“源1.0”API调用使用手册

尊敬的用户： 衷心感谢您选用了浪潮人工智能巨量模型“源1.0”API！ 本手册介绍了“源1.0”已开放API的接口说明和使用示例，可使使用者更好地了解本API支持的功能及快速使用方法，充分的发挥开放API的作用。 浪潮（北京）电子信息产业有限公司拥有本手册的版权。 本手册中的内容如有变动恕不另行通知。 如果您对本手册有疑问或建议，请向浪潮（北京）电子信息产业有限公司垂询。

浪潮（北京）电子信息产业有限公司

二零二二年四月

目录

序

“源1.0”，是浪潮人工智能研究院9月28日在京发布全球最大规模人工智能巨量模型。“源”的单体模型参数量达2457亿，超越美国OpenAI组织研发的GPT-3，成为全球最大规模的单体AI巨量模型。 本文将介绍如何进行“源1.0”API的调用。该API接口主要是针对外网开放，用于第三方用户根据自身需求获取推理结果。

0. 安装依赖

“源1.0”项目尽可能采用了目前python API调用所需的主流依赖库，如果您之前有过相关开发经验，将不需要进行额外安装，如果您的电脑和python环境处于初始化状态，可以采用如下命令安装或确认相关依赖：

```
pip install requests hashlib json
```

完成安装后在您的硬盘上任意位置，用如下命令将GitHub上的代码fork下来即可。

```
git clone https://github.com/Shawn-Inspur/Yuan-1.0.git
```

需要注意的是，`GitHub`上下载的代码包含了三个部分，`src`中是模型训练的代码，`sandbox`中是web示例开发使用的沙箱代码，需要额外安装`yarn`，`yuan_api`中是采用API进行推理的示例代码和工具。手册中将重点介绍

yuan_api的使用，并基于此简单介绍如何采用web沙箱进行示例应用创建。

1. 快速开始

我们开源了“源1.0”的训练代码以及采用API进行推理的代码，为了方便开发者社区的同仁快速上手，我们将推理的API代码进行便捷性封装。参照yuan_api/examples中的示例代码，可以快速完成不同的NLP应用开发。这里我们将带大家一起，通过构建一个对话机器人来快速上手源推理API开发过程。 首先请确认您的项目工作目录为yuan_api，然后在examples目录下新建一个python文件：dialog.py。

```
from inspurai import Yuan, set_yuan_account,Example
```

代码中首先从inspurai导入Yuan和Example这两个类，以及set_yuan_account函数。其中Yuan类中包含了用于提交API请求以及获取推理结果的各种函数，具体将在后文中进行解释。Example类用于构建输入输出示例，通过对yuan实例中添加example，即可实现one-shot或few-shot。

如果您的API审批申请已经通过，请用申请API时使用的账号和手机号来获得授权。

```
# 1. set account
set_yuan_account("用户名", "手机号")
```

初始化Yuan实例yuan，并对其加入一个样例：

```
yuan = Yuan(input_prefix="对话：“",
            input_suffix="”",
            output_prefix="答：“",
            output_suffix="”",)
# 3. add examples if in need.
yuan.add_example(Example(inp="对百雅轩798艺术中心有了解吗？",
                         out="有些了解，它位于北京798艺术区，创办于2003年。"))
```

其中input_prefix和input_suffix分别为输入的前缀和后缀。output_prefix和output_suffix为输出的前缀和后缀。如果对实例yuan添加了example，则会在提交query前在example的输入和输出部分分别添加前缀和后缀。例如上面添加的example，在提交query时将会被自动修改为以下的样式：

```
"对话：“对百雅轩798艺术中心有了解吗？”答：“有些了解，它位于北京798艺术区，创办于2003
年。”"
```

如果有需要，实例yuan中还可以继续添加更多example，实现few-shot。其他参数将在后面详细介绍。

至此一个问答机器人就已经完成了，接下来就可以提问了。我们把想问的问题放在prompt变量里，然后提交给yuan的API。

```
# 4. get response
prompt = "故宫的珍宝馆里有什么好玩的？"
```

```
response = yuan.submit_API(prompt=prompt,trun="”")
```

