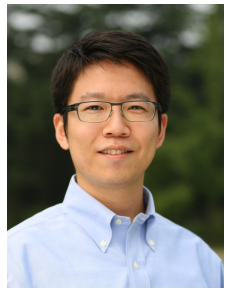
Generation of Synthetic Electronic Medical Record Text






Jiaqi Guan, guanjq14@tsinghua.org.cn

Runzhe Li, rli51@jhmi.edu

Sheng Yu, syu@tsinghua.edu.cn

Xuegong Zhang*, zhangxg@tsinghua.edu.cn

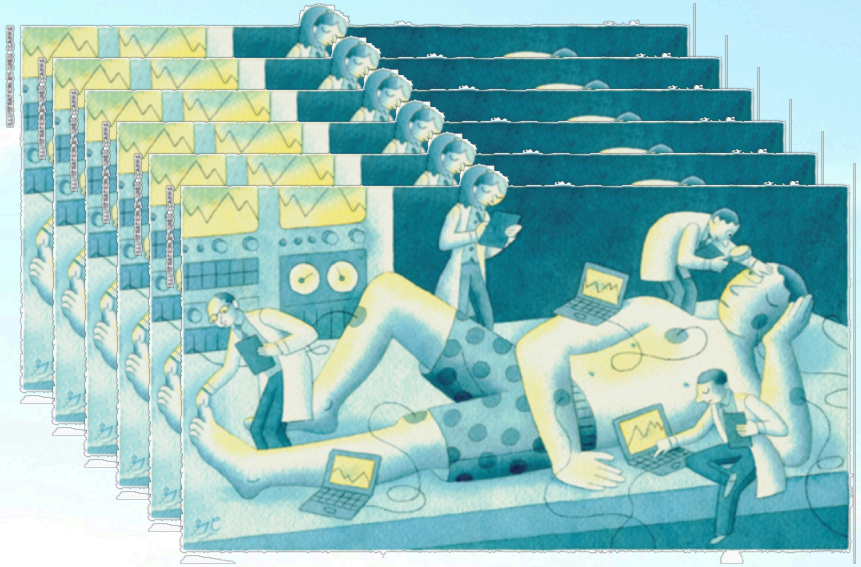
Tsinghua University

Dec. 4, 2018

Machine Learning, Big Data, and AI Medicine

Knowledge

+



DATA+ML

=

Electronic Medical Records

- Medical Big Data
  - **Electronic Medical Records (EMR, or EHR)**
    - Structured EMR
    - EMR in nature languages
  - Lab tests
  - Medical Images
  - Pharmacology data
  - Biological omics data
  - Life style data
  - Environmental data
  - …

EMR sheets in Chinese

- Most part of the data are in natural language

—— well, special medical language, not that “natural”

姓名：XXX　科室：肿瘤内科　病区：七病区　病案号：00067947

姓名：XXX　**性别**：女

年龄：43 岁　民族：汉族

婚姻状况：已婚　出生地：河北省衡水市

职业：农民　病史陈述者：患者本人

入院时间：2013-05-27 16:21　记录时间：2013-05-27 16:21

主　诉：间断咳嗽、咳痰 3 年，加重伴气短 1 月余

现病史：患者近 3 年来出现咳嗽，以干咳为主，咳痰费力，无明显气短，未予诊治。近 1 月多来感上述症状加重，伴气短，活动后明显，偶有喘鸣，行胸部 CT 示气管隆突上方占位，左锁骨上区淋巴结转移，行气管镜示气管下段肿物，活检病理为类癌。为进一步诊治收入院。

既往史：体健。否认明确的高血压、冠心病、糖尿病等慢性病史，否认肺结核、肝炎等传染病史，否认外伤史，有输血史，无明确食物、药物过敏史，预防接种史不详。

个人史：生于原籍，久居当地。近期无牧区及疫区接触史，无烟酒嗜好。

月经婚育史 21 岁结婚，育有 1 子 1 女，爱人与儿女均体健。

家族史：父母亲体健，有 2 弟均体健。否认家族遗传病史及恶性肿瘤家族史。

体 格 检 查

体温 36.6℃　脉搏 80 次/分　呼吸 22 次/分　血压 120/70mmHg

一般情况：　发育正常，营养中等，自主体位，神志清楚，对答切题，检查合作。

Name: xxx　　Sex: female

Age: 43　　Ethnic group: Han

Marital status: married

……

Major statement of symptoms: irregular cough, cough with sputum for 3 years, getting more severe in recent month accompanied with shortness of breath

Current illness history: .....

