

# 1200 万对话数据开放！清华 CoAI 课题组开源大规模对话数据和中文对话与训练模型

汉语堂2020-09-06 07:45

本文转载于：人工智能 THU

清华 CoAI 课题组开源 1200 万对话数据和中文对话预训练模型 CDial-GPT

## A Large-Scale Chinese Short-Text Conversation Dataset

Yida Wang<sup>1</sup>, Pei Ke<sup>2</sup>, Yinhe Zheng<sup>2,3</sup>, Kaili Huang<sup>2</sup>, Yong Jiang<sup>1</sup>, Xiaoyan Zhu<sup>2</sup>, and Minlie Huang<sup>2</sup> \*

<sup>1</sup> Tsinghua Shenzhen International Graduate School, Tsinghua University, China

<sup>2</sup> Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems. Beijing National Research Center for Information Science and Technology. Department of Computer Science and Technology, Tsinghua University, Beijing, China.

<sup>3</sup> Samsung Research China - Beijing (SRC-B), Beijing, China.

{wangyd18, kp17, hk116}@mails.tsinghua.edu.cn, yh.zheng@samsung.com, jiangy@sz.tsinghua.edu.cn, {zxy-dcs, aihuang}@tsinghua.edu.cn

基于 Transformer 的大规模预训练语言模型极大地促进了开放领域对话的研究进展。然而目前这一技术在中文对话领域并未被广泛应用，主要原因在于目前缺乏大规模高质量的中文对话开源数据。为了推动中文对话领域的研究，弥补中文对话语料不足这一问题，我们发布了一个包含 1200 万对话的大规模中文对话数据集 LCCC，并开源了在 LCCC 上预训练的大规模中文对话生成模型 CDial-GPT。开源地址：<https://github.com/thu-coai/CDial-GPT>

## LCCC 数据集的构建

LCCC (Large-scale Cleaned Chinese Conversation) 数据集有 LCCC-base 与 LCCC-large 两个版本,其中 LCCC-base 和 LCCC-large 中各包含 6.8M 和 12M 对话。这些数据是从 79M 原始对话数据中经过严格清洗得到的,也是目前所开源的规模最大、清洗最严格的中文对话数据集。

Type	Case
Platform tag	回复@精灵小宝贝: 我也失眠了
Advertisement	@张伟丽MMA 前来为DW持妆粉底液实力证言!
Generic form	我也是 我也是 我也是 啊啊
Dirty word	被 <b>小婊砸</b> 作的现在满身负能量
Special word	可以试试 <b>左氧氟沙星</b>
Name	<b>陈绍龙</b> 、你无朋友架啦
Symbol	#( ° - ° )#
Platform sign	文科574报哪里好?最好有师范英语【 <b>微信</b> 】
Not fluent	Q: 昨晚失眠了 A: 发恶梦扎醒,又发我最怕个蒋尸梦!
Incomplete information	Q: 江南小镇很美呀 A: <b>印象</b>
Time-sensitive	Q: 感觉没啥电视好看了 A: 琅琊榜, <b>就要大结局了</b>
External noun	Q: 假期要过了 A: <b>春节</b> 一过,好想立刻回到岛城
Irrelevant pairs	Q: 差点吧洗面奶当牙膏 A: 绿色是今年的流行色

表 1. 被过滤掉的噪音数据

开放领域对话数据的构建通常有三种方式:1、抽取剧本对话;2、人工众包构建对话;3、爬取社交媒体上用户的交流记录。使用第一种方式构建的对话在内容上依赖于特定剧情和场景,与日常对话有较大差异。使用第二种方式构建的对话质量最高,但是由于人力成本过高,无法使用这一方

式构建大规模数据集。使用第三种方式可以较为廉价地获取大规模对话数据，因此 LCCC 数据集中的原始数据主要使用第三种方式收集。

我们同时注意到，来自社交媒体的对话数据中存在各种各样的噪音(表 1)，为了保证 LCCC 中对话数据的质量，我们设计了如下数据获取和清洗策略：

### 1. 数据获取

我们的数据获取流程分为两个阶段。在第一个阶段，我们挑选了微博上由专业媒体团队运营的新闻媒体账号，然后收集了一批在这些新闻媒体下留言互动的活跃用户。在第二个阶段中，我们收集了这些活跃用户微博下的留言互动，并将其作为我们的原始数据。微博下的留言回复一般以一个树形结构展开，我们将这一树形回复结构中每一条从根节点到叶子节点的路径作为一个完整对话，最终共收集到了 79M 对话数据。

### 2. 数据清洗

为了保证数据质量，我们对收集到的原始对话数据进行了两个阶段的清洗。

第一阶段的清洗主要基于手工规则。这一阶段的主要目的是为了过滤掉对话数据中的明显噪声，如脏话、特殊符号、病句、复读机句式、广告、违法暴力信息等。在这一阶段中，我们花费了数周时间使用人工排查的方式优化规则。