其中trun为截断符，yuan API推理服务返回的生成结果有可能包含重复的答案，通过设置trun可以在第一次出现该字符时截断返回结果。因为我们在之前设置的输出后缀为“”,对于推理返回的结果，我们可以将trun设为“”,在返回完第一个完整输出时将其截断。因为截断时最后这个字符并不会被保留，为了保持我们对话机器人输出符号的对称性，我们人为在打印时加上后引号。（注：这种设计是必要的，因为对于更普遍的任务而言，加入的后缀是无意义的，仅作为语句分割用。我们并不希望这种字符被返回。）

为了能够连续进行对话，我们将上面提交prompt和返回结果的过程重构如下：

```
print("====问答机器人====")

while(1):
    print("输入Q退出")
    prompt = input("问：")
    if prompt.lower() == "q":
        break
    response = yuan.submit_API(prompt=prompt,trun="”")
    print(response+"”")
```

这样一个简单的问答机器人就开发完毕，您可以在命令行和他互动了！

2. 函数调用说明

上一节中我们已经使用一些函数，本节将具体说明本开源代码中所开放的函数用途和参数，可作为技术手册查阅。

2.1 账号设置

yuan推理API实例化相关的代码都放在inspurai.py文件中，这里将逐一对其中的函数和类进行介绍。

```
set_yuan_account(user, phone)
```

set_yuan_account函数将用户名和手机号设置为环境变量，并在后续推理时从环境变量中读取这两个参数，实现用户名和手机号验证。

参数名	含义	取值范围
user	用户名	API申请时填写的用户名
phone	手机号	API申请时填写的手机号

2.2 Yuan推理实例

Yuan类：

Yuan类为用户调用浪潮“源1.0”API最常用到的类。用户可以通过这个类设置用户信息、添加样例、提送API请求等。实例化方法如下：

```
yuan = Yuan(engine='base_10B',
            temperature=0.9,
            max_tokens=100,
            input_prefix='',
            input_suffix='\n',
            output_prefix='答:',
            output_suffix='\n\n',
            append_output_prefix_to_query=False,
            topK=1,
            topP=0.9)
```

为方便用户设置，所有参数均给定了默认值。

参数名	含义	取值范围
engine	大模型后台推理引擎，目前可选的推理引擎有基础模型，对话模型和翻译模型	'base_10B'：基础模型 translate'：翻译模型 'dialog'：对话模型
temprature	模拟退火温度参数。 值越大，使得概率分布越尖锐，模型的创造性越强，但生成效果不稳定。 值越小，模型的稳定性越强，生成效果稳定	float:[0,1]
max_tokens	最大生成token长度，数值越大，生成时间越长。不建议超过200。	int:[0~200]
input_prefix	输入序列前缀，如设置，将自动为query和每个样例的输入加上前缀	任意字符串
input_suffix	输入序列后缀，如设置，将自动为query和每个样例的输入加上后缀	任意字符串
output_prefix	输出序列前缀，如设置，将自动为每个样例的输出加上前缀	任意字符串
output_suffix	输出序列后缀，如设置，将自动为每个样例的输出加上后缀	任意字符串
append_output_prefix_to_query	如设置，将自动将设定的输出前缀添加到query序列的末尾	bool型
topK	挑选概率最高的 k 个 token作为候选集。 若k值为1，则答案唯一。 当topK为0时，该参数不起作用。	int：[0,-]

参数名	含义	取值范围
topP	token 的概率累加，从最大概率的 token 往下开始取， 当取到累加值大于等于topP时停止。 当topP为0时，该参数不起作用。	float：[0,1]