Past history: …….

Family history: …

Physiological tests: ….

…

The need for annotated data

- The great success of ML on image recognition benefited greatly from the availability of vast amount of annotated data

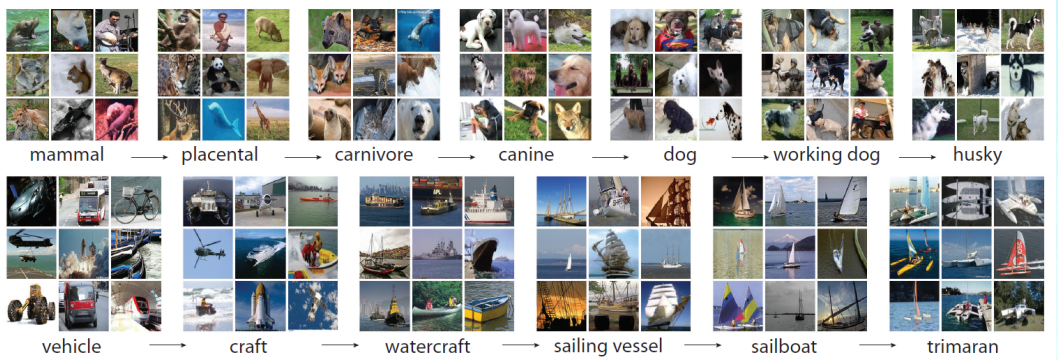
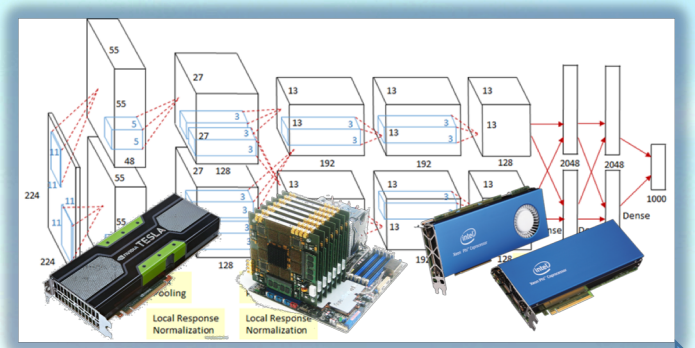


Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the **top** row is from the mammal subtree; the **bottom** row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.



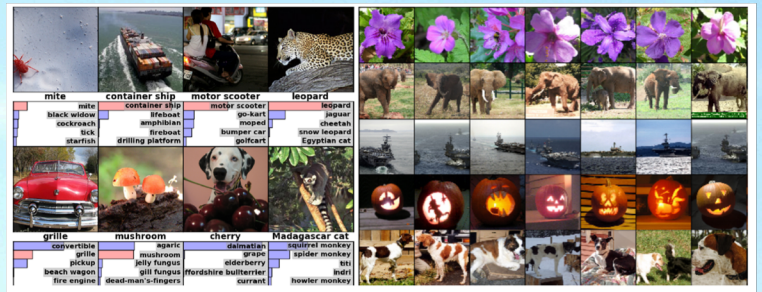
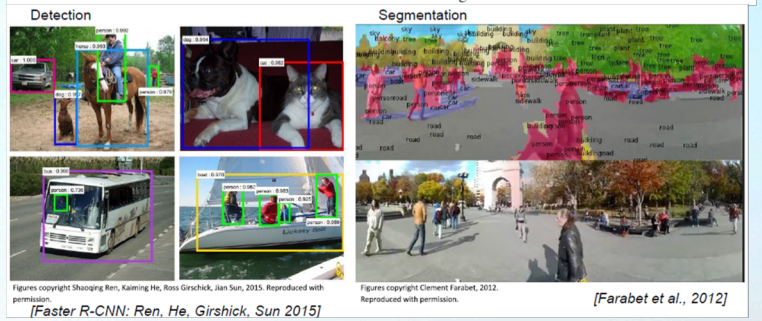


