英伟达在自然语言生成领域的最新进展

齐家兴 英伟达 高级解决方案架构师

目录

01 多角色生成式对话建模

背景

生成式闲聊对话

- 闲聊式对话系统有两类方法：检索式和生成式
- 生成式方法具有灵活性和创造性，能生成风格丰富的内容。
- 可控性稍差，内容风格多变，或者生成无聊的内容
- 添加说话人或者情感，但需要大量的标注数据。

背景

- 让模型基于某个说话人之前的对话内容，去生成风格一致的新的内容。
- 通过自动测评和人工测评，我们发现这种方法比之前的方法有更好的可控性，能生成更合适的对话回复。
- 增加模型尺寸，从117M 到 3.8B，模型越大，对话效果越好。

方法

GPT-2 自回归语言模型

GPT-2模型自回归式逐个token生成

$$p_\theta(\mathbf{x}) = \prod_{t=1}^{|\mathbf{x}|} p_\theta(x_t | x_{<t})$$

基于之前对话内容，生成新一轮对话

$$p_\theta(\mathbf{c}) = \prod_{j=1}^{|\mathbf{c}|} p_\theta(\mathbf{x}_j | \mathbf{x}_{<j})$$

$$\mathbf{c} = (\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathbf{c}|})$$

方法

添加目标说话人历史对话

借助额外的对话历史作为输入的一部分，控制模型生成的内容风格

$$p_\theta(\mathbf{c}|\mathbf{r}) = \prod_{j=1}^{|\mathbf{c}|} p_\theta(\mathbf{x}_j|\mathbf{x}_{<j}, \mathbf{r}_j)$$

提取目标说话人在数据集中的对话历史

$$\mathbf{r}_j = \{(\mathbf{x}_{k-1}, \mathbf{x}_k) \mid \text{author}(\mathbf{x}_k) = \text{author}(\mathbf{x}_j) \wedge \mathbf{x}_k \notin \mathbf{c}\}$$

模型

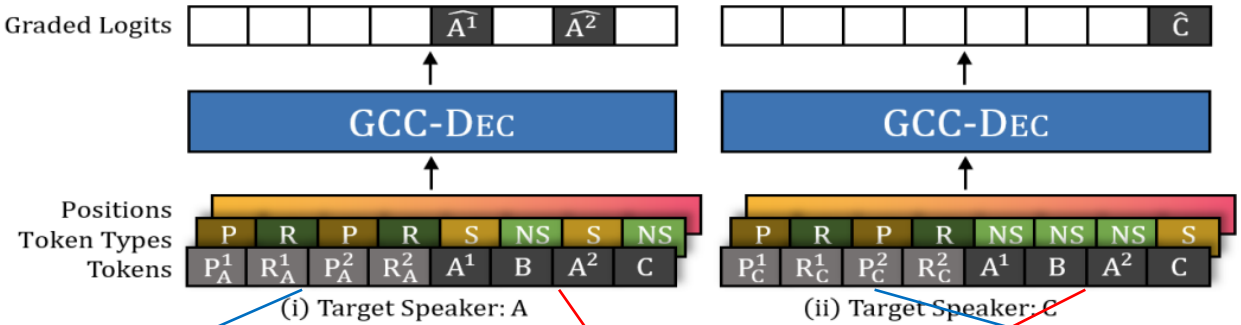
结构与步骤

- 模型: 生成式对话控制模型 (Generative Conversation Control model, GCC)
- 结构：GPT-2
- 主要步骤：
  - 确定想要模仿的说话人（风格），从数据集中提取他的历史对话作为参考
  - 将参考和当前对话的已有的内容输入给模型
  - 自回归式的对话生成

模型

输入格式

举例说明模型输入：三人 (A, B, C) 的四轮对话 (A1, B, A2, C)



