

# 让AI准备训练数据

## Question

生成一些json格式的feature-label数据项，用于retrieval-based chatbot的训练。

训练目的是让模型实现intent classification和entity recognize。

query的内容是学校设施相关的

示例：query: where is Nasi Kandar restaurant?

intent: ask for a position

entity: nasi kandar restaurant

如果我的例子给得有问题，请修改

最后输出你认为合适的json代码片段

## Answer

好的，你的例子结构是没问题的，非常适合用于 Intent Classification（意图分类）和 Entity Recognition（实体识别）的训练。你的例子中 `intent` 是用户查询的目的，`entities` 是查询中包含的关键信息。

对于学校设施相关的检索式聊天机器人训练，我们可以设定一些常见的意图和实体类型。

### 建议的意图 (Intents):

- `find_location` : 询问某个设施的位置。
- `ask_hours` : 询问某个设施的开放时间或服务时间。
- `ask_service` : 询问某个设施提供哪些服务或功能。
- `ask_contact` : 询问某个设施或部门的联系方式（电话、邮箱等）。
- `ask_availability` : 询问某个设施或资源是否可用（如会议室是否空闲，电脑是否可用）。
- `ask_existence` : 询问某个设施或服务是否存在。
- `ask_type` : 询问某个地点是什么类型的设施。
- `Maps` : 请求导航或指路（这个可能需要更复杂的处理，但可以作为意图识别出来）。

### 建议的实体类型 (Entity Types):

- `facility_name` : 具体设施的名称（如 Main Library, Student Center, Block A, Lecture Hall 101）。

- `facility_type` : 设施的通用类型 (如 restaurant, library, lab, office, restroom, parking lot, cafe)。
- `department_name` : 部门名称 (如 Computer Science Department, Admissions Office)。
- `service_name` : 服务的名称 (如 printing service, Wi-Fi access, book borrowing)。
- `time` : 时间相关的词语 (如 today, tomorrow, weekends, 3 PM)。
- `relative_location` : 相对位置描述词 (如 nearest, nearby, upstairs, downstairs)。
- `building_part` : 建筑的特定部分 (如 third floor, room 301, east wing)。
- `person_name` : 人名 (如果在询问某个人的办公室或位置时)。

## 你的例子修改建议:

你的例子非常好, `intent` 可以更标准化一些, 实体类型也更明确一点。

原例:

```
1  {
2    "query": "where is Nasi Kandar restaurant? ",
3    "intent": "ask for a position",
4    "entity": "nasi kandar restaurant"
5  }
```

修改后:

```
1  {
2    "query": "where is Nasi Kandar restaurant?",
3    "intent": "find_location",
4    "entities": [
5      {"text": "Nasi Kandar restaurant", "type": "facility_name"}
6    ]
7  }
```

这里我们将 `intent` 改为更常用的 `find_location`, 并将 `entity` 改为 `entities` 列表, 包含实体的文本和类型。

## 生成一些 JSON 格式的 feature-label 数据项:

下面是一些基于学校设施主题的 JSON 数据项示例, 包含不同的意图和实体类型:

```
1  [
2    {
3      "query": "where is the main library?",
4      "intent": "find_location",
5      "entities": [
6        {"text": "main library", "type": "facility_name"}
7      ]
8    },
9    {
```

```

10     "query": "What are the opening hours for the student center?",
11     "intent": "ask_hours",
12     "entities": [
13         {"text": "student center", "type": "facility_name"},
14         {"text": "opening hours", "type": "time"}
15     ]
16 },
17 {
18     "query": "Does the IT support office provide laptop repair service?",
19     "intent": "ask_service",
20     "entities": [
21         {"text": "IT support office", "type": "facility_name"},
22         {"text": "laptop repair service", "type": "service_name"}
23     ]
24 },
25 {
26     "query": "What is the contact number for the Admissions Office?",
27     "intent": "ask_contact",
28     "entities": [
29         {"text": "Admissions Office", "type": "department_name"},
30         {"text": "contact number", "type": "contact_info_type"} // 新增实体类
    型表示联系信息类型
31     ]
32 },
33 {
34     "query": "Is Meeting Room B on the 5th floor available now?",
35     "intent": "ask_availability",
36     "entities": [
37         {"text": "Meeting Room B", "type": "facility_name"},
38         {"text": "5th floor", "type": "building_part"},
39         {"text": "now", "type": "time"}
40     ]
41 },
42 {
43     "query": "Is there a swimming pool on campus?",
44     "intent": "ask_existence",
45     "entities": [
46         {"text": "swimming pool", "type": "facility_type"}
47     ]
48 },
49 {
50     "query": "What kind of facility is the Innovation Hub?",
51     "intent": "ask_type",
52     "entities": [
53         {"text": "Innovation Hub", "type": "facility_name"}
54     ]
55 },
56 {
57     "query": "Where is the nearest restroom?",

```

```
58     "intent": "find_location",
59     "entities": [
60         {"text": "nearest", "type": "relative_location"},
61         {"text": "restroom", "type": "facility_type"}
62     ]
63 },
64 {
65     "query": "Tell me the library hours for this Saturday.",
66     "intent": "ask_hours",
67     "entities": [
68         {"text": "library", "type": "facility_type"},
69         {"text": "this Saturday", "type": "time"}
70     ]
71 },
72 {
73     "query": "Can I print documents in the computer lab?",
74     "intent": "ask_service",
75     "entities": [
76         {"text": "print documents", "type": "service_name"},
77         {"text": "computer lab", "type": "facility_name"}
78     ]
79 }
80 ]
```

## 使用说明：

- 上面的 JSON 数组包含了 10 个数据项。
- 每个数据项是一个对象，包含 `query` (用户输入文本)、`intent` (识别出的用户意图)、`entities` (识别出的实体列表)。
- `entities` 是一个数组，每个元素是一个对象，包含 `text` (实体在 `query` 中的文本) 和 `type` (实体的类型)。

这只是一个很小的示例数据集。在实际训练中，你需要收集大量的用户查询数据，并手动或半自动地标注它们的意图和实体，以构建一个更大、更全面的数据集来训练你的模型。数据越多样化、覆盖的意图和实体越全面，模型的识别效果就会越好。