第二阶段的清洗主要基于分类器过滤。在这一阶段中，我们基于 BERT 训练了两个文本分类器，第一个分类器主要用于甄别那些无法通过规则检测的噪音，如：1、语义模糊、语法错乱或有严重拼写错误的语句；2、时效性太强的对话；3、与上下文语义不相关的回复。第二个分类器主要用于甄别那些需要依赖额外上下文信息，如图片或视频等，才能理解的对话。这两个分类器均使用人工标注数据训练，我们为其标注了共计 11 万对话数据，最终的分类器在人工标注的测试集上分别达到了 73.76%和 77.60%的准确率。我们通过 F1-score 选择阈值来过滤得到高质量的对话数据。

	<b>Single-Turn Multi-Turn</b>		<b>Single-Turn Multi-Turn</b>	
Raw dialogs	52,708,95.5	26,749,365	63,251,887	28,189,952
Cleaned dialogs	3,354,382	3,466,607	7,273,804	4,733,955
Utterances	6,708,554	13,365,268	14,547,608	18,341,167
Characters	68,559,727	163,690,614	162,301,556	217,776,649
Vocabulary size	372,063	666,931	662,514	690,027
Avg. words	6.79	8.36	7.45	8.14
Avg. turns	2	3.86	2	3.87

表 2. 数据统计信息，左侧为 LCCC-base，右侧为 LCCC-large

最终我们基于上述原始对话数据过滤得到了 6.8M 高质量的对话数据 LCCC-base。此外，我们还收集了目前已公开的其他对话数据，并使用同样的清洗流程，结合 LCCC-base 构造了包含 12M 对话的数据集 LCCC-large。表 2 展示了这两个数据集中单轮对话和多轮对话的详细统计信息。

### 中文对话预训练模型 CDial-GPT

为促进中文对话预训练模型的发展，我们基于 LCCC 数据集预训练了大规模中文对话生成模型 CDial-GPT。该模型的训练过程包含两个阶段，首先，我们在总计 5 亿字符、包含各类题材的小说数据上训练得到了一个中文小说 GPT 预训练模型，然后在该模型的基础上，我们使用 LCCC 中的对话数据继续对模型进行训练，最终得到了中文对话预训练模型 CDial-GPT。

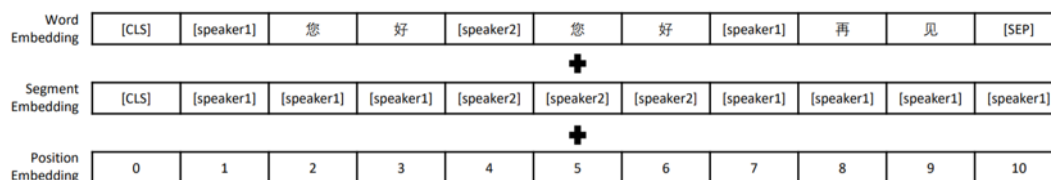


图 1. 输入编码示例

该模型拥有 12 层 Transformer 结构，我们按字分词，字典大小 13088，字向量维度 768，最长上下文长度为 513。我们沿用 TransferTransfo 的方式对对话进行建模，即把对话历史拼接为长文本，并使用段分割向量加以区分。具体来说：我们使用[CLS]字符标志文本起始，在段落后使用[SEP]字符表示段落结束，在段落中对相邻轮次对话使用[speaker1] [speaker2]交替分割，并在 segment embedding 中使用[speaker1]、[speaker2]进行编码。图 1 为输入数据示例。

## 模型效果评测

为了评估对话预训练模型的质量，我们在 440 万规模的中文对话数据集 STC 上对其进行了评测实验，并对比了现有的中文对话预训练模型和一些经典的非预训练对话模型。我们主要通过 PPL 这一指标来反映模型的拟合能力，PPL 越低表示模型的拟合能力越强。我们通过基于 n-gram 重合度的指标 BLEU 和基于 Embedding 相似度的指标 Greedy Matching 和 Embedding Average 来衡量对话回复与真实回复的相关性，并通过 Dist-n 指标来衡量生成回复的多样性。实验结果展示在表 3 中。可以看到我们的模型在绝大多数指标上达到了最好的效果。由于自动指标无法完全反映生成对话的质量，于是我们对各模型生成的对话进行了人工评测。