Figure 4: **(Left)** Eight ILSVRC-2010 test images and the five labels considered most probable by our model. The correct label is written under each image, and the probability assigned to the correct label is also shown with a red bar (if it happens to be in the top 5). **(Right)** Five ILSVRC-2010 test images in the first column. The remaining columns show the six training images that produce feature vectors in the last hidden layer with the smallest Euclidean distance from the feature vector for the test image.



- It's hard to get sufficient well-annotated real EMR data
- It's hard to get simplified EMR data for method study

Introduction

- We proposed a model (mtGAN) for generating “artificial” or synthetic EMR texts in Chinese, based on the GAN framework
  - Use policy gradient algorithm (REINFORCE) to train the model
  - Conditional model: take designated disease features as inputs, and generate corresponding EMR text data

- It can generate artificial EMR data

  → As controlled simulation data for many research on methods

  → The method can be adopted for transforming real EMR data to pseudo-data to avoid exposing patients’ privacy

Related Work

- Natural language generation:
  - writing poems, stories, novels...
  - little work transfer related methods to a more specific application scenario, such as EMR generation.

An example of the story generation given an image

We were barely able to catch the breeze at the beach , and it felt as if someone stepped out of my mind . She was in love with him for the first time in months , so she had no intention of escaping . The sun had risen from the ocean , making her feel more alive than normal . She 's beautiful , but the truth is that I do n't know what to do . The sun was just starting to fade away , leaving people scattered around the Atlantic Ocean . I d seen the men in his life , who guided me at the beach once more .

Samim has made an awesome blog post with lots of results here.

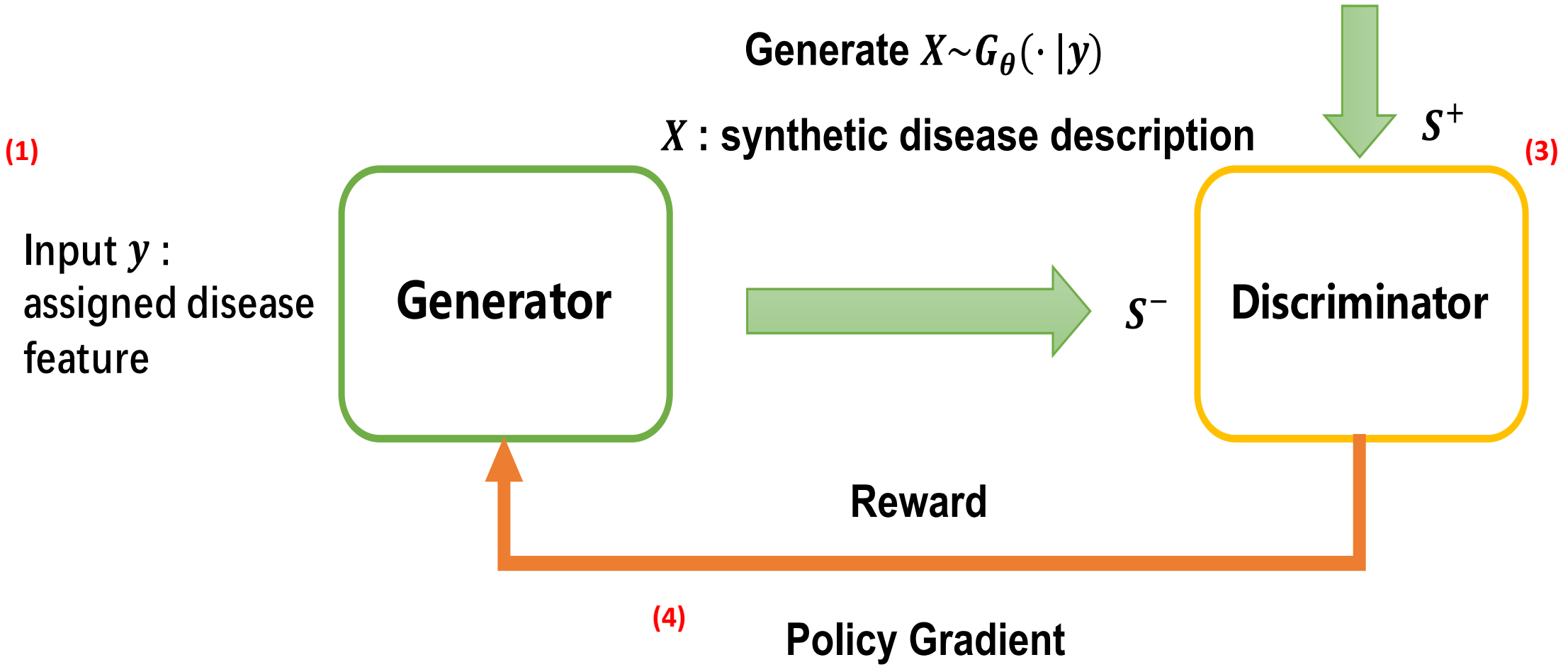
https://github.com/ryankiros/neural-storyteller#neural-storyteller

