Maluuba dataset

链接地址：https://datasets.maluuba.com/Frames/dl

该数据集基于假期预定的场景——具体来说，查找航班和宾馆。Maluuba 让两个人在聊天室中对话并收集了这些数据。一个人扮演用户，另一个人充当计算机。用户试图查找特价机票，另一个充当聊天机器人的人使用数据库检索信息。

对于该数据集的意图识别更加困难的，用户经常改变谈话主题。你可能同时讨论去滑铁卢、蒙特利尔、多伦多的计划。Maluuba 发布的数据集侧重于进行同时涉及多个主题的对话。

该数据集包含 1369 个有关旅行规划的多轮对话，共有 19986 个行对话，每个多轮对话约 14.6 行对话。

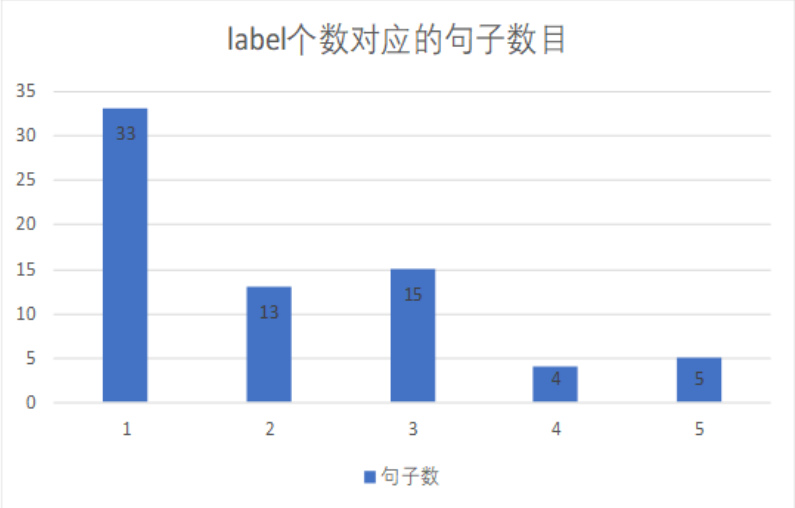
通过查看一些数据集的意图标注，如：

句子：tell me about the weather

意图：weather-query（Siri or Cortana）

通过人工标注的 50 句对话，5 个对话片段，其分析如下：

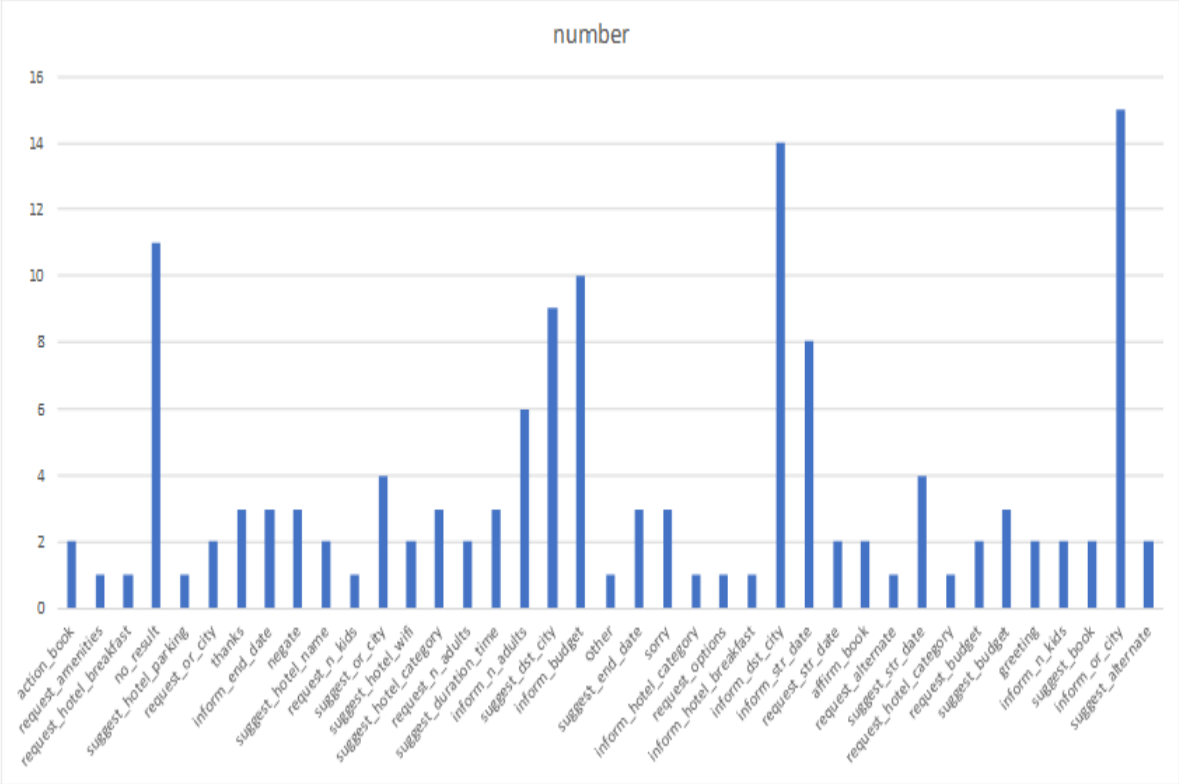
每个句子的 label 个数如下：



Label 的分布如下：

label	number
action_book	2
request_amenities	1
request_hotel_breakfast	1
no_result	11
suggest_hotel_parking	1
request_or_city	2
thanks	3
inform_end_date	3
negate	3
suggest_hotel_name	2
request_n_kids	1
suggest_or_city	4
suggest_hotel_wifi	2
suggest_hotel_category	3
request_n_adults	2
suggest_duration_time	3
inform_n_adults	6
suggest_dst_city	9
inform_budget	10
other	1
suggest_end_date	3
sorry	3
inform_hotel_category	1
request_options	1
inform_hotel_breakfast	1
inform_dst_city	14
inform_str_date	8
request_str_date	2
affirm_book	2
request_alternate	1

suggest_str_date	4
request_hotel_category	1
request_budget	2
suggest_budget	3
greeting	2
inform_n_kids	2
suggest_book	2
inform_or_city	15
suggest_alternate	2



举个例子：

User:Hi I'd like to go to Caprica from Busan, between Sunday August 21, 2016 and Wednesday August 31, 2016

意图：inform_dst_city（提出目的地） inform_or_city（提出出发地） inform_str_date（提出出发时间） inform_end_date（提出结束时间）

Wizard：And what would be your maximum budget for this trip?

意图：request_budget（询问预算）

User:Actually it's unlimited for this trip

意图：inform_budget（提出预算）

总结：1.句子的意图不仅仅与本句话相关，还与上文有关，必须结合上文分析，才能准确理解该句的意图。2.数据中，无关句子较少，基本为打招呼和感谢。3.文中部分标签带有依赖性，比如 inform_dst_city(目的地)和 inform_or_city(出发地)相互关联，inform_str_date(开始时间)和 inform_end_date(结束时间）相互关联。4.label 的分布不均衡，有些 label 的样本多，有些 label 的样本少。