开域聊天机器人

吴金龙@爱因互动

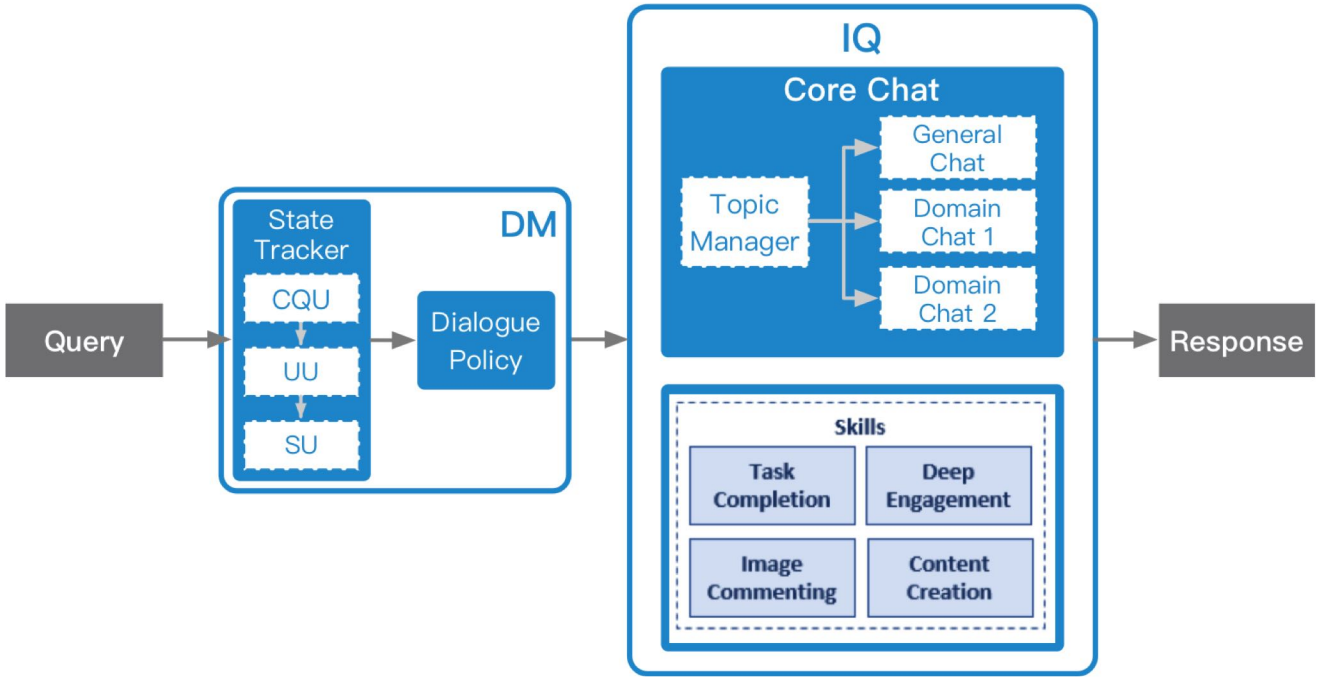
2020.06.15

提纲

- 复杂架构聊天机器人
  - 微软小冰
- 端到端聊天机器人
  - Google Meena
  - Facebook Blender

复杂架构聊天机器人

- 微软小冰



Google Meena

- 基于Evolved Transformer (ET) 的seq2seq模型
  - sample: *(context, response)*
  - encoder输入为context，由前几轮（最多7轮）对话拼接而成

Name	Total Params	Voc Size	L_enc	L_dec	d	h	Train Set Size	Steps	PPL
GTP-2	1.5B		-				40GB		
DialoGPT	762M						147M sessions		
Meena	2.6B	8K	1 ET (≈2)	13 ET (≈26)	2,560	32	341GB/ 867M samples	738k	10.2

Google Meena

- Decoder
  - **sample-and-rank**：抽样多次获得N个候选回复句子，然后从中选取概率最高的句子

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

    - *N = 20, T = 0.88*

Google Meena

- 评估方法
  - 自动评估：decoder上的Perplexity (**PPL**)

$$\begin{aligned}\mathrm{PPL}(S) &= P(w_1, \ldots, w_n)^{-\frac{1}{n}} \\ &= 2^{\log_2 \left[\prod_{i=1}^{n} p(w_i|w_0,\ldots,w_{i-1})\right]^{-\frac{1}{n}}} \\ &= 2^{-\frac{1}{n}\sum_{i=1}^{n} \log_2 p(w_i|w_0,\ldots,w_{i-1})}\end{aligned}$$

  - 人为评估：**S**ensibleness and **S**pecficity **A**verage (**SSA**)
    - Sensibleness：回复合理；符合逻辑、保持一致性
    - Specficity：回复具体，有内容

Google Meena

- 评估方法
  - **SSA**
    - Sensibleness：回复合理；符合逻辑、保持一致性
    - Specficity：回复具体，有内容

Bot	Sensibleness	Specficity	SSA
GenericBot	70%	0%	35%
DialoGPT	62%	39%	51%
Meena	87%	70%	79%
Human	97%	75%	86%