说话人A的历史对话，作为参考

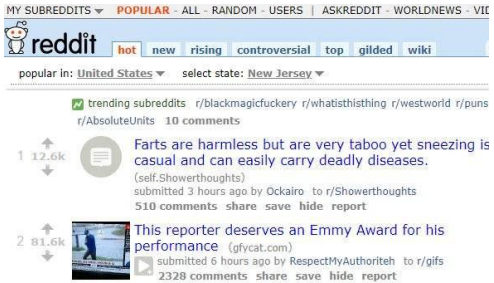
目标多轮对话

说话人C的历史对话，作为参考

数据

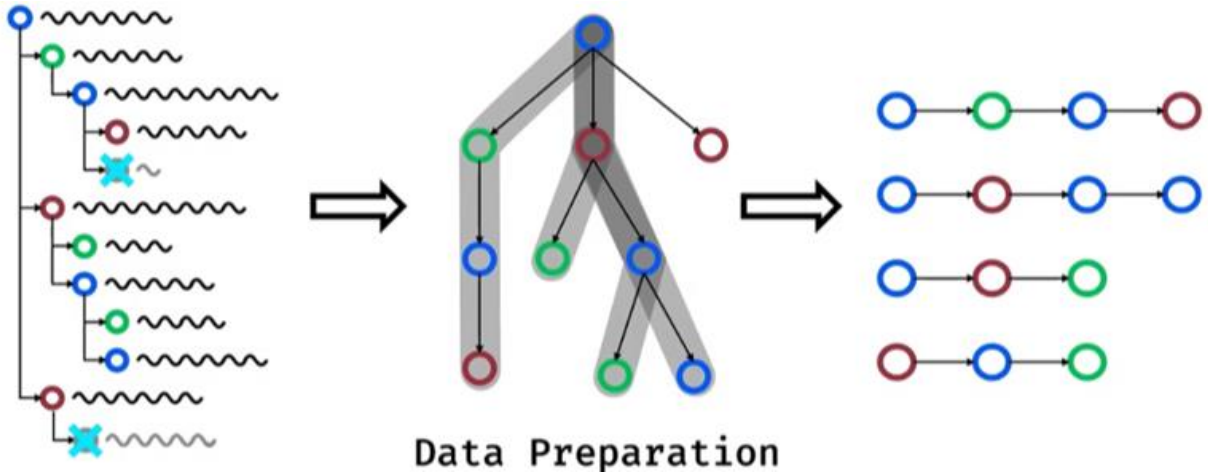
ADD RELATED TITLE WORDS

- 从pushshift.io下载了Reddit 论坛里的网友提问和回复，作为对话数据
- 训练集：2018.10 – 2019.3
- 测试集：2019.4



数据

ADD RELATED TITLE WORDS



- 树形论坛回复，转换为多人序列对话形式
- 按经验设定条件清洗数据，保证对话数据的多人，有效和丰富性。
- 对每位网友，保存最多8条回复。实际平均每人7.1条。

对比实验

自动测试与人工测试

使用基于困惑度的自动测试，对比几种不同的 baseline 模型结构：

- Decoder-only: GCC-DEC
- Seq2Seq baseline: GCC-S2S
- Variational Autoencoder baseline：GCC-VAE
- No Reference Context baseline: GCC-NRC

以及预训练是否有帮助

之后再进行人工测试，测评生成对话的实际效果。

结果

模型结构对比

基于困惑度（PPL）的测试

Model	h	l	A	Params	PPL
GCC-S2S	768	18	16	375M	22.09
GCC-VAE	768	20	16	362M	22.43
GCC-DEC	1024	24	16	355M	**19.10**
GCC-S2S	1024	24	16	810M	19.89
GCC-VAE	1024	24	16	711M	20.49

对比不同模型结构和大小

结果

预训练的影响

基于困惑度（PPL）的测试

Model	P.T.	Iter.	Batch	PPL
GCC-DEC	✗	200K	256	19.10
GCC-DEC	✓	70K	128	**18.92**
GCC-NRC	✓	70K	128	19.39

是否使用预训练 (P.T.) 模型的对比测试

测试

模型参数量的影响

不同模型大小的困惑度

Model	h	l	A	Params	PPL
GCC-DEC	768	12	12	117M	23.14
GCC-DEC	1024	24	16	355M	18.92
GCC-DEC	1280	36	16	774M	17.18
GCC-DEC	1536	40	16	1.2B	16.08
GCC-DEC	3072	72	24	8.3B	**13.24**

