基于预训练的个性化对话

作者：张荣升、毛晓曦、席亚东

网易伏羲实验室

摘要

赋予机器人特定的个性是人机对话系统里至关重要的方向。但是由于大规模个性化对话语料的缺失，以及如何在自然语言中使用个性化特征是困难的，所以这个问题的研究还有很多需要深入和改进的地方。在这篇文章中，我们提出了一种，给定个性化特征和大规模多轮对话语料，基于预训练的方法生成个性化回复的对话系统。这个预训练的对话系统分为三个阶段，分别为预训练语言模型初始化参数、利用大规模对话语料预训练对话模型、利用个性化的对话语料微调对话模型。通过这种方式，我们最后在 SMP2019 中文人机对话技术评测任务二中，自动指标评测获得第二名（BLEU: 0.0061，Perplexity: 292.67, Distinct: 0.2160），并在最终人工评测中获得第一名（流畅性：0.742，个性化：0.180，相关性：0.544）。

关键词：

人机对话，个性化对话，预训练

1.引言

构建类人的对话系统是目前人工智能的重要方向，通过赋予机器人不同的个性化特征，如年龄，性别，地理位置，兴趣爱好等，可以极大的提高对话中的用户参与感。这种风格化的对话系统可以和用户更自然的交流并能够更容易的获得用户的信任。

目前的建立个性化对话机器人的方法可以分为两类。第一类是个性化特征没有给定，在这些场景下，个性化的特征隐式的表现在对话语料中[1,2]，个性特征使用一个向量编码加入到模型里。在这些模型里，由于所有个性化的用户信息用

一个向量表示，我们通常不知道哪种个性化的信息被抓住了。第二类是明确的个性化特征给定了。这些特性可能是结构化的数据[3]，也可能是非结构化的描述[4，5，6]。但是这些模型受限于人工标注的数据或者是众包创作的对话，因此不能广泛用于现实中的大规模对话数据集。

近年来，基于大规模语料预训练的方法（如 Elmo,GPT,BERT,GPT2,XLNet）在各类 NLP 的任务上都取得了较大的成就[7,8,9,10]。通过使用预训练的模型同样在个性化的对话生成模型上取得了初步的成就，例如在小的个性化数据集上微调预训练的模型[4]。但是这种对话数据搜集的时候是为了使对话包含更多个性化的数据，但是这样的设定与实际情况种的对话场景并不符合。实际情况中的对话语料是个性稀疏的，对话双方没有在对话中刻意的去使用个性化的特征，这与众包标注的个性化对话数据有本质的区别。

SMP2019ECDT 任务二个性化对话的数据是大规模的真实对话场景语料，并且提供了对话的参与者结构化的个性标签。但是这些实际的大规模语料里可能之友少部分存在与个性化相关对话。针对个性稀疏的问题，我们提出了一种有效的基于预训练的方法，从大规模个性稀疏的对话语料中去更多的生成个性化的回复。我们的预训练方法主要有三步：

1. 利用大规模的中文语料预训练 GPT 语言模型。并用预训练的语言模型的参数初始化 seq2seq 的 encoder 和 decoder 参数。
2. 利用所有官方提供的个性稀疏对话语料在第一步的基础上，训练 seq2seq 模型，使模型学习关于对话的特性。
3. 利用一定的策略从所有对话语料里提取出小规模的与个性化相关的多轮对话。再在小规模的对话语料上训练 seq2seq 模型，使模型更容易学到对话与个性特征之间的关系。

2.模型及方法介绍

该部分从数据集的处理、模型结构和训练细节上阐述我们的系统。

2.1 数据集

2.1.1 数据集样例

如图 1，数据集中的每个 session 为一组多轮对话，并且对每个参与者给定了其对应的个性化特征，包括性别，地域和兴趣标签。数据集一共有 543 万多轮对话，每一组对话可能会涉及到部分个性化的信息，也可能不包含个性化的信息。

```
----------------------------------------
s1 半斤白酒已下肚
s2 能赏我口饭吃吗
s1 好啊，问题是..... 你过得来吗
s2 你可以开灰机来接我吗
s1 白机中不？
s2 是飞机吗
s1 那也得等姐有money 了
s2 蚂蚁老多了
s1 弟弟，咱俩私聊去
s2 不，弟弟没空，忙着呢
s1:(性别：女，年龄：80后，地域：'河南 商丘'，兴趣标签：'育儿百科')
s2:(性别：男，年龄：00后，地域：'上海 黄浦区'，兴趣标签：'快乐大本营;开朗;旅游;娱乐')
----------------------------------------
```