添加样例：

```
add_example(ex)
```

添加示例到Yuan实例。

参数名	含义	取值范围
ex	Example类的实例	Example类

删除样例：

```
delete_example(id)
```

根据样例id删除样例。

参数名	含义	取值范围
id	Example实例的id	int:uuid

按照id获取实例中的样例：

```
get_example(id)
```

参数名	含义	取值范围
id	Example实例的id	int:uuid
return	返回样例	

返回实例中所有的样例：

```
get_all_examples()
```

参数名	含义	取值范围
return	返回样例字典	关键字为样例id，值为样例字典

将所有样例拼接到输入序列，应为私有函数，不直接使用：

```
get_prime_text()
```

参数名	含义	取值范围
return	拼接后的字符串	

获取推理引擎：

```
get_engine()
```

参数名	含义	取值范围
return	实例所用的推理引擎	字符串

获取模拟退火温度参数：

```
get_temperature()
```

参数名	含义	取值范围
return	temp值	float:[0,1]

获取最大token设置：

```
get_max_tokens()
```

参数名	含义	取值范围
return	最大输出token数量	int:[0,-]

将样例和query拼接成输入序列：

```
craft_query(prmopt)
```

参数名	含义	取值范围
prmopt	除样例外，用户输入的内容	任意字符串
return	query序列	任意字符串

将前后缀添加到Query和样例中：

```
format_example(ex)
```

参数名	含义	取值范围
ex	样例实例	

获取大模型推理API得到的原始结果：

```
response(query,engine='base_10B',max_tokens=20,temperature=0.9,topP=0.1,topK=1):
```

函数会异步调用两个API，首先提交query请求到后台，获取请求id，然后按照id轮询“答复”API端口，直到获取到推理生成的结果或超时。

参数名	含义	取值范围
query	包含样例、用户输入、前后缀在内的query字符串。	字符串，长度小于2048
engine	大模型后台推理引擎，目前可选的推理引擎有基础模型，对话模型和翻译模型	'base_10B'：基础模型 'translate'：翻译模型 'dialog'：对话模型
temprature	模拟退火温度参数。 值越大，使得概率分布越尖锐，模型的创造性越强，但生成效果不稳定。 值越小，模型的稳定性越强，生成效果稳定	float:[0,1]
max_tokens	最大生成token长度，数值越大，生成时间越长。不建议超过200。	int:[0~200]
topK	挑选概率最高的 k 个 token作为候选集。 若k值为1，则答案唯一。 当topK为0时，该参数不起作用。	int:[0,-]
topP	token 的概率累加，从最大概率的 token 往下开始取，当取到累加值大于等于topP时停止。 当topP为0时，该参数不起作用。	float:[0,1]
return	大模型声称的原始内容	任意字符串

删除特殊字符：

```
del_special_chars(msg)
```

将大模型生成的内容中的特殊字符剔除，如'<unk>', '<eod>', '#', '▄', '▁', '▂', ' '等

参数名	含义	取值范围
msg	规范化后的生成文本	模型返回的内容
return	剔除后的msg	

将prompt提交到API，并返回处理后的生成结果：

```
submit_API(prmopt, trun='▬')
```

提供用户端使用的方法，将用户输入的内容传递到Yuan的API接口，并返回处理后的生成结果。

参数名	含义	取值范围
prmopt	问题或其他用户输入的内容（不包括样例）	任意中文字符串
trun	截断符，设置后结果将在生成内容中第一次出现截断符的地方截断	任意字符