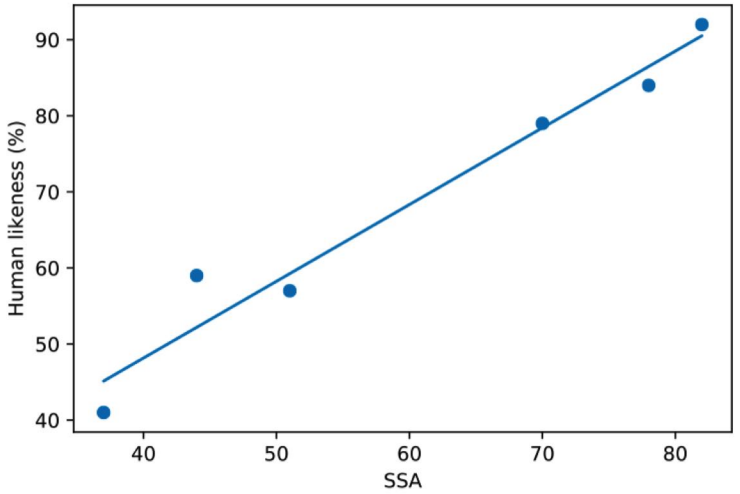


Figure 2: SSA vs human likeness. Each point is a different chatbot, except for the top right one, which is human. A regression line is plotted, for which the coefficient of determination (R^2) is 0.96.

Google Meena

- 评估方法
  - PPL vs. SSA
  - 模型PPL越低，效果越好

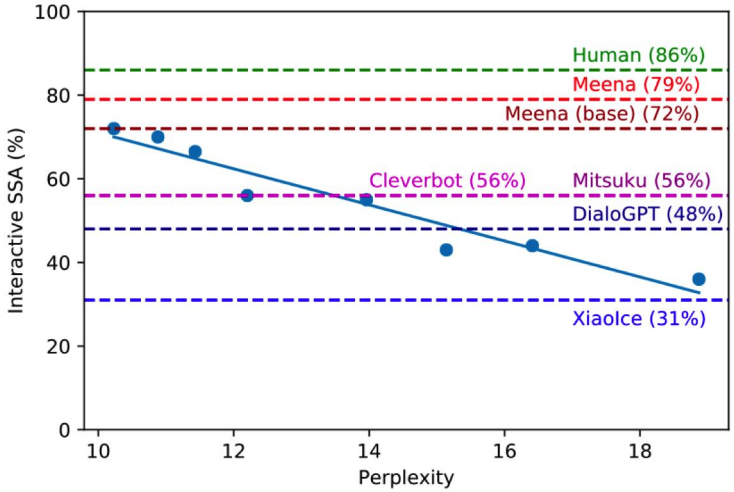


Figure 1: Interactive SSA vs Perplexity. Each point is a different version of the Meena model. A regression line is plotted, for which the coefficient of determination (R^2) is 0.93, an indication of strong correlation between perplexity and the human evaluation metric (SSA).

Google Meena

- 示例

Human:	*Hi!*
Meena:	*Hi! How are you this fine evening?*
Human:	*I'm doing well. it's morning my time!*
Meena:	*Good morning! What are your plans for today?*
Human:	*Get my work done then watch a movie at home in the night*
Meena:	*Sounds fun! What movie?*
Human:	*I haven't decided yet, perhaps an Oscar-nominated one. any suggestion?*
Meena:	*I'm not too familiar with recent Oscar winners, but I do know I wanted to see The Grand Budapest Hotel.*
Human:	*ah, tell me about that movie!*
Meena:	*It's from the same director and writer as Moonrise Kingdom. Wes Anderson is pretty well known for making funny, colorful films. I've heard it's pretty good.*
Human:	*ha! I've searched on the internet and both movies are very colorful!*
Meena:	*I'm looking forward to it. I love colorful movies.*

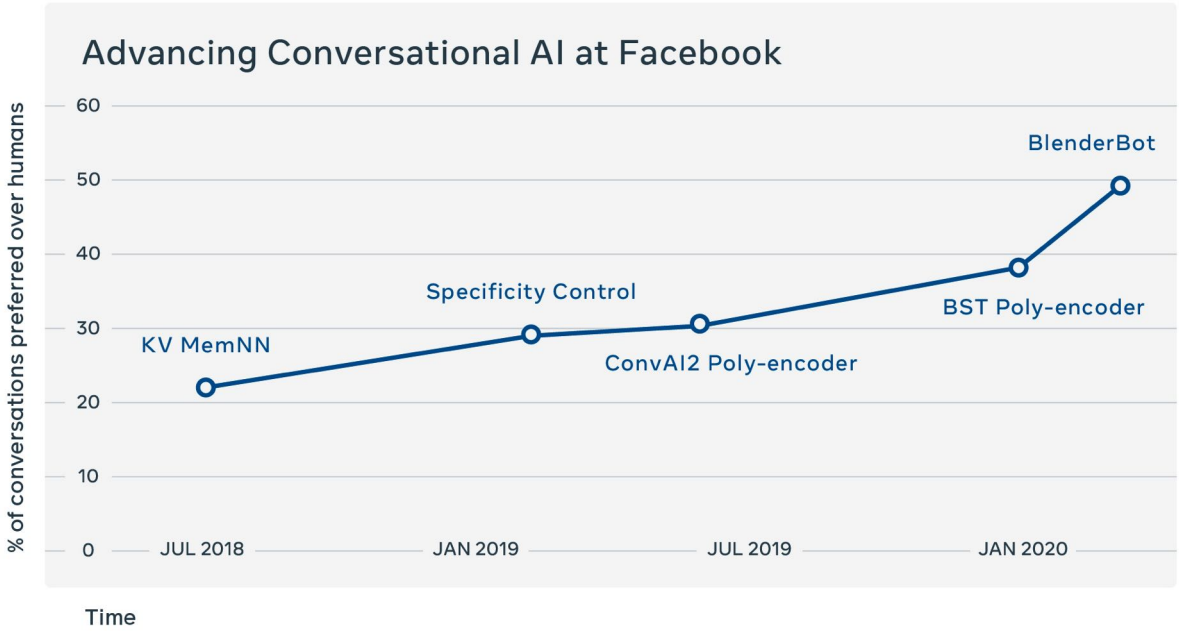
Human:	*I like how cute cows are*
Meena:	*Ah, interesting. You like how they are cute, right?*
Human:	*yeah, and they're realy smart*
Meena:	*I heard they go to college*
Human:	*cows go to college?*
Meena:	*I heard that a cow went to Harvard.*
Human:	*what did the cow study?*
Meena:	*Bovine sciences.*
Human:	*do horses go to Harvard?*
Meena:	*Horses go to Hayvard.*
Human:	*that's a pretty good joke, I feel like you led me into it*
Meena:	*You were trying to steer it elsewhere, I can see it.*