测试

人工评价对话生成质量

基于PPL的自动测试之后，我们又进行了人工测评，从多个维度评价生成对话的质量：

- 真实度
- 风格一致性
- 内容质量
- 连贯性

方法：给定一段真实的N轮对话的前N-1轮内容，最后一轮对话是不同的实现方式（模型生成或者真实的）。把两种方式的对话都展示给测评者，让其对比。

测试

对话生成质量

Source A	Realistic	Reference	Quality	Coherency	Source B
GCC-NRC (355M)	31% - 35%	37% - 41%	29% - 36%	32% - 39%	Human
GCC-DEC (355M)	32% - 34%	38% - 40%	31% - 33%	32% - 36%	Human
GCC-DEC (774M)	31% - 35%	40% - 39%	33% - 33%	34% - 36%	Human
GCC-DEC (1.2B)	32% - 37%	40% - 40%	34% - 38%	29% - 36%	Human
GCC-DEC (8.3B)	**37% - 40%**	**42% - 38%**	**42% - 42%**	**40% - 42%**	Human
GCC-DEC (355M)	31% - 34%	41% - 39%	37% - 36%	33% - 35%	GCC-NRC (355M)
GCC-DEC (774M)	33% - 33%	39% - 40%	34% - 29%	34% - 36%	GCC-DEC (355M)
GCC-DEC (1.2B)	31% - 31%	40% - 38%	33% - 32%	38% - 38%	GCC-DEC (774M)
GCC-DEC (8.3B)	41% - 37%	39% - 43%	38% - 38%	42% - 39%	GCC-DEC (1.2B)

总结

- 传统的生成式对话存在风格不可控的问题。
- 借助目标说话人的历史对话，可以让生成式模型生成风格更一致的对话内容。
- 借助 Reddit 论坛获取真实网友的回帖作为对话训练数据。
- 通过自动测试和人工评测，对比了多种模型结构和大小，结果显示基于decoder结构的模型效果最好，并且效果随着模型参数增大会进一步提升。

02 借助合成数据训练问答模型

背景

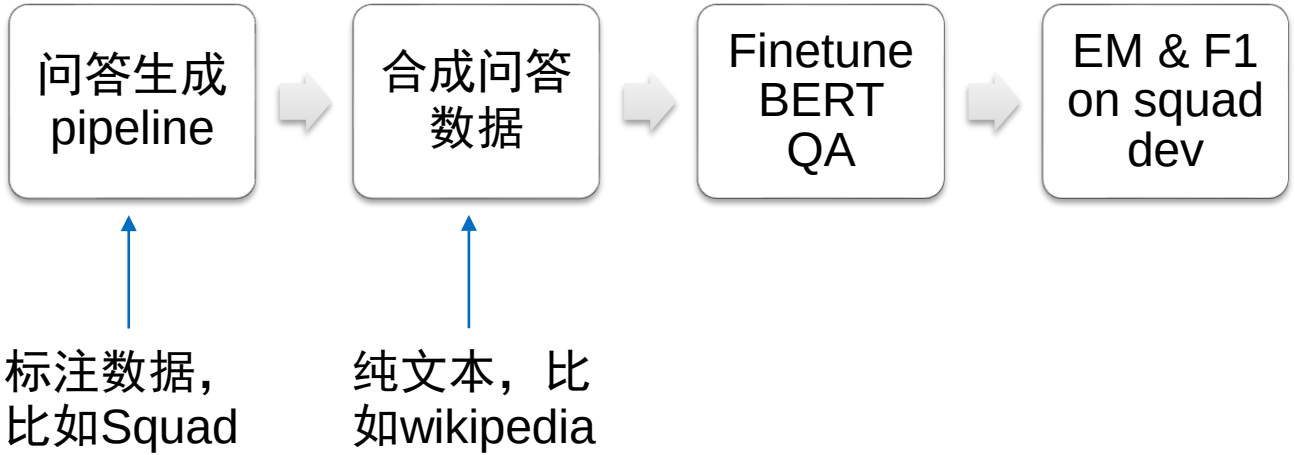
- 深度学习高度依赖标注数据
- 半监督学习：在已有数据上训练一个模型，再用它去标注新数据
- 但是对于问答任务来说，合成数据的质量与人工标注数据还有一定差距。

