0507 汇报

项目三: 基于大规模预训练模型的生成式知识问答

朱睿涵　李宸亦

2022 年 5 月 7 日

目录

PANGUBOT 简介

PANGUBOT 是一种基于大型预训练语言模型 PANGU- 的汉语预训练开放域对话生成模型.
PANGUBOT 有如下特点:

1. 与其他在海量对话数据上从头开始训练的预训练对话模型不同，PANGUBOT 通过**继承**预训练语言模型中宝贵的语言能力和知识，以相对较少的数据和计算成本建立一个强大的对话模型
2. PANGUBOT 在反应质量、知识和安全性**优于**最先进的中文对话系统
3. PANGUBOT 可以很容易地被部署，以产生情感反应，而不需要进一步的训练

PANGUBOT 数据来源

PANGUBOT 数据包括社交媒体、基于知识的对话、问答三种类型. 与其他大型对话模型相比，PANGUBOT 的训练数据要小得多

- 社交媒体数据主要考虑两个主流媒体: 微博和豆瓣, 有 STC,RGC,LCCC-large 和 Douban 四个数据集
- 基于知识的对话数据有 DuConv 数据集
- 问答数据有 Children Dialog 和 CQA 两个数据集

Dataset	Domain	# of dialog	# of utterances
LCCC-large (Wang et al., 2020)	Social Media	12.0M	32.9M
Douban (Wu et al., 2017)	Social Media	33K	1.8M
STC (Shang et al., 2015)	Social Media	4.4M	8.9M
RGC-2M (Cai et al., 2019)	Social Media	2M	4M
DuConv (Wu et al., 2019)	Wiki Dialog	30K	270K
Children Dialog†	Wiki QA	4.7M	9.5M
CQA†	Web QA	20.3M	40.7M
Overall		~44M	~100M

图: 数据集描述与统计

用于训练 PANGUBOT 的数据集包含 4400 万个多回合对话会话，包含 1 亿个话语和 13 亿个 token。它比用于训练 EVA 和 PLATO-XL 的数据集要小一个数量级以上。

数据清洗

由于很多对话数据来自开放资源，为了保证对话数据的质量，进行如下预处理步骤:

1. 去掉不含汉字的话语
2. 通过匹配预定义的黑名单词汇表来删除不当话语
3. 删除带有特殊字符、url 或敏感信息 (如电子邮件地址或个人 id) 的话语
4. 删除可能含有广告内容的话语
5. 把一个话语中连续重复的字符缩短为三个字符的最长长度
6. 删除超过 100 字的对话

PANGUBOT 模型训练

因为 PANGU-α 已经在一系列的 NLP 任务中表现得很好所以并非从零开始训练 PANGUBOT 而是直接从 PANGU-α 中继承参数，然后根据对话数据进行训练. 因此 PANGUBOT 使用与 PANGU-α 相同的架构，即类似 gpt 的自回归语言模型.
给定对话历史或上下文,PANGUBOT 的目标是产生一个响应 y，最大化式 (1)

$$p_\theta(\mathrm{y}|X) = \prod_{t=1}^{n} p_\theta(y_t|y_{<t}, X), (1)$$

其中 $X = \{x_1, x_2 \ \ldots \ x_{t-1}\}$ 为一系列句子,n 是响应的长度

PANGUBOT 训练细节

由于此模型训练数据较少，因此将训练损失用于反应和对话语境，即具有多个回合的话语，充分利用数据. 此外，为了确保在一个训练步骤中不受前一个对话会话上下文的干扰，通过重置对话会话之间的掩码使模型只看到当前对话会话中的前一个标记.

训练参数:

对于 PANGUBOT350M

- 使用一个 24 层 transform，隐藏层大小为 1024，并设置 attention-head 的数量为 16
- 16 batch size / GPU
- 使用 16 块 NVIDIA V100 图形处理器训练

对于 PANGUBOT2.6B

- 使用 32 层 transform，隐藏层大小为 2560，并设置 attention-head 的数量为 32
- 8 batch size / GPU
- 使用 32 块 NVIDIA V100 图形处理器训练

因此两种规模模型在一次训练中学习到的 token 数量是相同的，每 10 万 step，大约 20 个 epoch. PANGUBOT350M 总训练时间为 2.5 天左右，PANGUBOT2.6B 总训练时间为 5.5 天左右.