Models	Size	PPL	BLEU-2	BLEU-4	Dist-1	Dist-2	Greedy Matching	Embedding Average
Attn - Seq2Seq	73M	34.2	3.93	0.90	0.0085	0.1191	0.6584	0.8338
Transformer	113M	22.10	<b>6.72</b>	3.14	0.0088	0.1397	0.6606	0.8355
GPT2 - chitchat	88M	-	2.28	0.54	<b>0.0103</b>	<b>0.1625</b>	0.6154	0.7894
GPT <sub>Novel</sub>	104M	21.27	5.96	2.71	0.0080	0.1172	0.6612	0.8334
CDialGPT <sub>LCCC-base</sub>	104M	18.38	6.48	3.08	0.0083	0.1268	0.6621	0.8354
CDialGPT2 <sub>LCCC-base</sub>	104M	22.76	5.69	2.50	0.0077	0.1087	0.6624	0.8346
CDialGPT <sub>LCCC-large</sub>	104M	<b>18.23</b>	6.63	<b>3.20</b>	0.0083	0.1271	<b>0.6625</b>	<b>0.8363</b>

表 3. 自动指标评估

Models	+2	+1	+0	Score	Kappa
Seq2Seq	27.1%	21.4%	51.4%	0.756	0.4544
Transformer	42.4%	23.6%	34.0%	1.084	0.4429
GPT2 – chitchats	24.3%	19,1%	56.6%	0.677	0.3941
CDialGPT <sub>LCCC-base</sub>	46.7%	21.8%	31.5%	1.152	0.3954
CDialGPT <sub>LCCC-large</sub>	<b>48.6%</b>	<b>24.5%</b>	<b>27.0%</b>	<b>1.217</b>	0.4203

表 4. 人工评价

在人工评测中我们主要考虑 3 个方面：1、语法性，也就是生成语句的流畅性；2、相关性，即生成语句与对话上文的相关性；3、信息量，即生成结果自身含有的信息量。具体来说，不符合语法性或对话上文不相关的生成结果我们给予 0 分；语句流畅、和对话上文相关但信息量不足的生成结果给予 1 分；语句流畅、和对话上文相关并且信息量充足的生成结果给予 2 分，我们将人工评测结果展示在表 4 中。实验结果显示，我们的预训练模型拥有出色的生成质量，可以在生成信息量充足的回复的同时，保持较高的流畅性与相关性，优于其他基线模型，一些生成例子展示在表 5，6 中。

Role	Response	Role	Response
User	别爱我没结果	User	生命的意义是什么
Bot	没结果的爱是多么的痛。	Bot1	生命的意义是过得比你想象的要好
User	但是爱过就可以了	Bot2	也许吧
Bot	有时候，爱情就是一场游戏。	Bot1	嗯，生命的意义是过得比你想象的要好
User	输的总是我	Bot2	说的好
Bot	我爱你，就像我爱你一样。	Bot1	好什么
User	你可拉到吧	Bot2	过得比你想象的好
Bot	你可拉到吧。	Bot1	好吧
User	别学我说话		
Bot	我说话就是你不行。		

表 5. 人机交互示例（左）和 模型互相交互示例（右）

Post	response
<p>这样洗手洗得干净吗?!?(°_°)</p> <p>作为一名伟大的吃货你必须知道的事情。</p> <p>惊现特种喵!这身手。可以维护世界和平了吧。</p> <p>大家热的不行,干脆比比谁的舌头漂亮超萌的考拉。不怕痒,挠吧。</p> <p>你觉得最好看的英剧 Top 3 是什么?</p> <p>你觉得最好看的英剧 Top 3 是什么?</p> <p>当你迷茫的时候。</p> <p>终于知道自己为什么失眠了。</p> <p>无论身在何方,我心里满满都是你!</p> <p>安东尼刷屏,可惜小手太小,就是投不进啊</p> <p>披萨饼盒瞬间变笔记本支架,这个太酷了!</p> <p>孩子们踢完球后去游泳。好漂亮的泳池!</p> <p>从来扬州风流地,最是自在小盘古。</p> <p>凤凰古城,千年一醉,太美了。</p> <p>无论你多么讨厌你的学校,当你离开久了,你还是會想念它。</p>	<p>那是必须的,不过洗手的时候会有点脏。</p> <p>大半夜的不睡觉,你是在诱惑我吗?</p> <p>这猫太有才了,这猫很有想象力啊!</p> <p>好像是你的舌头,呵呵</p> <p>我的小心脏要跳出来了</p> <p>我的英剧 top 3 是《老婆婆》</p> <p>最好看的是《宇宙大战》</p> <p>当你迷茫的时候,找点事做做</p> <p>为什么会失眠?你不是想太多了吗?</p> <p>此时此刻,我只想说一句:好可爱!</p> <p>这小子太牛了,真想给他一拳。</p> <p>这个真的可以有。但是不是很实用</p> <p>有好多入啊,我要去游泳。</p> <p>小盘古。还是扬州好啊</p> <p>好美的凤凰古城,希望有机会能去看看。</p> <p>还是学校好,没有负担,没有压力</p>

表 6.在 STC 微调的生成结果示例

目前 CDial-GPT 模型以及 LCCC 数据集都已公开,我们提供了训练以及微调代码,可以方便地应用于各种数据和下游任务上。读者可以通过以下链接获取数据集以及模型:<https://github.com/thu-coai/CDial-GPT>