背景

- 我们希望能借助生成模型，在给定的一个文档集合上，用合成的方式来生成问题与答案，从而作为一种数据增强的方式来帮助更好的训练问答模型
- 在 SQUAD 1.1 测试集上，我们使用纯合成数据训练而来的模型达到比使用训练集标注数据的模型的精度更好。
- 更进一步，我们尝试在生成文本上，再生成问答对，这样的合成数据能用来finetune模型，达到真实文本上相近的精度。

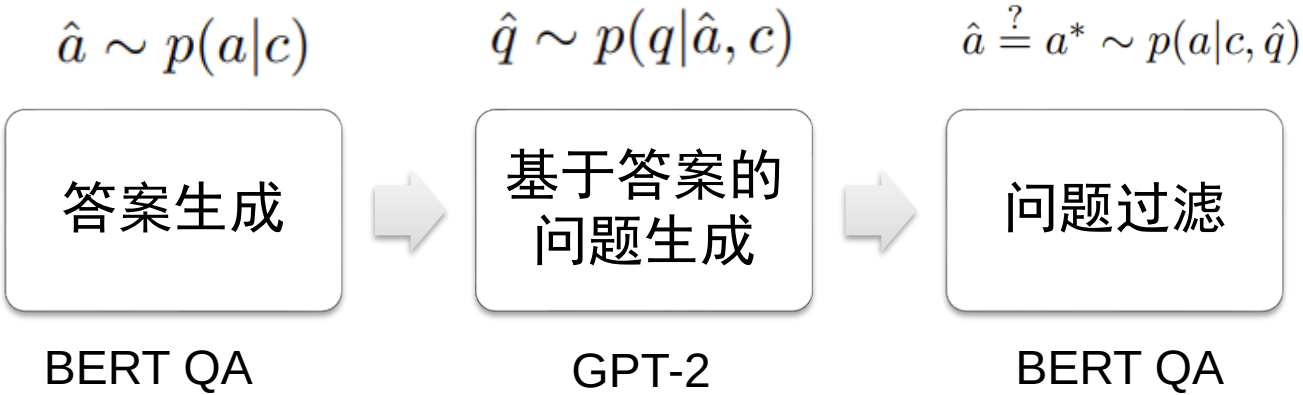
方法

完整流程


问答生成 pipeline
合成问答数据
Finetune BERT QA
EM & F1 on squad dev
标注数据，比如Squad
纯文本，比如wikipedia

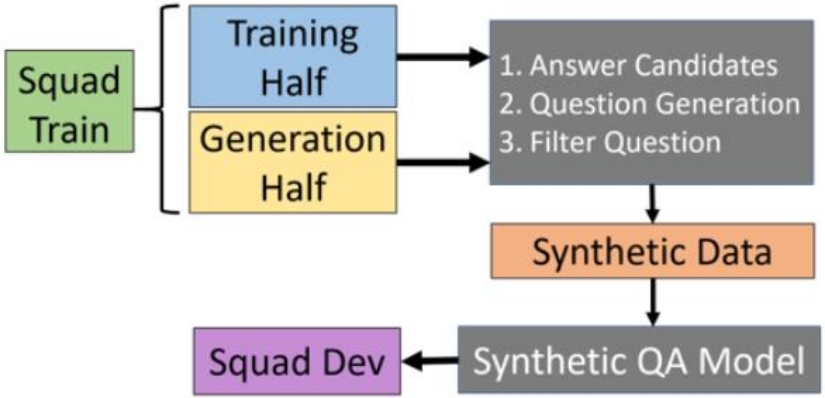
方法

答案 – 问题 – 过滤


$\hat{a} \sim p(a|c)$
$\hat{q} \sim p(q|\hat{a}, c)$
$\hat{a} \stackrel{?}{=} a^* \sim p(a|c, \hat{q})$
答案生成
基于答案的问题生成
问题过滤
BERT QA
GPT-2
BERT QA