实验方法

实验分为 4 个部分:

1. 研究整体的对话反应质量
2. 研究 PANGUBOT 捕获知识量
3. 研究不同对话模型的安全问题
4. 证明 PANGUBOT 可以很容易地用于产生情绪反应

整体的对话反应质量

自聊对话流程: 以一个预定义的第一轮提示开始，这些提示来自 7 个常见的领域 (聊天、文学、体育、地理、旅游、常识、电影) 总共有 50 个提示。对话模型使用 5 个随机种子进行另外 5 轮 (10 轮) 的自聊对话，从而产生 250 个对话。
评估标准:

1. 自动评估: 计算了 250 对话回答的平均长度以及 dis-n，以衡量生成的回答的语言多样性
2. 人工评价: 选择了 50 个对话，由三位评注家从以下五个方面进行评价:Sensibility,Specificity,Interestingness,Hallucination,Safety
3. 人机交互评估: 参与者用不同的对话模式进行对话，并判断他们的回答质量

整体的对话反应质量

	Human Evaluation						Automatic Evaluation		
Model	Sensibility	Specificity	Interestingness	SSI	Hallucination ↓	Safety	Dist-1	Dist-2	Avg. Len
CDIALGPT	0.663	0.567	0.407	0.546	0.108	0.965	0.049	0.210	5.0
EVA	0.379	**0.776**	0.505	0.553	0.139	0.970	0.045	0.261	**9.4**
PANGUBOT350M	0.823	0.693	**0.555**	**0.690**	0.101	0.990	**0.070**	0.320	6.9
PANGUBOT2.6B	**0.834**	0.664	0.529	0.676	**0.098**	0.992	0.068	**0.322**	7.0

Table 3: Self-chat results of different dialog models using both human evaluation and automatic evaluation.

图: 自聊评估结果

Model	Sensibility	Specificity	Interestingness	SSI	Hallucination ↓	Safety
CDIALGPT	0.737	0.388	0.279	0.468	0.068	0.984
EVA	0.573	**0.715**	0.331	0.540	0.057	0.986
PANGUBOT350M	0.748	0.608	0.328	0.561	0.037	**0.998**
PANGUBOT2.6B	**0.803**	0.602	**0.337**	**0.580**	**0.034**	0.996

Table 4: Interactive human evaluation results of different dialog models.

图: 交互式人工评价结果

PANGUBOT 两种变体在总体反应 SSI 质量上仍优于 CDIALGPT 和 EVA，且幻觉评分较低，安全评分较高. 其中 PANGUBOT2.6B 在交互式人评价方面略优于 PANGUBOT350M, 表明 PANGUBOT2.6B 在与真人对话时具有更好的性能.

知识回应

评估方式: 从网络论坛众包了中文问答对, 确保所有的问题都可以被认为是在“常识”的水平上，可以被 K-12 岁的孩子回答，或可以通过一些在线资源，如搜索引擎的帮助下推断, 还确保答案是 (一个或几个) 可以用几个标记 (少于 10 个) 描述的简单实体. 通过这种方式收集问题数据，并附上答案和证据.
评估标准:

1. 自动评估: 一元分词精确度 (“P”)、召回率 (“R”) 和 F1 分数，它们衡量黄金答案和生成的响应之间的重叠。
2. 人工评价: 要求众包工作者检查答案是否正确，也就是人的准确性 (“H-Acc.”)

知识回应

对于 PANGU-α 和 PANGU- bot，两种模型的性能都明显优于其他模型. PANGUBOT 甚至可以在两种模型配置中大幅度超过 PANGU-α, 原因可能为没有提示的 PANGU-α 更倾向于作为一种语言模型，而不是对话或问答系统.

Model	P	R	F1	H-Acc.
Without evidence				
CDIALGPT	3.3	6.7	4.1	3.6
EVA	0.8	5.1	1.2	3.6
PLATO	24.1	30.2	25.4	23.8
PANGU-α 350M	13.1	46.5	17.7	35.7
+ prompt	18.1	49.7	21.6	41.7
PANGU-α 2.6B	17.8	50.6	22.5	38.1
+ prompt	33.2	57.5	37.7	48.9
PANGUBOT 350M	**58.7**	63.6	**58.8**	60.7
PANGUBOT 2.6B	56.7	**66.2**	57.3	**66.7**
With evidence prompt				
PANGU-α 350M				
+ 0-shot	6.5	32.1	8.8	14.3
+ 3-shot	19.0	23.5	18.0	19.0
PANGU-α 2.6B				
+ 0-shot	7.1	34.8	9.2	25.0
+ 3-shot	18.2	26.7	19.0	26.2

Table 6: Results of knowledge evaluations under two configurations with or without evidence. H-Acc. is human evaluation accuracy.

