预训练模型在开放式对话领域的应用

作者：钱胜杰

彩云科技&句子互动

摘要

目前开放领域的对话生成已经取得了极大的进展和大量的关注，但是大部分情况下，由于缺少高质量的语料限制了模型的泛华性。为了克服这种缺陷，我们在对话的相似领域中收集了大量的单语语料，例如小说，使用 GPT[1]模型架构，预训练了一个语言模型。经过预训练的过程，模型本身学习到了大量和对话相关的知识。在对话生成任务上再去使用预训练模型，不仅提高了模型的泛华性，而且增强了模型的收敛速度。

关键词：对话、预训练

1.引言

在最近几年来，非任务型对话吸引了大量的学术和工业领域的兴趣。目前存在的方法分为两种，一种是检索式对话，通过从一个存在的对话语料库中选择一个最合适的回复，另一种是生成式对话，通过神经网络自回归的生成一个回复。因为检索式方法受限于事先建立好的语料索引，生成式对话变得越来越受关注。由于传统生成式模型很难生成长的，多样的回复，并且这些模型往往需要大量的对话语料去训练，这就导致大量的成本。随着最近预训练模型的发展，不仅是关注与表示层的 BERT[2]，还是生成式的 GPT，或者两者都兼顾的 MASS[3]和 XL-Net[4] 。这些模型显示预训练模型在自然语言处理中表现的高效性能，受此启发，我们首先使用收集到的小说语料预训练一个中文 GPT，然后在训练好的模型上进行微调，取得了不错的效果。

2.模型及方法介绍

SMP2019 中文人机对话技术测评，任务二：个性化竞赛。在对话场景下，已知对话上下文和所有对话参与者的个性化属性，要求生成符合给定个性化特性与上下文逻辑的回复 R。

所谓个性化属性由一系列键值对（如<性别， 男>， <年龄， 90 后>）描述，数据样例如下描述：

```
s1 黄昏，肃静的时光。我喜欢这种环境。
s2 有时间您也来逛逛吧
s1 好的，我都希望回中国。还希望下次在中国常驻工作时一个星期起马可以休息一天最好连休！这样就可以多看中国的美丽风景。
s2 那一定不是广本
p1:(性别：男，年龄：80后，地域：'海外 日本'，爱好标签：'重口味')
p2:(性别：女，年龄：90后，地域：'广东 广州'，爱好标签'音乐;自由;旅游;吃货')
```

图 1

S1 和 S2 是对话信息，p1 和 p2 是双方的个性化信息。模型在生成一个回复的时候，不仅需要考虑对话的上文信息，还需要考虑当前说话者的个人信息，例如图 1 中，在生成第三句话的时候就需要考虑当前说话者的地域信息。于是如何把这种个性化信息编码进模型是要考虑的问题。

由于主办方的测评脚本中提供了一个字典，并且是基于 jieba 分词。于是为了计算方便，我们模型的词表都是基于中文词的。

2.1 模型

传统的对话生成模型一般都是 seq-to-seq 模型，编码器用来编码对话的历史信息，解码器根据编码器的语义表示和之前已经生成的词信息来预测下一个词（字）。

由于我们想使用预训练好 GPT，于是如果采用编码器加解码器的架构，那么预训练 GPT 模型的作用就会被降低。所以整体的方案就是使用 GPT 模型去做对话生成任务，通过对对话历史和个性化信息编码，使模型可以生成下一句回复。

2.1.1 预训练

数据方面，我们一共收集大约 15G 的小说数据，但是由于时间和计算资源的影响，我们并没有全部用上，大约用了 4G 左右的数据大约有 10 亿 token 数量。

2.1.2 模型编码结构

图 2

模型在生成一个回复的时候应当考虑到对话历史信息，我们可以看到全部的历史对话，也可以只考虑最近的历史对话信息。如果对话历史轮数偏多，会影响模型注意力，一般情况下，一个对话的话题在三四轮以内。另一个方面，太长的数据会导致模型训练速度变慢，并且 SMP 数据的平均对话轮数 3 轮不到，太长的会影响整体的模型。于是在训练的时候，我们控制了对话轮数在 6 轮以内。

再次之前，也有人使用英文 GPT 去做对话[5]，但是我们模型和他们的有些不一样。如图 2，模型也需要考虑回复者的个性化信息，为了统一编码，我们把不同的个性化信息给拼接成一个句子，为了区分不同个性化信息的标签，模型额外的给每个标签加了一个类型编码，如图 2 的 fuse position embedding，灰色的是

标签编码，蓝色的是位置编码，之后所有的不同 embedding 相加得到每一个词的 embedding 信息送入到预训练 GPT 中。

3.实验结果及分析

	Bleu	diversity	Ppl
V1	0.01063	0.1647	135.5
V2	0.0105	0.163	120.3
V1&V2 Ensemble	0.0116	0.1429	107.5

表 1：V1 是模型 1，V2 是模型 2。可以看出来，模型 2 的 PPL 相比于模型 1 PPL 降了快 25 个点。

Top-P	Temperature	Bleu	Distinct
0.9	0.7	0.00529	0.243
0.9	1	0.002	0.378
TOP-K			
3	1	0.004417	0.1918
3	0.7	0.007	0.1886
3	0.9	0.0074	0.1885
5	1	0.00632	0.2166
Beam-size			
5	1	0.01	0.16
5	0.7	0.00959	0.1718
5	0.9	0.00996	0.1712
10	0.7	0.0101	0.1753

表 2

模型的词表是由 3 万比赛词表+12 万最高频的词，一共词表大小 15 万。由于时间紧张和计算资源缺乏，在预训练模型还未收敛时拿最好的 checkpoint 去微调对话模型，这是模型 1。最后几天又拿最新预训练模型去微调，获得模型 2。模型 1 和模型 2 的参数设置一样。使用官方的测试集数据，如表格 1 所示。

如表格 1 所示，收敛情况不同的预训练模型对最后模型的生成能力影响很大，基本上更好的预训练模型会得到更好的微调模型。并且可以看出，BLEU 指标和 Diversity 指标并不能很好的两者兼顾，最后提交的模型是两者的 ensemble，提升了最终模型的性能。

生成方式的不同也会很大程度上影响最终模型的生成效果，我们评测了 Beam-Search、top-p sampling、top-k sampling 在验证机上的表现。如表格 2 所示。选择最好的生成参数之后，我们又在 top-p 和 beam-search 上进行人工测评，一共 6 个人参与测评，最终决定选择 bean-search，size 为 5 作为最终的提交。

4.总结

在第一轮的自动测评中获得了第一名，大概是由于我们用了预训练模型，并且也扩张词了表的大小。但是在人工测评中，获得了第二，和第一名相比，我们模型的个性化能力不高，大概是由于在生成的时候的干预不多，完全是模型的输出。

由于主办方的词模型限制，导致我的预训练模型的作用被降低了。因为相比于字模型，UNK 比率大大上升，提高整体训练的难度。在此模型上模型的泛华能力会变弱，并且不同分词工具也会带来不确定性。

总的来说，第一次参加对话相关的比赛，收货很多。

参考文献

1. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI
2. Jacob Devlin · Mingwei Chang · Kenton Lee · Kristina Toutanova 2019 · north american chapter of the association for computational linguistics

3. Song K, Tan X, Qin T, et al. MASS: Masked Sequence to Sequence Pre-training for Language Generation[C]. international conference on machine learning, 2019: 5926-5936.
4. Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding.[J]. arXiv: Computation and Language, 2019.
5. Wolf T, Sanh V, Chaumond J, et al. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents.[J]. arXiv: Computation and Language, 2019.