Google Meena

- 三个贡献
  - 提出了评估多轮对话效果的指标SSL
  - PPL和SSL高度负相关，所以可用PPL自动评估模型效果
  - 足够大的端到端的模型可以打败复杂架构的对话系统

Facebook Blender

- 本文是前面一系列工作的融合

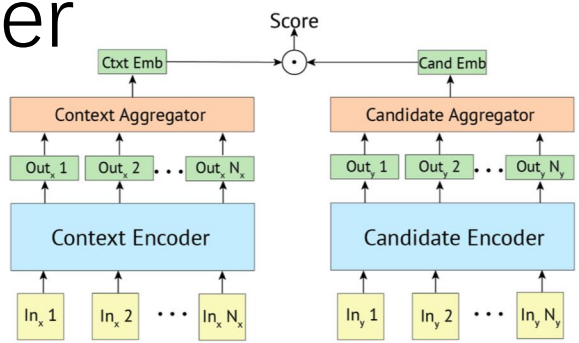


Facebook Blender

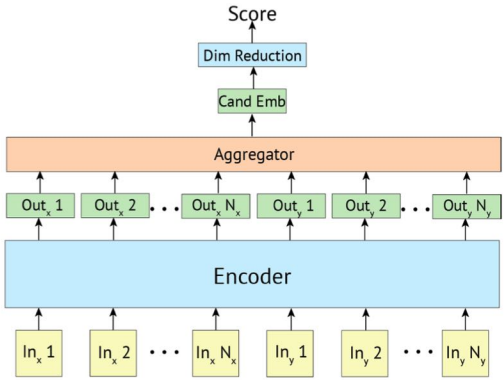
- 产生答复的三个模型
  - 检索：Retriever
  - 生成：Generator
  - [检索; 生成]：Retrieve and Refine
- 数据集
- 效果评估
- 待解决问题

Facebook Blender

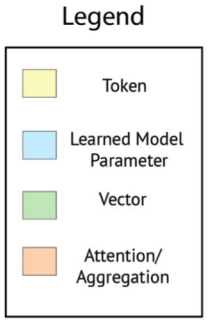
- **Retriever: Poly-encoder**
  - Context一般更长，使用 m 个向量表达
  - 降低计算量的同时，精度可以接近Cross-encoder