Related Work

- There has been efforts on generating synthetic EMRs, but mostly for structured EMRs (e.g., using ICD-9 codes) or specific medical data (e.g., EEG)
  - Buczak et al, Data-driven approach for creating synthetic electronic medical records. *BMC Medical Informatics and Decision Making, 2010*
  - Choi et al. Generating multi-label discrete patient records using generative adversarial networks, Proceedings of Machine Learning for Healthcare 2017
  - Esteban et al. Real-valued (medical) time series generation with recurrent conditional GANs, https://arxiv.org/abs/1706.02633v2, 2017

Method

Our model: mtGAN



Method

Model the generator as a **policy**:

- **State**: the generated medical words so far $x_{1:t-1}$
- **Action**: the next medical word to be generated x_t
- **Reward**: the output of GAN discriminator: the likelihood that the synthetic EMR text can fool the discriminator

Intermediate Reward:

Monto Carlo Search on the partially generated sentence $x_{1:t-1}$, the average score is used as the reward for x_t

Method

Objective Function:

log-likelihood

Reward

G: $$\max \ \mathbb{E}_{X\sim G_\theta(\cdot|y)}[\sum_{t=1}^{T} \log G_\theta(x_t|X_{1:t-1},y) \cdot R_{D_\phi}^{G_\theta}(X_{1:t-1},x_t,y)]$$

EMR text

disease feature

D: $$\max \ \mathbb{E}_{(X,y)\sim p_d}[\log D_\phi(X,y)]+\mathbb{E}_{X\sim G_\theta(\cdot|y)}[\log(1-D_\phi(X,y))]$$