图(1): 数据集示例

2.1.2 数据集预处理

为了适应模型，我们分别处理对话的历史和结构化的个性化数据。

对对话历史我们直接进行拼接，并用分隔符间隔。例如“*你想表达什么，我听着<p>单纯的字面上的意思，just 想而已。木有要做<p>你存心找死。让我这种单身的情何以堪<p>找个男朋友吧，你也单身够了*”，其中<p>为分隔符。

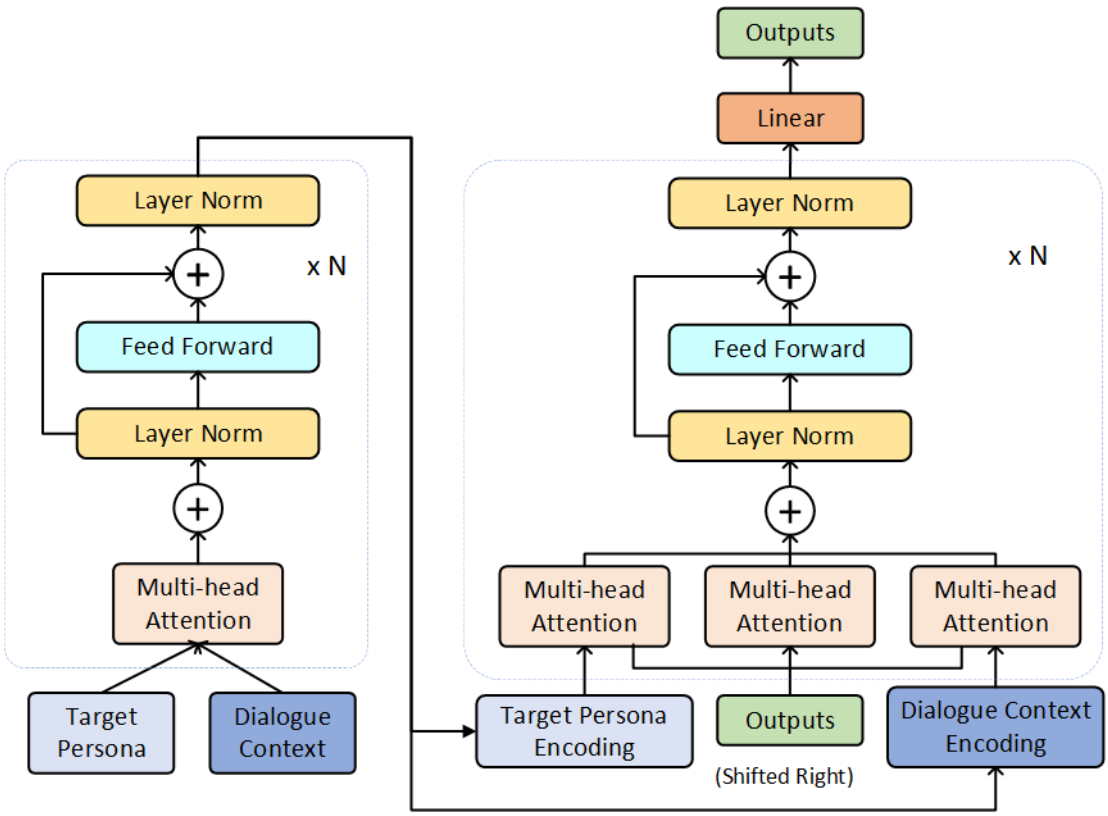
对个性化的特征，数据集中是类似结构化的数据，但是我们在处理时将其处

理为字符串的形式。例如“*标签:游戏动漫;双子座;宅;音乐;90 后;WOW 台服众,地点:安徽;合肥,性别:男*”，这个时候将个性化的信息作为整体字符串作为模型的输入。