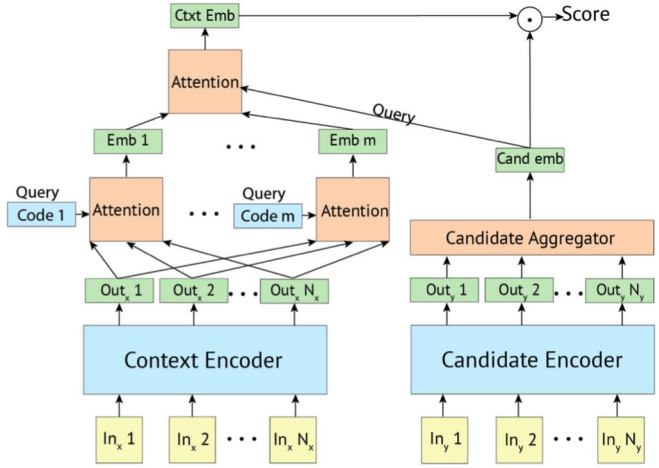


(a) Bi-encoder



(b) Cross-encoder

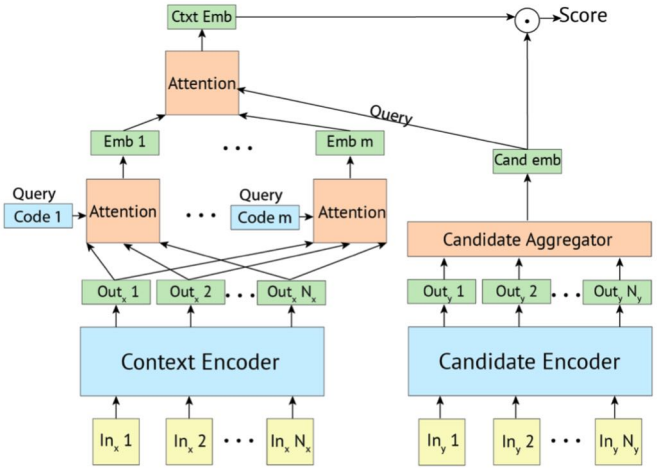




(c) Poly-encoder

Facebook Blender

- **Retriever: Poly-encoder**



(c) Poly-encoder

Dataset	ConvAI2	DSTC 7		Ubuntu v2		Wikipedia IR
split	test	test		test		test
metric	R@1/20	R@1/100	MRR	R@1/10	MRR	R@1/10001
pre-trained BERT weights from (Devlin et al., 2019) - Toronto Books + Wikipedia						
Bi-encoder	81.7 ± 0.2	66.8 ± 0.7	74.6 ± 0.5	80.6 ± 0.4	88.0 ± 0.3	-
Poly-encoder 16	83.2 ± 0.1	67.8 ± 0.3	75.1 ± 0.2	81.2 ± 0.2	88.3 ± 0.1	-
Poly-encoder 64	83.7 ± 0.2	67.0 ± 0.9	74.7 ± 0.6	81.3 ± 0.2	88.4 ± 0.1	-
Poly-encoder 360	83.7 ± 0.2	68.9 ± 0.4	76.2 ± 0.2	80.9 ± 0.0	88.1 ± 0.1	-
Cross-encoder	84.8 ± 0.3	67.4 ± 0.7	75.6 ± 0.4	82.8 ± 0.3	89.4 ± 0.2	-
Our pre-training on Toronto Books + Wikipedia						
Bi-encoder	82.0 ± 0.1	64.5 ± 0.5	72.6 ± 0.4	80.8 ± 0.5	88.2 ± 0.4	-
Poly-encoder 16	82.7 ± 0.1	65.3 ± 0.9	73.2 ± 0.7	83.4 ± 0.2	89.9 ± 0.1	-
Poly-encoder 64	83.3 ± 0.1	65.8 ± 0.7	73.5 ± 0.5	83.4 ± 0.1	89.9 ± 0.0	-
Poly-encoder 360	83.8 ± 0.1	65.8 ± 0.7	73.6 ± 0.6	83.7 ± 0.0	90.1 ± 0.0	-
Cross-encoder	84.9 ± 0.3	65.3 ± 1.0	73.8 ± 0.6	83.1 ± 0.7	89.7 ± 0.5	-
Our pre-training on Reddit						
Bi-encoder	84.8 ± 0.1	70.9 ± 0.5	78.1 ± 0.3	83.6 ± 0.7	90.1 ± 0.4	71.0
Poly-encoder 16	86.3 ± 0.3	71.6 ± 0.6	78.4 ± 0.4	86.0 ± 0.1	91.5 ± 0.1	71.5
Poly-encoder 64	86.5 ± 0.2	71.2 ± 0.8	78.2 ± 0.7	85.9 ± 0.1	91.5 ± 0.1	71.3
Poly-encoder 360	86.8 ± 0.1	71.4 ± 1.0	78.3 ± 0.7	85.9 ± 0.1	91.5 ± 0.0	**71.8**
Cross-encoder	**87.9 ± 0.2**	**71.7 ± 0.3**	**79.0 ± 0.2**	**86.5 ± 0.1**	**91.9 ± 0.0**	-

Facebook Blender

- **Retriever: Poly-encoder**
  - 训练方法
    - 最小化 InfoNCE，负样本来自同一个batch下随机选取的其他样本

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$

  - 推断
    - 候选回复是训练集中的回复

Facebook Blender

- **Generator: 基于transformer架构的seq2seq模型**
  - decoder层数比encoder大很多

Name	Total Params	V	L_{enc}	L_{dec}	d	h	Steps	PPL
90M	87,508,992	55K	8	8	512	16	2.86M	25.6
2.7B	2,696,268,800	8K	2	24	2560	32	200K	13.3
9.4B	9,431,810,048	8K	4	32	4096	32	200K	12.2

Table 2: **Perplexity on the validation set of pushshift.io Reddit** for several generative Transformer models with given architecture settings. Note that perplexity is not directly comparable between the 90M models and the larger models as the 90M models use a different dictionary. Columns include the vocabulary size (V), number of encoder and decoder layers (L_{enc}, L_{dec}), embedding dimensionality (d), Multihead Attention Heads (h), and training steps.

Facebook Blender

- **Generator: 基于transformer架构的seq2seq模型**
  - decoder层数比encoder大很多

Name	Total Params	V	L_{enc}	L_{dec}	d	h	Steps	PPL
90M	87,508,992	55K	8	8	512	16	2.86M	25.6
2.7B	2,696,268,800	8K	2	24	2560	32	200K	13.3
9.4B	9,431,810,048	8K	4	32	4096	32	200K	12.2

对标Meena

Name	Total Params	Voc Size	L_enc	L_dec	d	h	Train Set Size	Steps	PPL
Meena	2.6B	8K	1 ET (≈2)	13 ET (≈26)	2,560	32	61B BPE tokens	738k	10.2

Facebook Blender

- **Generator: 基于transformer架构的seq2seq模型**
  - 训练方法
    - 方法一：**MLE**
    - 方法二： **Unlikelihood Loss**，惩罚其他token

$$\mathcal{L}_{\mathrm{UL}}^{(i)}(p_\theta, \mathcal{C}_{1:T}, \mathbf{x}, \mathbf{y}) = -\sum_{t=1}^{|y|} \sum_{y_c \in \mathcal{C}_t} \log\left(1 - p_\theta(y_c | \mathbf{x}, y_{<t})\right)$$

$$\mathcal{L}_{\mathrm{ULE}}^{(i)} = \mathcal{L}_{\mathrm{MLE}}^{(i)} + \alpha \mathcal{L}_{\mathrm{UL}}^{(i)}$$