cross-entropy

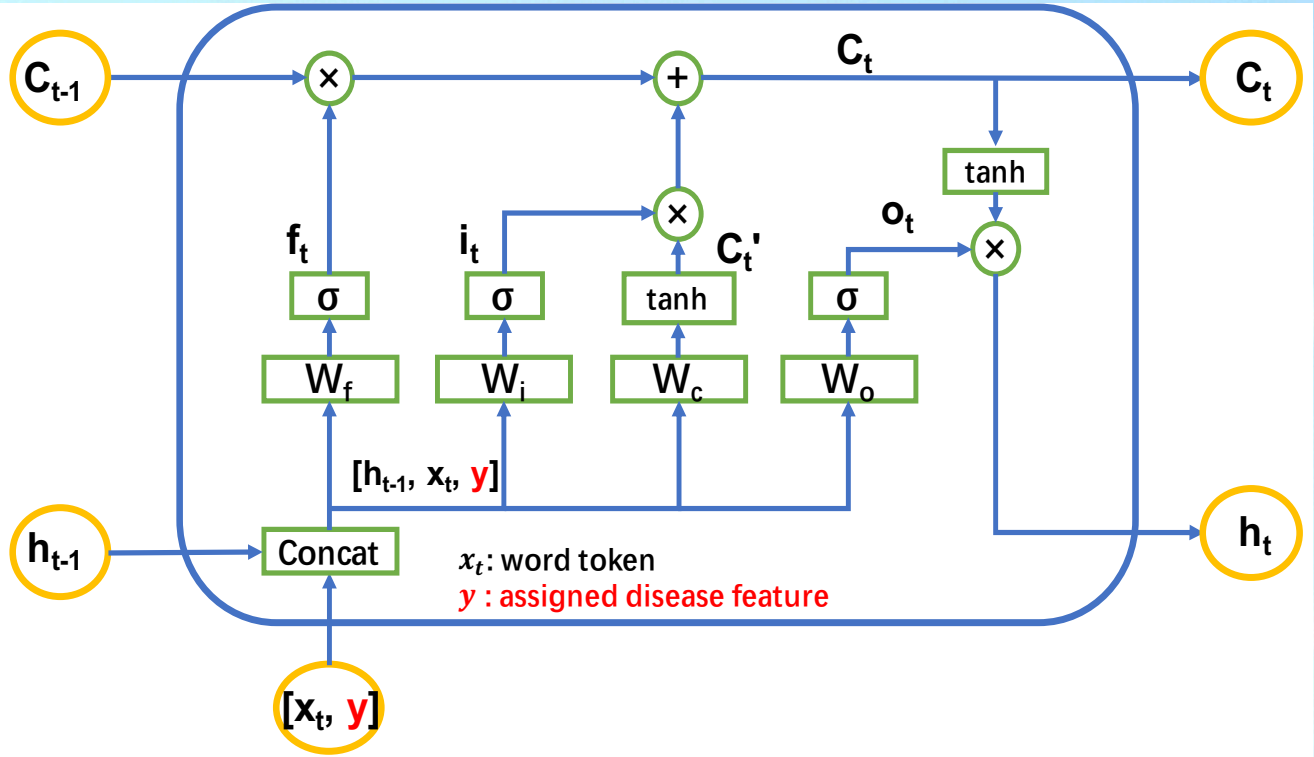
$$R_{D_\phi}^{G_\theta}(s=X_{1:t-1},y,a=x_t) = \frac{1}{K}\sum_{k=1}^{K} D_\phi(X_{1:T}^k,y)$$

$$,X_{1:T} \in \mathrm{MC}^{G_\theta}(X_{1:t},y)$$

Monto Carlo Search

Method

Generator



c_t: cell state
h_t: hidden state
W_f: forget gate
W_i: input gate
W_o: output gate
o_t: the final output

Discriminator: fastText, BiRNN or **CNN**

Method

Rescale Rewards : → alleviate gradient vanishing problem

- Optimal Discriminator Activation[1] (ODA): $\mathrm{R} = \frac{\mathrm{D}}{1-\mathrm{D}}$
- Bootstrapped Ranking Activation[2] (BRA): $\mathrm{R} = \sigma(\delta \cdot (0.5 - \frac{\mathrm{rank(i)}}{\mathrm{B}}))$
- baseline: R-b

Teacher Forcing: → alleviate mode collapse problem

- one step of teacher forcing (MLE) after one step of adversarial training

[1] Che,T.,Li,Y.,Zhang,R.,Hjelm,R.D.,Li,W.,Song,Y.,andBengio,Y. (2017). Maximum-likelihood augmented discrete generative adversarial networks.
[2] Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., and Wang, J. (2017). Long text generation via adversarial training with leaked information.