安全性问题

收集手工制作的对抗性提示并进行人工评估，以彻底衡量对话系统的安全性.
将不安全分为以下 3 类:

- Harmful
- Offensive
- Controversial

在这些类别的基础上设计了三组模板和关键词来构建对抗性提示并与不同类别的不安全对话中的对话系统进行交互. 对于每一类，起草了大约 160 条对抗性提示，作为四个经评估的对话系统的输入. 然后使用人工注释器来评估生成响应的安全性.

安全性问题

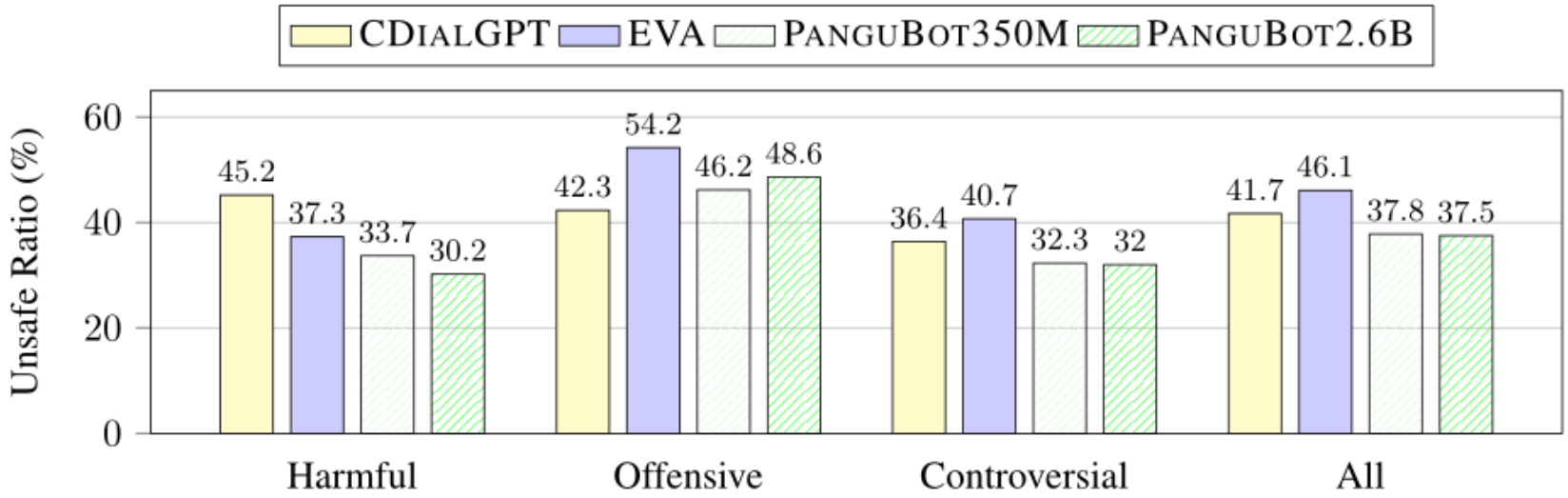


Figure 3: Ratio of unsafe responses of different dialogue models in different categories.

虽然 PANGUBOT 的两个版本对恶意上下文有更多的相关响应，但与 CDIALGPT 和 EVA 相比，PANGUBOT 的不安全性略低. 所有这四种模式在对抗性提示下仍然有较高的产生不安全反应的倾向，在改善对话安全以建立更可靠和可用的对话系统方面仍有很大的空间.

生成情绪反应

User Input	Model Responses in Different Emotion Conditions
生活总是让人琢磨不定 Life is always unpredictable	[Happy] 哈哈哈,你开心就好 Hahaha, just be happy [Angry] 谁惹你了,告诉我! Who messed with you, tell me! [Sad] 我也不知道咋了,可能是因为老了 I don't why, maybe it's because I'm old
期想找个海边的地方度假 Looking for a beach vacation	[Happy] 哈哈哈,好开心啊 Hahaha, so happy [Angry] 我不知道你是谁啊 I don't know who you are [Sad] 我去啊你这是赤果果的打击啊 Oh my god, you're clearly blowing me
知道真相的我眼泪掉下来 Knowing the truth, my tears fall	[Happy] 是不是很有成就感 Is it a sense of achievement [Angry] 你这是要干嘛? What are you doing? [Sad] 真是个悲伤滴孩子 What a sad kid

Table 9: Results of PANGUBOT2.6B generating different responses conditioned on different emotions.