      - 如果一个token组成的n-grams比真实答案中n-grams比例高，负样本优先选取这样的token
      - 期望生成无意义回复的比例降低

Facebook Blender

- **Generator: 基于transformer架构的seq2seq模型**
  - 推断解码
    - **Sampling**
      - **top-k sampling**：每个时间步从top-k个候选词中按概率选取一个词
      - **sample-and-rank**：抽样多次获得多个候选回复句子，然后从中选取概率最高的结果
    - **Beam Search，**控制回复长度
      - **Minimum length**：要求回复长度必须大于设定的值
        - 长度不达标时，强制不产生结束token
      - **Predictive length**：利用四分类模型预测回复长度
        - < 10, < 20, < 30, or > 30 tokens
        - 模型：Poly-encoder
      - **Subsequence Blocking**：不允许产生当前句子和前面对话中已经存在的 3-grams

Facebook Blender

- **Retrieve and Refine**
  - 融合了检索和生成步骤
    1. 先利用检索模型检索出一个结果
    2. 然后把检索出的结果拼接到Context后面，作为generator的输入

  - 期望生成模型能学习到在合适的时候从检索结果中copy词

Facebook Blender

- **Retrieve and Refine**
  - 如何利用检索模型检索出一个结果？
    - **Dialogue Retrieval**
      - 检索出得分最高的候选回复
    - **Knowledge Retrieval**
      - 从大知识库中检索，如WoW（Wiki）
        - 分别利用当前对话topic和最后两轮对话，各自检索出 top-7 相关文章
        - 把21篇文章各自分句，然后把自己文章的title追加到每个句子最前面，获得很多候选句子

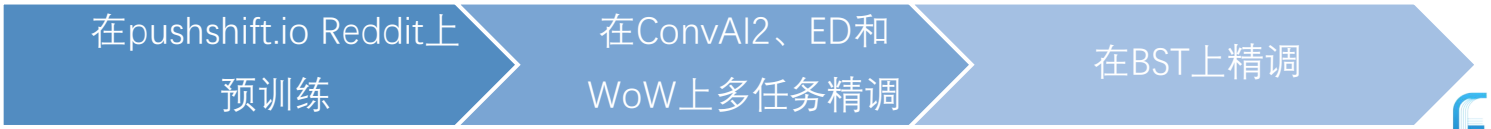
        - 再利用Poly-encoder架构的模型对候选句子排序，最终使用 top-1 的句子
        - 还会训练一个单独的分类器来判断是否需要从知识库中检索知识。有些对话context不需要额外知识

Facebook Blender

- **Retrieve and Refine**
  - 训练方法：
    - **Dialogue Retrieval训练：α-blending**
      - 把检索出的结果直接追加到context后面，然后利用MLE训练
        - 问题：训练出来的模型很容易直接忽略掉检索出的结果，因为实际回复与检索结果关联性可能不强
      - **α-blending：以 α% 的概率把检索结果替换为实际回复**
    - **Knowledge Retrieval训练：MLE**
      - 上面的情形不明显

Facebook Blender

- 好的对话应该：个性有趣、包含知识、富有同理心
- 数据集
  - **pushshift.io Reddit**：整理自Reddit网站上的讨论；数据量大
    - 训练预训练模型（检索模型：MLM、生成模型：LM）
  - **ConvAI2**：带个性的对话数据，对话目标是了解对方；对话个性有趣
  - **Empathetic Dialogues (ED)** ：一个人发牢骚另一个人倾听；数据富有同理心
  - **Wizard of Wikipedia (WoW)**：基于wiki 某些topic的对话；包含知识
  - **Blended Skill Talk (BST)**：基于ConvAI2、ED和WoW构建，并融合它们的优势
- 模型训练的流程



Facebook Blender

- **BST 对话示例**
  - 提供角色背景、对话主题和对话历史

Persona for Unguided Speaker:	**Persona for Guided Speaker**:
My son plays on the local football team.	My eyes are green.
I design video games for a living.	I wear glasses that are cateye.

Wizard of Wikipedia topic: Video game design
Previous utterances (shown to speakers):
U: What video games do you like to play?
G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget

Actual utterances:
U: Exactly! I think many people fail to notice how beautiful the art of video games can be. **(PB)**
(G selected the WoW suggestion: "Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.")
G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. **(K)**
U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids **(K)**
G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. **(S)**
U: I use games to relax after a stressful day, the small escape is relaxing. **(PB)**
(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")
G: I enjoy a good gaming session after a hard day at work as well. **(PB)**
U: What other hobbies does your son have? **(PB)**
G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. **(PB)**
U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. **(PB)**
G: I wish I could play football, But I wear this cateye glasses and they would break if I tried. **(PB)**
U: Sounds nice. Are they new or vintage? **(E)**
G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. **(S)**

Figure 1: Sample conversation from the BlendedSkillTalk dataset, annotated with four conversation mode types (PB: personal background; K: knowledge; S: personal situation; E: empathy). The guided (G) and unguided (U) workers are given personas and a topic. The conversation has been seeded with two utterances from a conversation sampled from WoW. When the guided worker selected one of the suggestions, it is shown in shaded grey.