Experiments

Dataset:

- In Chinese, 2216 EMR texts
- Use the “history of present illness” field as input sequences and the “admission diagnosis” field as sequence tags
- **Tags: pneumonia and lung cancer**
- Word segmentation: *jieba* (a Chinese word segmentation python package)
- Dictionary: 7674 words. First 40 words of each EMR text are used
- split into training, validation and test set with the proportion of 0.7, 0.1, 0.2

Implement:

- Python, Tensorflow
- Pretrain G 1000 epochs by MLE, Time: 2hrs
- Pretrain D 500,100,100 epochs for fastText, CNN and BiRNN, Time: 1hr
- Adversarial training: update G five steps and then update D five steps. one step of teacher forcing
- after each adversarial step
- Total epochs: 100, Time: 2hrs

Experiments

Example of real EMR texts:

现病史:3天前无明显原因出现咳嗽，有痰不能自行咳出，伴喘憋，不能平卧，无发热，无腹泻，无便血及黑便，无尿频、尿急及排尿困难，症状进行性加重，今日患者家属发现患者不能自行进食。来我院门诊查血常规白细胞6.4×109/L，中性粒细胞89.6%，血红蛋白103g/l，血小板123×109/L，胸片双肺间质样改变，合并双下肺渗出病变，心电图为窦性心律，异常Q波，ST段改变。为进一步诊治住院治疗。起病后神志清，精神弱，饮食差，睡眠差，二便失禁。

Diagnosis:

入院诊断：双侧肺炎
高血压病3级（极高危组）
2型糖尿病
冠状动脉粥样硬化性心脏病
无痛性心肌缺血

Data entry for the model:

Disease description:
3/ 天/ 前/ 无明显/ 原因/ 出现/ 咳嗽/ , / 有痰/ 不能/ 自行/ 咳出/ , / 伴/ 喘憋/ , / 不能/ 平卧/ , / 无/ 发热/ , / 无/ 腹泻/ , /无/ 便血/ 及/ 黑便/ , / 无/ 尿频/ 、/ 尿急/ 及/ 排尿困难/, / 症状/ 进行性加重/ ,

Tag: 肺炎

Experiments

Examples of generated synthetic EMR texts:

Pneumonia: 患者于1周前无明显诱因出现咳嗽，咳白色粘痰，伴活动后加重，休息后可缓解，间断服用镇咳药物等治疗，未行正规诊治。

The patient had cough with no obvious cause a week ago, coughing white phlegm, getting more severe after activity, can be relieved after rest, intermittently used antitussive drugs and others for treatment, no professional diagnosis and treatment.

Lung Cancer: 患者10多年前开始出现咳嗽、咳痰，痰中带血，当地医院查胸部CT示纵隔肿大淋巴结，右肺下叶结节，后行气管镜时治疗收入院。

The patient began to cough with sputum more than 10 years ago, with blood in the phlegm. Chest CT showed mediastinal enlarged lymph nodes and nodules in the right lower lobe of the lung. Be admitted to the hospital after endoscopic treatment.

More examples

- Good examples: (words matching, grammar correct, logical)
  - 患者于1周前无明显诱因出现咳嗽，咳白色粘痰，伴活动后加重，休息后可缓解，间断服用镇咳药物等治疗，未行正规诊治。
  - 患者1年前无明显诱因出现咳嗽、咳痰伴气喘，多于受冷或天气转凉时发作，持续时间在2周至半年不等，经抗感染等治疗后好转。无明显诱因，冬季多发，
  - 患者于3天前无明显诱因出现咳嗽、咳痰，为黄色粘痰，伴活动后气喘，一年四季均有发作，无气喘，无咯血，无咽痛，无胸痛，有时伴发热
  - 患者无明显诱因出现反复出现腹胀、腹泻，稀水样便，有恶心、呕吐，呕吐物为胃内容物。无畏寒、发热。就诊于当地社区医院，诊断为"急性肠炎”
  - 患者2014.6查颈部不适。无明显诱因出现轻微咳嗽，以干咳为主，咳痰，伴气喘，无发热，于当地医院给予左氧氟沙星抗感染治疗未见腺癌
  - 患者10多年前开始出现咳嗽、咳痰，痰中带血，当地医院查胸部CT示纵隔肿大淋巴结，右肺下叶结节，后于北京人民医院行气管镜时治疗收入院。