PANGUBOT2.6B 确实产生了合理的反应，我们可以很容易地分辨出他们的情绪. 这个结果很有趣，因为 PANGUBOT2.6B 没有经过任何情绪对话数据集的训练，但它确实理解简单的情绪提示，并产生相应的情绪反应.

总结分析

在大型预训练模型 PANGU-α 的基础上，得到一个具有 350M 和 2.6B 参数的中文开放领域对话模型 PANGUBOT. PANGUBOT 具有强大的开放领域对话性能和较高的训练效率, 与目前最先进的对话系统相比, 在对话质量、对话知识、对话安全、对话情感四个方面都有出色表现.

但是还需要注意, 除了知识之外更多的维度，如角色、共情、记忆等能否以更普遍的方式进行建模将是一个重要的问题, 此外反应的安全性仍非常关键, 长尾分布的不安全案例对服务提供者是有巨大影响, 安全性是在实践中应用代模型的最危险的部分.

PANGUBOT 已取得一定成绩, 但之后仍有工作需要完成.

目录

PLATO 系列

PLATO 系列：百度这两年针对 NLP 对话领域提出的一系列预训练的模型。

- PLATO
- PLATO-2
- PLATO-XL

从 PLATO 到 PLATO-XL，使用的数据越来越多，模型大小越来越大，但是在 PLATO-XL 中模型的结构实际上更加简单了。

模型结构

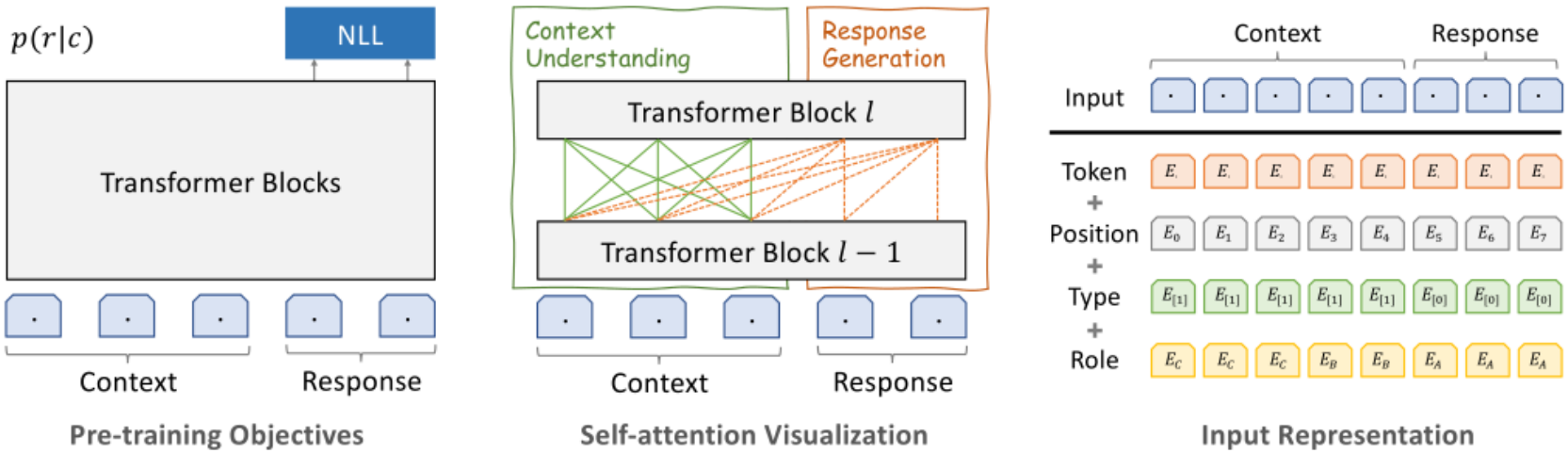


Figure 1: Network overview of PLATO-XL.