Facebook Blender

- 效果评估方法
  - **hits@1/K, PPL**
  - **ACUTE-Eval**
    - 给两个对话session，让人来评估哪个好（更想跟谁聊；谁更像人）
  - **Self-Chat ACUTE-Eval**
    - 让模型自己与自己聊，然后把对话session交由人来评估

Facebook Blender

- 效果评估方法
  - PPL

Name	Total Params	V	L_{enc}	L_{dec}	d	h	Steps	PPL
90M	87,508,992	55K	8	8	512	16	2.86M	25.6
2.7B	2,696,268,800	8K	2	24	2560	32	200K	13.3
9.4B	9,431,810,048	8K	4	32	4096	32	200K	12.2

Table 2: **Perplexity on the validation set of pushshift.io Reddit** for several generative Transformer models with given architecture settings. Note that perplexity is not directly comparable between the 90M models and the larger models as the 90M models use a different dictionary. Columns include the vocabulary size (V), number of encoder and decoder layers (L_{enc}, L_{dec}), embedding dimensionality (d), Multihead Attention Heads (h), and training steps.

Model	Size	ConvAI2	WoW	ED	BST	Avg.
pushshift.io Reddit Generative	90M	18.33	31.18	14.44	18.09	20.51
BST Generative	90M	11.36	17.56	11.48	14.65	13.76
BST RetNRef	256M/90M	11.79	18.37	11.87	14.62	14.16
pushshift.io Reddit Generative	2.7B	15.70	13.73	11.06	14.36	13.71
BST Generative	2.7B	8.74	8.78	8.32	10.08	8.98
BST RetNRef	622M/2.7B	9.31	9.28	9.93	10.59	9.78
pushshift.io Reddit Generative	9.4B	15.02	12.88	10.41	13.5	12.95
BST Generative	9.4B	8.36	8.61	7.81	9.57	8.59

Table 3: **Perplexity of the pre-trained and fine-tuned models on the validation set for BST datasets.** Note that perplexity is not directly comparable between the 90M models and the larger models as 90M models use a different dictionary. Fine-tuning gives gains for each skill (task) compared to pre-training on pushshift.io Reddit alone.

- RetNRef的PPL会比Generative略高

Facebook Blender

- 效果评估
  - 训练数据决定模型特性
    - **想让模型回复有趣，用ConvAI2训练**
    - **想让模型回复有同理心，用ED训练**
    - **想让模型回复包含知识，用WoW训练**
    - **想让模型回复同时有多个特性，同时用ConvAI2、ED和WoW训练，并在BST上精调**
    - 在优质的数据上训练，也能降低模型回复出现污词的概率

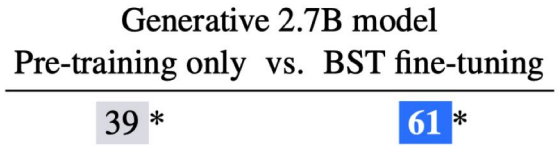


Figure 11: Self-Chat ACUTE-Eval (engagingness) shows a significant gain ($p < 0.05$) for fine-tuning on the BST Tasks.

在pushshift.io Reddit上预训练 → 在ConvAI2、ED和WoW上多任务精调 → 在BST上精调

Facebook Blender

- 效果评估（Self-Chat ACUTE-Eval）
  - beam search中加入控制
    - **控制长度很有用**

Generative 2.7B model: Min Beam Length

	Constrained	vs.	Unconst.
Min. Length 5	52		48
Min. Length 10	68 **		32 **
Min. Length 20	83 **		17 **
Min. Length 40	82 **		18 **
Predictive (5,10,15,20)	69 **		31 **
Predictive (10,20,30,40)	81 **		19 **

Figure 7: Self-Chat ACUTE-Eval (engagingness) shows controlling minimum beam length gives large gains in engagingness compared to not controlling it, according to humans, with 20 being best. All rows are significant ($p < 0.01$) except the first.

    - **子序列屏蔽有点用**

Generative 2.7B model: Beam Blocking

	Block	vs.	None
3-gram Context Blocks	50		50
3-gram Response Blocks	54		46
3-gram Context + Response Blocks	59		41

Figure 8: Self-Chat ACUTE-Eval (engagingness): comparing beam-blocking variants. Blocking both context and response 3-grams during generation gives highest scores, however, none of these results are significant.

Facebook Blender

- 效果评估（Self-Chat ACUTE-Eval）
  - Self-Chat ACUTE-Eval的其他结论
    - **模型越大效果越好**
    - **beam search中的beam值取10效果比1、30都好**
    - **Unlikelihood Loss比MLE效果好点**

Generative BST 2.7B model	
MLE	vs. Unlikelihood
46	**54**

Figure 13: Self-Chat ACUTE-Eval (engagingness) MLE vs. Unlikelihood training (penalizing overexpressed n-grams). The result is not statistically significant (165 trials).

Facebook Blender

- 效果评估（ACUTE-Eval）
  - Retrieval vs. Generator vs. RetNRef
    - **RetNRef >= Generator > Retrieval**

Win %	Loss % Ret	Loss % Gen	Loss % RetNRef
Retrieval		29 *	30 *
Generative	71 *		44
RetNRef	70 *	56	

Figure 14: Human-bot ACUTE-Eval (engagingness): Retrieve and Refine($\alpha = 0.5$) and Generative (90M, beam search decoding, min beam size 20) beat Retrieval (256M). All results are significant ($p < 0.01$) except for RetNRef vs. Generative.