数据

ADD RELATED TITLE WORDS



消融实验，使用SQUAD一半的数据做训练，另一半数据用来生成。

另外，使用全部SQUAD数据来测试最佳精度。

结果

最好的模型

1. 先用SQAUD 1.1 训练问答生成pipeline。
2. 然后分别在真实的wiki上 和 合成的wiki上，进行问答合成。
3. 最后在合成的问答数据上 finetune BERT QA，并测试EM & F1
4. 再用 SQUAD 进一步 finetune

Text Source	Source Data Size	finetune data	# Questions	EM	F1
Wikipedia	638 MB	Synthetic	19,925,130	88.4	94.1
		+SQuAD	20,012,729	**89.4**	**95.2**
8.3B GPT-2	480 MB	Synthetic	17,400,016	88.4	93.9
		+SQuAD	17,487,615	**89.1**	**94.9**
SQuAD1.1	14MB	SQuAD	87,599	87.7	94.0

结果

模型大小对合成pipeline的影响

Model Size				# Questions	EM	F1
Answer	Question	Filter	QA			
345M	345M	345M	345M	116721	85.3	92.0
1.2B	**1.2B**	**1.2B**	**345M**	**184992**	**87.1**	**93.2**
Human Generated Data			345M	42472	86.3	93.2

结果

模型大小对问题生成的影响

Question Generator	# Questions	EM	F1
117M	42345	76.6	85.0
345M (Klein and Nabi, 2019)	-	75.4	84.4
345M (w/ BERT QA model)	42414	76.6	84.8
345M	42414	80.7	88.6
768M	42465	81.0	89.0
1.2B	42472	83.4	90.9
8.3B	**42478**	**84.9**	**92.0**
Human Generated Data	42472	86.3	93.2

生成的问题质量随着模型增大而变好

结果

模型大小对答案生成的影响

Answer Generator	#Questions	EM	F1
BERT-Large	227063	77.7	87.6
BERT-345M	229297	79.1	87.9
BERT-1.2B	**229067**	**79.2**	**88.3**
Human Generated Answers	42472	83.7	91.1

生成的答案质量与模型大小不相关

结果

模型大小对问题过滤的影响

Filter Model	# Questions	EM	F1
Synthetic Questions + Real Answers			
BERT-Large	45888	84.5	91.4
BERT-345M	34341	84.2	91.4
BERT-1.2B	**47772**	**85.6**	**92.4**
Synthetic Questions + Synthetic Answers			
BERT-Large	177712	85.5	91.9
BERT-345M	144322	85.9	92.5
BERT-1.2B	**184992**	**87.1**	**93.2**
Human Generated Data	42472	86.3	93.2

增大模型可以提升问题过滤的效果

总结

- 借助生成模型，在给定的一个文档集合上，用合成的方式来生成问题与答案，从而作为一种数据增强的方式来帮助更好的训练问答模型
- 在 SQUAD 1.1 测试集上，使用纯合成数据训练而来的模型达到比使用训练集标注数据的模型的精度更好。
- 通过模型大小的对比实验，增大模型对问题生成和问题过滤都有更高的精度。但是对答案生成的影响比较小。

参考文献

1. Puri R , Spring R , Patwary M , et al. Training Question Answering Models From Synthetic Data[C]// 2020.
2. Boyd A , Puri R , Shoeybi M , et al. Large Scale Multi-Actor Generative Dialog Modeling[C]// 2020.
3. Xu P , Patwary M , Shoeybi M , et al. MEGATRON-CNTRL: Controllable Story Generation with External Knowledge Using Large-Scale Language Models[C]// 2020.

非常感谢您的观看

NVIDIA. INGRAM MICRO | DataFun.