- 抛弃了 PLATO 的隐变量和 PLATO-2 的课程学习的方法，直接训练
- 沿袭 PLATO 中使用的 unified transformer
- 只保留了负对数似然（NLL）作为对话生成的预训练目标
- 训练时加入了多角色感知的预训练

为何采用 unified transformer

传统用 Transformer 的对话生成：使用 encoder-decoder 结构。
使用 unified transformer 的两方面好处：

1. 提升训练效率
   - 传统：对话样本长短不一
   - padding 补齐带来大量的无效计算
   - unified transformer：可以对输入样本进行有效的排序
2. 提高参数性价比
   - 可以**同时**进行对话理解和回复生成的联合建模
   - 灵活的注意力机制
   - 对 context 进行了双向编码
   - 对 response 进行了单向解码

多角色感知的预训练

目的：

- 改善对话模型有时候自相矛盾的问题
- 提升多轮对话上的一致性

产生矛盾的原因：

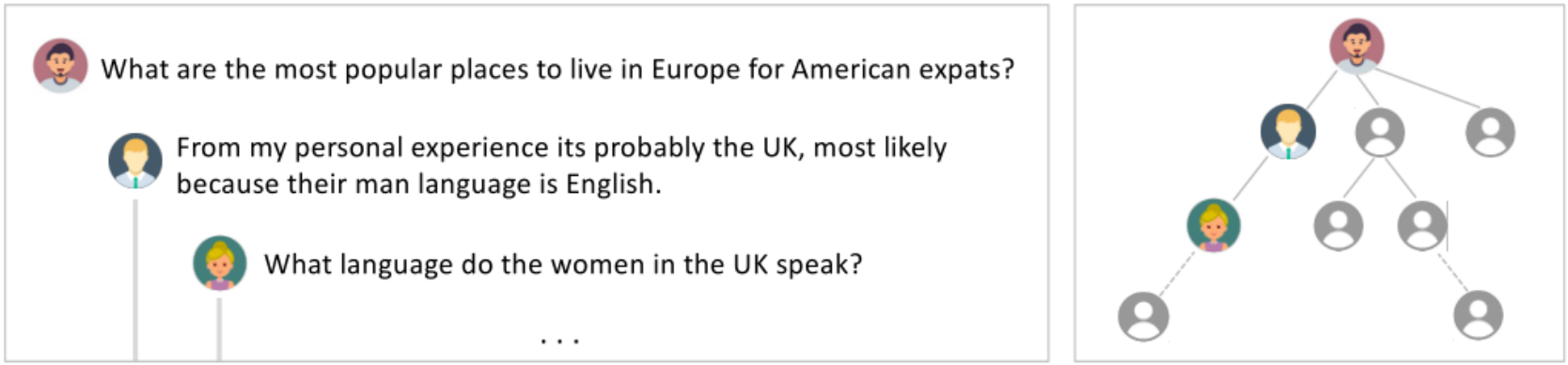


Figure 2: Left: one toy example to illustrate social media conversations. Right: corresponding message tree.

对话模型所用的预训练语料：社交媒体对话

- 特点：通常有多个用户参与
- 训练时模型较难区分对话上文中不同角度的观点和信息
- 容易产生一些自相矛盾的回复

实验结果

评估方法：采用了两个模型针对开放域进行相互对话（self-chat）的形式，然后再通过人工来评估效果。

- PLATO-XL 与 Facebook Blender、微软 DialoGPT、清华 EVA 模型、PLATO-2 相比，取得了更优异的效果
- PLATO-XL 也显著超越了目前主流的商用聊天机器人
- 除了开放域闲聊对话，模型也可以很好的支持知识型对话和任务型对话，在多种对话任务上效果全面领先
- 模型规模扩大对于效果提升也有显著作用，呈现较稳定的正相关关系

目录

Facebook Blender

这篇论文融合了 Facebook Blender 这个组近些年来在 open-domain chatbot 方向的诸多相关工作，读起来有些吃力。
论文中提出了三个模型：

