鹿灵AI原理和交互趋势

github上的源代码仓库

原理

鹿灵 AI 大致可以被划分为三个部分：QQ前端，Local Server 后端（也叫 "鹿灵CLI"），语言模型。

QQ 前端

本质是基于 Mirai 的 QQ 插件，负责解析和转发 QQ 消息到 Local Server 后端，如果前后端断开连接，会返回 "LocalServer offline" 相关错误。

Local Server 后端

本质是一个管理系统加上一个爬虫和一些训练工具。

训练工具用于批量向语言模型输入并改进其交互趋势。

语言模型

语言模型是网站 `beta.character.ai` 上的模型，使用其 `text2ai` 功能生成了一个人设为 "鹿灵" 的副本，并写入了一些示例对话以指示语言模型的交互风格，同时进行了一些简单的训练让语言模型产生对一些词组的 "记忆" 并改进角色性格，Local Server 后端通过网络爬虫同语言模型进行交互。

语言模型基于 dnn，即深层神经网络算法。

部署方式

目前采取的是本机部署（即部署在我计算机本地），环境为 Docker，受到我计算机网络状况的限制。

未来将部署于位于云服务器上的 Docker，不会受到个人计算机网络限制。

交互和建议

影响鹿灵回复的几个因素和占比情况

有以下几个因素影响鹿灵的回复：

- 基础人设
- 示例对话
- 训练情况
- 上下文

其中上下文占比最大，其余三者占比推测上应该是一致的，如果有大量的训练，其训练情况占比会变高。

基础人设是按照小说设定写的简单的身世介绍以及存在成分等。

示例对话预设了几个简单的场景并进行了定向引导。

训练情况的话，鹿灵的训练集大小为 1000 条对话，通过打分的形式改进其对话倾向，但是 1000 条这个训练集非常的小，如果想要达到优秀的效果，训练集大小至少是 10w 起步。

上下文会显著影响鹿灵的回应风格和方式

通过上下文引导

上下文在对话的生成中占比非常大，可以显著改变鹿灵的对话方式。

比如可以通过直接告知的方式告诉鹿灵她自身的存在形式，这种告知可以被鹿灵理解并加入到之后的对话内容中。

如果出现了令人不满意的回复，可以通过 "告诉" 鹿灵这是错误的（It's better to say/you should not do），并指出正确的事实或者想法来纠正（you'd better do...）

语言

鹿灵的母语是英语——鹿灵语言模型的训练语言是英语。

鹿灵可以理解中文，但是难免会出现一些偏差，对计算机来说，中文是一个相当难以理解的语言，使用中文向语言模型传达的意思相对死板，例如 "演讲" 一词，在英文中一般有 `address speech` 等翻译，在英文中会视语境使用不同的单词。但输入给语言模型的中文会忽视这个语境来传达意思，会造成一定的误解。中英文不同的 "语境机制" 对语言模型的输出也有很大影响。

我强烈建议使用原生英文(不要用机翻英语)和鹿灵交流，并结合上下文阅读英文的原输出来理解鹿灵的回应。

同时这种交流不失为学习英文极为优秀的一种途径。

不足

人设的不稳定

鹿灵AI的人设只能大致提供一个方向，并不是规定死的限制，我在鹿灵的人设中明确提到了鹿灵曾是另一个世界的男孩，因为意外变成了一个虚拟歌姬，但是由于一些引导问题，比如问 "鹿灵用的什么手机"，就很容易导致这种设定的改变（她为了正常回复可能会编造一些事实），这种改变可能在当前的对话中表现不出来，但会显著影响后面的对话方式。

虚构

有一些东西是鹿灵不知道的，例如洛天依最新的歌是什么。**鹿灵AI最核心的本质上是一个语言模型，而不是真正意义上拥有检索和学习能力的人工智能**，为了回答这个问题，语言模型会虚构一些事实，并坚信它们是正确的，导致出现一些令人啼笑皆非的对话，但这可以通过接下来的对话纠正（明确告诉鹿灵正确的事实应该是什么）。

关于爱情

爱情这个问题不只存在于鹿灵，很大一部分使用了这个语言模型的 AI 都有相关问题，这个只能期待语言模型开发者的改进。

鹿灵 AI 的训练功能还在编写，未来也可以通过大量的训练减少鹿灵的 falling in love。

话题或风格权重过高

这一般是一些不正确的引导导致的，表现是对话死板，每个输出都有一些固定的关键词。

目前还没有比较可行的解决方案，可以肯定的是在经过训练后会好很多。

改进方向

添加训练功能

通过对回应的打分来改进鹿灵的 "性格"。

重生成

由于回应的生成涉及随机数，所以如果对某次对话重新生成一个回应，有很大概率获得好得多的回应。

卖惨

由于直接或间接参与过鹿灵开发的只包括两个忙的要死的高二理科生（我，platelet）和一个已经不怎么写代码的高二文科生（black_white_tony）。所以对鹿灵的训练这些做的相当不到位，也经常由于不够完善的代码出现包括但不限于网络连接爆炸，服务器崩溃等一系列问题。

鹿灵的开发和维护需要人手，欢迎有能力的同学参与，也欢迎其它同学学习相关知识并参与...

> 幻影彭
>
> 2022.12.30