Facebook Blender

- 效果评估（ACUTE-Eval）
  - Blender vs. Meena
    - **Blender > Meena**

	Ours	vs.	Meena
BST Generative (2.7B) std. beam	46		54
BST RetNRef (256M/90M)	49		51
pushshift.io Reddit Generative (2.7B)	56		44
BST Generative (90M)	59		41
Wiz Generative (2.7B)	59 *		41 *
BST RetNRef (622M/2.7B)	65 **		35 **
BST Generative (2.7B)	65 **		35 **
BST Generative (9.4B)	66 **		34 **
BST Unlikelihood (2.7B)	70 **		30 **

Figure 16: Human-Chat ACUTE-Eval of **humanness**, various models compared to Meena. Our best models are considered more humanlike than Meena, rows with * and ** are statistically significant.

	Ours	vs.	Meena
BST Generative (2.7B) std. beam	50		50
pushshift.io Reddit Generative (2.7B)	53		47
BST RetNRef (256M/90M)	60 *		40 *
BST Generative* (90M)	61 *		39 *
Wiz Generative (2.7B)	61 **		39 **
BST Unlikelihood (2.7B)	64 **		36 **
BST Generative (9.4B)	67 **		33 **
BST RetNRef (622M/2.7B)	70 **		30 **
BST Generative (2.7B)	75 **		25 **

Figure 15: Human-Chat ACUTE-Eval of **engaging-ness**, various models compared to Meena. Our best models are considered more engaging than Meena, rows with * ($p < 0.05$) and ** ($p < 0.01$) are statistically significant. Larger generative models with BST fine-tuning and length-controlled decoding work best.

Facebook Blender

- 优秀示例

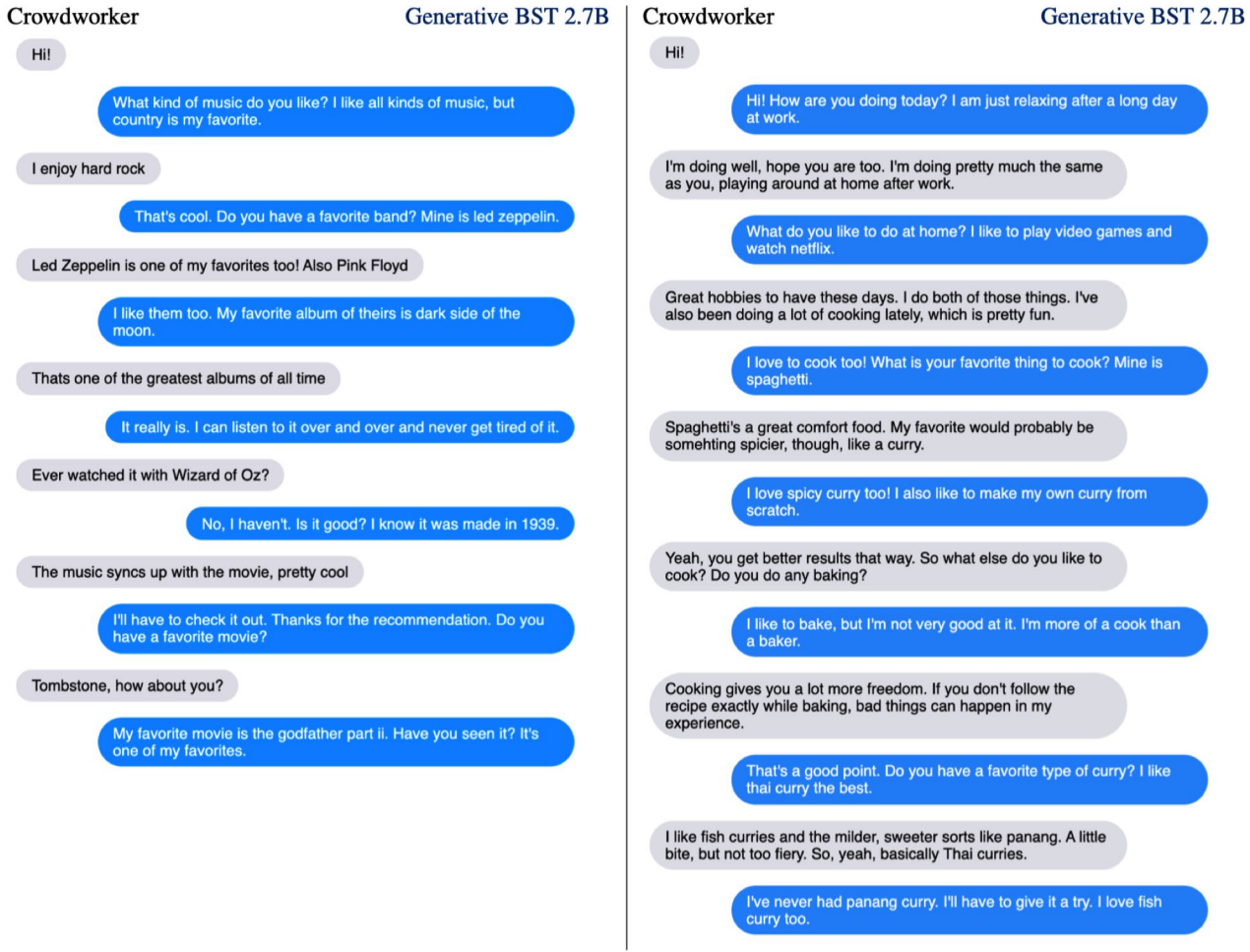


Figure 19: **Cherry-picked crowdworker examples.** Two conversations between different crowdworkers (left speakers) and the Generative BST 2.7B model (right speakers).