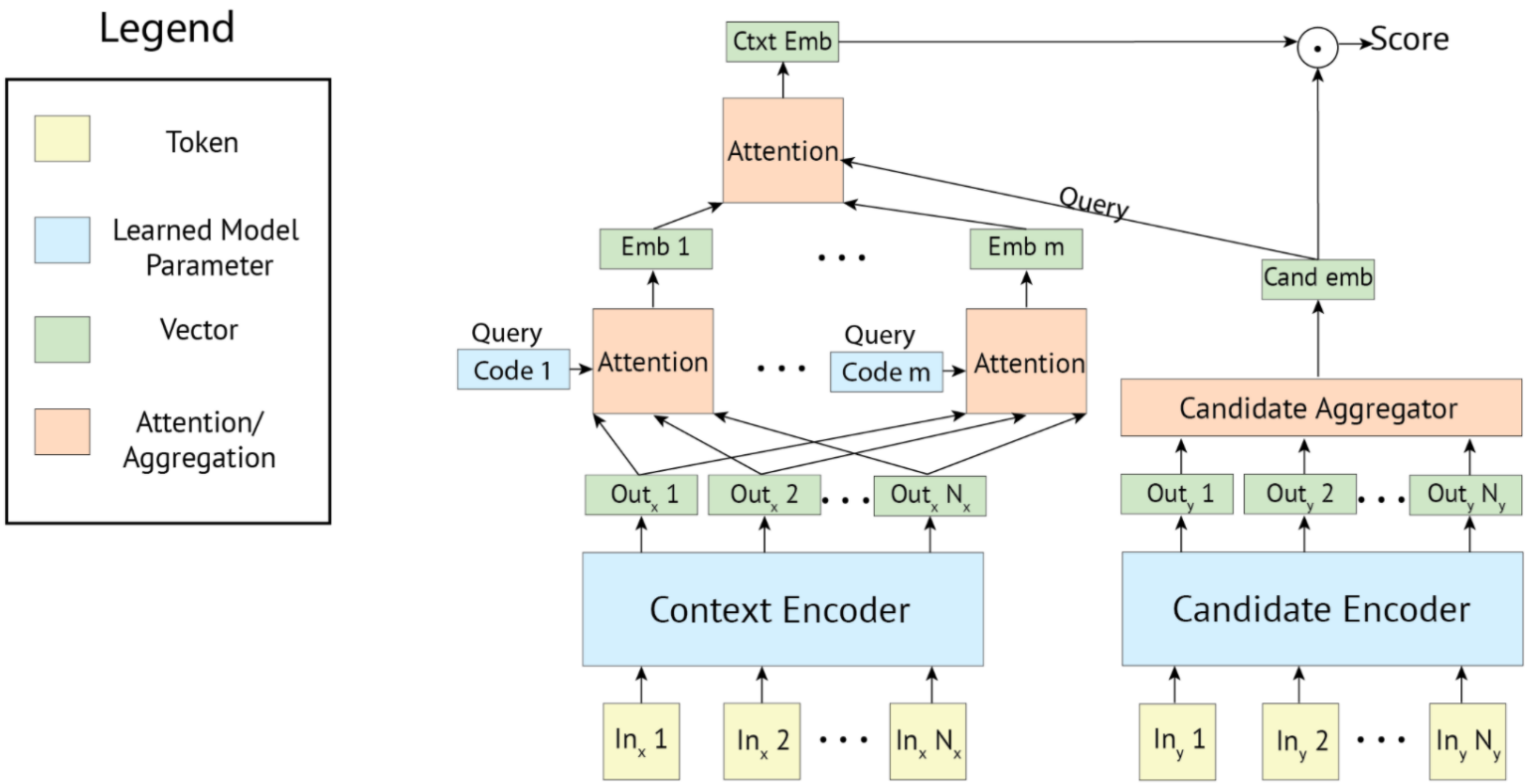
- 检索模型（Retriever）
- 生成模型（Generator）
- 检索 + 生成（Retrieve and Refine）

检索模型（Retriever）

从候选集中选取最合适的句子作为 bot 当前的答复。

- 训练时，候选集只有给定的一句 response
- 推断时，候选集由训练集中的所有 response 组成

打分 / 排序模型使用他们在之前的论文中提出的 Poly-encoder 模型。



实验表明：m 越大效果越好，当然模型打分也越耗时。

生成模型（Generator）

模型结构：

- 标准的 seq2seq 结构，只是用了标准的 Transformer 层
- encoder 层数少，decoder 层数多的设计

Blender 使用 beam search，但是加入了一些限制方法：

- 限制生成 response 的最小长度：
  - Minimum length
  - Predictive length
- 屏蔽重复的子序列

beam search 的这些限制方法实际上都是锦上添花，如果模型本身的质量不高，上述的方法作用也不大。

生成模型（Generator）

训练方法：

- Seq2seq 模型标准的训练方法 MLE
- Unlikelihood Loss

Unlikelihood Loss（UL）是作者在之前的论文中提出的一种损失函数，在提高正确的 token 概率的同时，降低其他 token 的概率。

UL 的关键

如何选取这些被打压的负 token。

- 作者选的是那些容易组合成常见 n-grams 的 tokens。
- 目的：期望降低生成无意义 response 的比例。

检索 + 生成（Retrieve and Refine）

Retrieve and Refine (RetNRef) 融合了检索和生成两种方法，是作者在 18 年的论文中提出的。

RetNRef

1. 先利用检索模型检索出一个结果
2. 把检索出的结果拼接到 context 后面，用一个特殊的分割符和 context 分隔
3. 整体作为 generator 模型的输入