2.1.3 个性化对话抽取

由于原始的大规模对话里，很多都是与个性无关的，所以会影响模型学习个性特征与对话场景的关系。所以针对该问题，我们利用一定的启发式策略筛选出一批与个性化相关的对话。策略如下：如果一组对话里出现（男、女、哥、姐、弟、妹），我们则认为其与性别个性相关；如果出现地点里的省份或城市则认为与位置个性相关；如果出现兴趣标签里的任意标签，则认为这种对话与兴趣个性相关。按照这种方式，我们一共过滤出 88 万组与个性特征相关的对话。

2.2 模型结构



图(2): 个性化对话生成模型结构

我们的模型由基本的 seq2seq 模型组成，如图（2），encoder 和 decoder 都由 transformer block 组成，其中 encoder 端对 persona 和 history 分别编码。然后将编码的结果送入 decoder 端和已经解码出的部分序列通过 multi-head attention 的方式进行信息融合，并解码得到输出。

另外，我们模型的 encoder 和 decoder 是由预训练的中文 GPT 的参数做初始化，并在训练的过程中共享 encoder 和 decoder 的参数。

2.3 训练设定

2.1.3 预训练 GPT 模型训练细节

我们预训练的中文 GPT 模型的框架与原始 GPT 结构[9]相同，包括 12 层的 transformer block，隐层状态的向量维度为 768，12 个 attention head。词汇大小为 13084，句子长度为 512，dropout 设置为 0.1。另外使用 Adam 作为梯度更细算法，最大的学习率为 2.5e-4，warm-up 的大小设置为 2000。此外我们预训练 GPT 的语料来自于我们搜集的 1.3G 的中文小说语料，并在 8 块 Nvidia GTX 1080TI 上训练了 70 个 epochs。

2.1.4 seq2seq 模型训练细节

Seq2seq 模型的训练主要分为三步：

第一、利用训练好的 GPT 模型初始化 encoder 和 decoder 的参数，并在训练的过程中共享参数。

第二、利用所有官方提供的 543 万组对话，训练 seq2seq 模型，尽管个性特征有可能并没有体现在对话中。这一阶段主要让模型学习

encoder 和 decoder 端的关系映射。该过程一共训练 2 个 epoch。

第三、用提取出的 88 万个性化明显对话语料训练 seq2seq 模型，因为这些对话有较强的个性特征，所以模型更容易学习如何使用个性化的特征。

上面三个阶段的训练各有自己的特点，第一阶段语言模型预训练阶段是为了更好地学习文本的通顺性，第二阶段大规模对话数据训练时为了更好的学习输入和输出之间的映射关系，第三阶段的个性化数据上的训练是为了更好的让模型学会用个性化的特征。

3.实验结果及分析

实验结果主要分为自动指标评估和人工评估两部分，对于每一种评估方式，分别在随机选取的对话测试集和有偏选取的个性化测试集上进行评估，并取两者的平均值作为最后的分数。

3.1 自动评测指标

自动评测指标主要包括：

BLEU：评估输出回复相对于标准回复的 n-gram 重合度。

Perplexity：评估模型所输出结果的流畅性。

Distinct：评估输出回复的多样性。

	BLEU	**Perplexity**	**Distinct**
biased	0.0084	224.10	0.297
random	0.0039	361.25	0.134
average	**0.006**	**292.67**	**0.216**

表(1): 我们模型随机和有偏的评测结果

表(1)为我们最后随机部分和有偏部分的结果，可以看出有偏部分测试集的

各项指标是要远高于随机部分的指标的，这是由于有偏部分是人工挑选的，并且都是包含个性的对话，分布空间较小，并且我们模型在第三阶段是特别针对有个性化的数据进行微调训练过的，所以指标相对随机的部分较好。