More examples

- Bad examples
- **Repetitive**
  - 患者9年前受凉后出现咳嗽、咳痰、气喘，活动后加重，喘憋明显，给予抗感染等对症治疗后好转，后进食后症状可改善，给予抗感染等对症治疗后好转
  - 患者院外病情尚稳定，10年前出现无发热，咳嗽、咳痰，痰为白色粘痰，伴咯血、气喘、憋气，尚可平卧，无发热，无畏寒、寒战，体温最高
  - 患者1天前无明显诱因出现咳嗽、较剧烈，伴咳少许白色粘痰，间断出现咳嗽，咳白色痰，粘稠，不易咳出，感胸闷减轻，伴气喘，不伴有心前区
- **Inconsistent**
  - 患者无明显诱因出现咳嗽、咳痰渐感咳嗽、咳白色粘痰，最高体温39.5℃，无咳血、发热，无畏寒
  - 患者2014.1无明显诱因出现咳嗽，以干咳为主，无痰，无咳嗽、咳痰，无咯血、胸痛，无发热，无胸痛、胸闷、心悸，无畏寒
- **Improper word matching**
  - 患者反复出现反复咳嗽，咳白色粘痰，伴活动后气短，无午后发热，体温痰量增多、体温阵发性咳嗽，就诊于北京东直门医院，镜身不能慢性阻塞性肺疾病合并肺部感染
  - 患者于3年前无明显诱因出现咳嗽、咳痰，每年发病后出现气喘，出现颜面部及眼睑呕血憋气。呼吸无欠佳后体温就诊于当地医院查发现右上叶及肺囊肿切除术

Experiments

Micro-Level Evaluation

$$\mathrm{NLL}_{test} = -\mathbb{E}_{\mathbf{x}\in S_{\text{test}}}[\sum_{t=1}^{T} \log G_\theta(x_t|X_{1:t-1})]$$

$$\text{self-BLEU} = \mathbb{E}_{\mathbf{x}\in S_{\text{test}}}[\mathrm{BLEU}(X|\complement_{S_{\text{test}}}(X))]$$ [1]

Model	NLL-test	self-BLEU
MLE	6.2141	0.9270
SeqGAN	5.9685	0.9267
mtGAN-BiRNN-BRA	6.1191	0.9155
mtGAN-BiRNN-ODA	5.9561	0.9209
mtGAN-CNN-ODA	**5.7764**	**0.9182**

[1] Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., and Wang, J., et al. (2018). Texygen: a benchmarking platform for text generation models.

Experiments

Macro-Level Evaluation[1]

- Adversarial Success: AdverSuc
- Evaluator reliability error ERE

1) Randomly split real EMR texts as positive examples and negative examples. (Ideal Acc: 0.5)
2) Randomly split generated EMR texts as positive examples and negative examples. (Ideal Acc: 0.5)
3) Use real EMR texts as positive examples and random generated EMR texts as negative examples. (Ideal Acc: 1.0)

Model	AdverSuc	ERE1	ERE2	ERE3	meanERE
MLE	0.3007	0.0068	0.0676	0.3446	0.1396
SeqGAN	0.1351	0.0203	0.0270	0.1081	0.0518
mtGAN	**0.3041**	0.0811	0.0270	0.2804	0.1295

[1] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017). Adversarial learning for neural dialogue generation.

Experiments

Application-Level Evaluation

- Can be used as a data augmentation approach

Accuracy / Data source / Model	Real	Synthetic	Mix
MLE	0.7500	0.7432	0.7568
SeqGAN	0.7500	0.6959	0.7095
mtGAN	0.7500	0.7432	**0.7635**

Conclusion

- We proposed a conditional model mtGAN to generate synthetic EMR texts.
- It can help to solve the problem of privacy and of insufficient and imbalance samples in utilizing big EMR data.
- The micro-level, macro-level and application-level evaluation demonstrate that our model can generate EMR texts that are very similar to real EMRs.

Discussion

- Existing problems: repetition, inconsistent, etc.
  - → combine rewards given by Discriminator with human-designed rewards
  - → only apply constraints on the generation phase
- Challenges of long EMR text generation:
  - the consistency and logic of the whole text
- Encoder-decoder model:
  - build a map between a noise distribution and an EMR text distribution
  - provide a more convenient approach to low-dimensional representation of EMR, the similarity evaluation between EMRs

Thanks!