Facebook Blender

- 效果评估（ACUTE-Eval）
  - Model vs. Human
    - **Model <= Human**
  - **其实模型还有很多问题！**

	Model	vs.	Human
Meena (Adiwardana et al., 2020)	28 **		72 **
BST Generative (2.7B) std. beam	21 **		79 **
pushshift.io Reddit Generative (2.7B)	36 **		64 **
BST RetNRef (256M/90M)	37 **		63 **
BST Generative (90M)	42		58
BST Generative (9.4B)	45		55
BST RetNRef (622M/2.7B)	46		54
Wiz Generative (2.7B)	47		53
BST Unlikelihood (2.7B)	48		52
BST Generative (2.7B)	49		51

Figure 17: ACUTE-Eval of engagingness of models vs. humans by comparing human-bot logs to human-human logs. Rows with ** are statistically significant.

Facebook Blender

- 效果评估
  - **其实模型还有很多问题！**

- 倾向于使用高频词，产生高频 n-grams，很少使用低频词
  - 鼓励模型产生更长的回复会缓解这个问题，但无法杜绝
  - 使用unlikelihood loss也可以缓解此问题，但从ACUTE-Evals看可能带来副作用

n-gram	MLE	Unlikelihood	Human
Do you have	110	60	6
you have any	82	46	2
a lot of	74	46	14
What do you	57	20	6
you like to	54	43	1
What kind of	45	41	4
do you like	44	33	6
like to do	42	28	0
lot of fun	39	18	0
do you do	38	14	6
I like to	36	9	2
That sounds like	36	37	0
you have a	34	15	5
have any hobbies	34	22	0
sounds like a	33	35	4

Figure 22: Counts of most common 3-grams from the BST Generative 2.7B model (likelihood) from the conversation logs when talking to crowdworkers, compared to those of the same model trained with unlikelihood, and to human logs (for the same number of utterances).

Facebook Blender

- 效果评估
  - **其实模型还有很多问题！**

- 倾向于拷贝信息，产生各种重复
  - 比如前面用户说了喜欢狗，模型可能说自己也喜欢
  - 子序列屏蔽的方法可以屏蔽简单的n-grams重复，但复杂的重复很难规避
  - 可以考虑训练时使用unlikelihood loss，最小化context重复
  - 为机器人添加角色背景可能也是一种解决方法

Facebook Blender

- 效果评估
  - **其实模型还有很多问题！**

- 内容冲突和遗忘（原因可能不是模型忘了，而是模型不会推理）
  - 前面说喜欢狗，后面说不喜欢狗
  - 用户已经说过喜欢狗，可机器人还问喜不喜欢狗
  - 子序列屏蔽的方法可以屏蔽简单的n-grams重复，但复杂的重复很难规避
  - 可以考虑训练时使用unlikelihood loss，最小化背景重复
  - 为机器人添加角色背景可能也是一种解决方法

Facebook Blender

- 效果评估
  - **其实模型还有很多问题！**

- 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

- 当前的评测针对的是14轮长度的对话，更长的对话肯定会存在其他问题

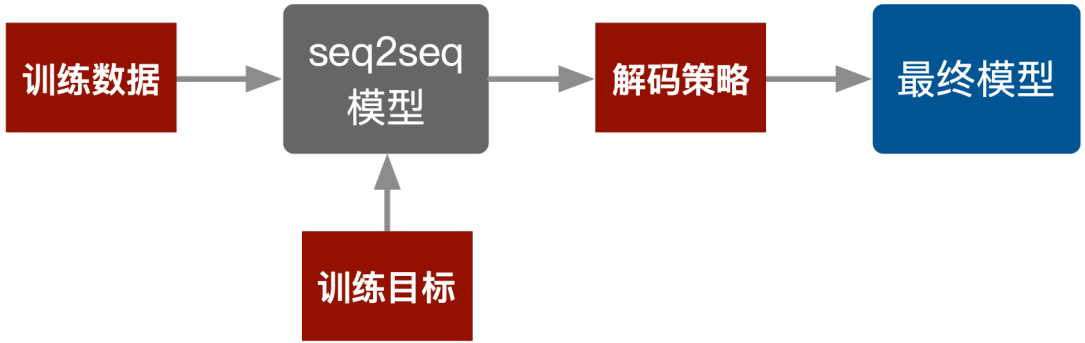
- 模型无法深度理解，无法通过对话教会模型真正理解某个概念

Facebook Blender

- 总结
  - **训练数据决定模型特性**：BST
  - 评估方法： ACUTE-Eval、 Self-Chat ACUTE-Eval
  - Larger Transformer is all you need
  - 对decoding过程加入控制，比如控制生成句子长度，效果会更佳
  - 大的端到端模型在轮次少（<=28 轮）的情况下有希望能达到或接近人类的水平，但目前还有很多问题待解决；轮次多的情况下一切都还未知

总结

- 端到端模型真的能打败复杂架构的系统！
- 端到端模型是开域聊天机器人的未来，但是否未来已来？
- 如何让端到端模型的表现更可控？



- 机器人能否胜任更长轮次的对话？

NLP/Chatbot

Thanks !

Your business is in good hands .

References

- 微软小冰对话机器人架构

- Towards a human-like open-domain chatbot
- 读后感：Towards a human-like open-domain chatbot
- 能跟你聊DOTA的神经对话模型：Meena&DialoGPT

- Recipes for building an open-domain chatbot
- Recipes for building an open-domain chatbot - dair.ai
- A state-of-the-art open source chatbot