3.2 人工评测指标

人工测评过程中在“流畅性”“个性化”和“上下文相关性”三个维度对生成结果进行打分，分数从[-1, 0, 1]中取值。该过程由三个评测人员独立进行，最终队伍的排名使用三位评测人员打分结果的平均值确定。图(3)列举了人工评测的各队伍分数。

	biased_test_set			random_test_set			Average			Rank
	流畅性	个性化	相关性	流畅性	个性化	相关性	流畅性	个性化	相关性	
Golden_response（原微博回复）	0.828333	0.65	0.77	0.67833	0.0466	0.79333	0.753332	0.3483	0.781665	
网易伏羲实验室	0.75333	0.23833	0.52	0.73	0.12166	0.56833	0.741665	0.179995	0.544165	1
彩云科技&句子互动	0.57	0.105	0.63667	0.455	0.025	0.73	0.5125	0.065	0.683335	2
华南理工大学-CIKE实验室	0.42166	0.07833	0.49333	0.4	0.01333	0.53333	0.41083	0.04583	0.51333	3
中国科学院深圳先进技术研究院	0.46833	0.03	0.365	0.43666	0.01333	0.46333	0.452495	0.021665	0.414165	4
NEU NLP LAB	0.655	0.03	-0.18833	0.675	0.02833	-0.00833	0.665	0.029165	-0.09833	5
WRFML LAB	0.0767	0.0317	-0.12353	0.0818	-0.00333	-0.06844	0.07925	0.014185	-0.09599	6
东北大学	0.04333	0.02833	-0.26166	0.00333	-0.025	-0.375	0.02333	0.001665	-0.31833	7
复旦大学大数据学院	-0.49333	-0.00166	-0.89666	-0.39	-0.00166	-0.85166	-0.44167	-0.00166	-0.87416	8

从人工评测的结果来看，我们的流畅性和个性化的指标远高于其他队伍，甚至在随机部分的测试集上超过了 golden response，这主要得益于我们在个性化数据上的微调，另外我们的流畅性高主要得益于语言模型和对话模型在第一、第二阶段的预训练。我们的相关性比第二名稍微差一点，可能是因为在微调的过程中过分专注于个性特征的使用，而减少了对上下文的关注。

4.总结

这篇文章中我们针对大规模的个性特征稀疏的对话语料，基于预训练的方法，提出了一种三段训练的方式来尽可能的生成符合用户个性化特征的回复，该方法通过前两阶段预训练的语言模型和对话模型的方法保证了生成语言的流畅性和对话之间的逻辑性，最后通过个性化数据上的微调，尽可能的生成与个性化信息相关的回复。该方法在 SMP2019ECDT 评测中取得了人工指标评测第一，自动指

标评测的第二名。

参考文献

[1] Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.;and Dolan, B. 2016. A persona-based neural conversationmodel. InACL, 994–1003

[2] Kottur, S.; Wang, X.; and Carvalho, V. 2017. Exploringpersonalized neural conversational models. InIJCAI, 3728–3734.

[3] Qian, Q.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Assigning personality/identity to a chatting machine for co-herent conversation generation. InIJCAI.

[4] Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; andWeston, J. 2018. Personalizing dialogue agents: I have adog, do you have pets too? InACL, 2204–2213

[5] Mazar´e, P.-E.; Humeau, S.; Raison, M.; and Bordes, A.2018. Training millions of personalized dialogue agents. InEMNLP.

[6] Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2018.Transfertransfo: A transfer learning approach for neural net-work based conversational agents. InNIPS2018 CAI Work-shop.

[7] Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark,C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextual-ized word representations. InProceedings of the 2018 Con-ference of the North American Chapter of the Associationfor Computational Linguistics: Human Language Technolo-gies,

Volume 1 (Long Papers), 2227–2237. New Orleans,Louisiana: Association for Computational Linguistics.

[8] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018.Bert: Pre-training of deep bidirectional transformers for lan-guage understanding.arXiv preprint arXiv:1810.04805.

[9] Radford,A.;Narasimhan,K.;Salimans,T.;andSutskever, I.2018.Improving language un-derstandingbygenerativepre-training.URLhttps://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/ Languageunderstanding paper. pdf.

[10] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; andSutskever, I. 2019. Language models are unsupervised mul-titask learners.OpenAI Blog1(8).