2.3 样例实例

`Example`类：

用于给推理query添加示例，实例化方法如下：

```
example = Example(inp="", out="")
```

参数名	含义	取值范围
inp	样例输入	任意字符串
out	样例输出	任意字符串

`Example`实例化方法通常与`Yuan.add_example()`函数一起使用，为`Yuan`的实例添加样例。

获取样例输入：

```
get_input()
```

返回example的样例输入。

获取样例输出：

```
get_output()
```

返回example的样例输出。

获取样例id：

```
get_id()
```

返回样例的id

以字典形式返回样例：

```
as_dict()
```

返回以字典形式返回样例

关键字	含义	取值范围
input	样例输入	字符串
output	样例输出	字符串
id	样例id	uuid编码

3. 底层API接口

为了便于用户使用，第2节介绍的函数和方法为对底层API的高级封装。本节将介绍模型推理API的底层接口，用户可以基于本节内容，自定义更适合自身业务的API逻辑。

本节介绍的底层API保存在`url_config.py`当中。

目前底层API开放两个基础URI：

```
SUBMIT_URL = "http://api-
air.inspur.com:32102/v1/interface/api/infer/getRequestId?"
REPLY_URL = "http://api-air.inspur.com:32102/v1/interface/api/result?"
```

SUBMIT_URL用于提交用户的query到后端，返回查询id。

REPLY_URL用于根据查询id，异步查询后台是否完成针对用户query的推理生成。

生成后端服务验证所需要的token：

```
header_generation()
```

token会放置在head中传递给后台API接口。token会从环境变量中读取用户申请/设定的用户名和手机号，并添加时间戳进行加密。即使token被泄露，一天后token也会失效，需重新生成。

参数	含义	取值范围

参数	含义	取值范围
return	url请求的headers	md5加密的字符串

提交request：

```
submit_request(query,temperature,topP,topK,max_tokens,engine)
```

将最终的query通过`requests.get()`方法提交到uri。

参数	含义	取值范围
query	包含样例、用户输入、前后缀在内的query字符串。	字符串，长度小于2048
engine	大模型后台推理引擎，目前可选的推理引擎有基础模型，对话模型和翻译模型	'base_10B'：基础模型 'translate'：翻译模型 'dialog'：对话模型
temprature	模拟退火温度参数。值越大，使得概率分布越尖锐，模型的创造性越强，但生成效果不稳定。值越小，模型的稳定性越强，生成效果稳定	float:[0,1]
max_tokens	最大生成token长度，数值越大，生成时间越长。不建议超过200。	int:[0~200]
topK	挑选概率最高的 k 个 token作为候选集。 若k值为1，则答案唯一。 当topK为0时，该参数不起作用。	int:[0,-]
topP	token 的概率累加，从最大概率的 token 往下开始取，当取到累加值大于等于topP时停止。 当topP为0时，该参数不起作用。	float:[0,1]
return：requestId	请求id，用于异步查询生成结果	

获取生成结果：

```
reply_request(requestId,cycle_count=5)
```

根据requestId查询“结果返回”API接口是否有结果生成。

参数	含义	取值范围
requestId	submit_request生成的请求id	字符串
cycle_count	轮询次数，每次间隔3s	int:[1,-]

参数	含义	取值范围
return: response_text	返回生成的结果	字符串

4. 应用示例

这里我们汇总了一些简单的应用示例配置方法，以便用户参考。其中未提及的参数均采取了默认值。

序号	应用	模型	prompt模板	输入前缀	输入后缀	输出前缀	截断符	输入示例	few-shot
0	对话生成	dialog	问：“用户输入” 答：“	问：“	”	答：“	”	故宫有什么好玩的?	支持
1	内容续写	base_10B	用户输入	无	无	无	默认	徐凤年刚走入京大校门，已经有学生会迎新的同学走到了他面前，	不建议
2	诗词生成	base_10B	以“用户输入”为题作一首诗：”	以“	”为题作一首诗：“	无	”	清风	推荐
3	关键词抽取	base_10B	为以下正文提取关键词。正文：用户输入；关键词：	为以下正文提取关键词。正文：	；	关键词：	。	帮我写一首诗，描写春天到了，百花盛开。	支持
4	中英互译	translate	将下列英文/中文翻译成中文/英文。英文/中文：用户输入 中文/英文：“	将下列英文/中文翻译成中文/英文。英文/中文：	无	中文/英文：“	”	自然派的哲学家也被称为“苏格拉底之前的哲学家”。	不建议

更多应用尽请期待。