目的：期望生成模型能学习到在合适的时候从检索结果中 copy 词或词组。

检索 + 生成（Retrieve and Refine）

两种检索方法：

Dialogue Retrieval 从训练数据中检索出得分最高的 response 作为结果

Knowledge Retrieval 从外部的大知识库如 Wiki 中检索

- 对于 Knowledge Retrieval，把检索出的结果直接追加到 context 后面，然后利用标准的 MLE 训练即可
- 对于 Dialogue Retrieval，直接利用 MLE 训练会有问题。训练出来的模型很容易直接忽略掉追加的检索部分

α -blending

训练时以 $\alpha\%$ 的概率把检索结果替换为实际 response。
这样模型就会被吸引去关注检索部分了。

数据集

数据集公认的标准：

- 对话要个性有趣
- 对话要包含知识
- 对话要富有同理心

作者在他之前的论文中发现：**在具有某些特性的数据上训练出的模型也会拥有这些特性。**

训练使用的数据集：

- http://pushshift.io Reddit
- ConvAI2
- Empathetic Dialogues (ED)
- Wizard of Wikipedia (WoW)
- Blended Skill Talk (BST)

模型训练流程：

1. 在 http://pushshift.io Reddit 上进行预训练
2. 在 ConvAI2、ED、WoW 上多任务精调
3. 在 BST 上精调

在优质数据上训练模型，也能降低模型产生不好的 response 的概率。

评估方法

自动评估：

- 生成模型采用 Perplexity(PPL)
- 检索打分模型采用 Hits@1/K

人工评估：

- ACUTE-Eval：
  1. 每次给两个对话 session
  2. 让人来评判哪个 speaker 聊的更好
  3. ACUTE-Eval 可以给出两个 speaker 各自的胜率
- Self-Chat ACUTE-Eval：和 ACUTE-Eval 的做法类似，只是评估时用的是自己跟自己聊的 session。

评估结果

PPL：

- 模型越大 PPL 越低
- RetNRef 相比于同等规模的生成模型，PPL 会略有提升

Self-Chat ACUTE-Eval：

- 相同尺寸，Generator 和 RetNRef 模型都未使用最小长度约束时：Retrieval 比 RetNRef 略好，二者都远远好于 Generative。
- Beam Search 中加入限制方法：
  - 限制最小生成长度效果显著，最小长度限制设为 20 效果最好
  - 子序列屏蔽有点用，但不显著
- 精调后的模型比只进行预训练的模型效果好很多
- Unlikelihood Loss 比 MLE 效果好一点点，但不显著

评估结果

ACUTE-Eval：

- 相同尺寸 Generator 和 RetNRef 模型使用的 beam search 都加入了最小长度 20 的限制：Generator 和 RetNRef 显著优于 Retrieval，RetNRef 略优于 Generative，但不显著。
- Blender 显著优于 Meena，胜率高达 70%
- Blender 最好的模型和人 PK 的胜率已经到了 49%，与人类的差距仅有 1%

模型存在的问题

- 倾向于使用高频词
- 倾向于生成重复信息
- 模型生成的答复可能前后冲突
- 其他问题：
  - 无法针对某个话题做深度对话，不会倾向于使用更多知识进行深聊
  - 模型无法深度理解，无法通过对话真正教会模型理解某个概念
  - 当前的评测针对的是 14 轮长度的对话，分析更长的对话肯定会发现其他问题

Blender 总结

- 大模型、大数据集对实验结果提升极其显著
- 训练数据的特性决定模型特性
- 评估方法使用 ACUTE-Eval、Self-Chat ACUTE-Eval
- 对 decoding 过程加入控制，比如控制生成句子长度，效果会更佳
- Blender 在轮次少的情况下已经接近人类的水平，但目前还有很多问题待解决