**面向音乐领域的命令理解任务**

# 任务描述

对话系统是自然语言处理中一个重要的研究方向，也是人机交互的一种重要形式。其中，用户话语（utterance）可以根据意图的不同进一步分为聊天、问答、命令等。对于任务完成式系统而言，能够正确解析用户命令是完成指定任务的基础。因此，本任务主要关注口语对话系统中的命令理解问题。

在本次任务中，我们主要关注音乐领域，并进一步将音乐领域的命令理解分为两个子任务：1）音乐领域意图判断；2）音乐领域槽填充。接下里，分别介绍两个任务的目标。

* 1. 音乐领域意图判断

音乐领域意图判断的目标为判断用户的某条话语（utterance）是否表达了一个音乐领域内的意图。另外，在实际对话系统中，用户的一条话语往往存在相当大的歧义和模糊。因此，在本次测评中，系统需要通过一个话语片段（包括同一用户连续4个话语），判断最后一条话语片段是否表达了音乐意图。

* 1. 音乐领域槽填充

用户的某条话语（utterance）被标注为音乐领域意图后，为了完成该意图，需要将该话语中提及的相关参数提取出来，在此，话语中的相关参数被称为“槽”（slot）。比如说，在音乐领域中，最常见的槽是“歌手”和“歌曲”等。将相关参数识别出来的任务，称为槽填充。

本次评测任务的主要目标是针对对话系统中用户的真实话语，判断该条话语是否表达了音乐领域的意图，如果是，则完成该意图的槽填充。所有语料来自原始的对话系统用户日志，并经过了筛选和人工的意图判断和槽填充。

**输入**：用户的一个话语片段，其中包括3句话语。

**输出**：最后一条话语是否是音乐意图，并且填充相关槽。

**示例**：

*样例1*

输入：

23123, "来首刘德华的歌 2017-10-17 19:41:53","你叫什么名字  2017-10-17 19:42:51","请播放周杰伦的稻香   2017-10-17 19:43:53"

输出：

23123,{"song":"稻香","artist":"周杰伦"}

*样例2：*

输入：

57643, "请播放张学友的歌    2017-10-17 05:12:53","你叫什么名字  2017-10-17 05:12:40","今天我被人打了。    2017-10-17 05:13:53"

输出：

57643, {"code":"No-music"}

# 数据描述

* 1. 数据格式说明
		1. 输入数据格式说明

每一行的数据格式包含了四列，内容如下:

**utterance\_snippet\_id,** **"utterance\_1","utterance\_2","utterance\_3"**

字段说明：

* utterance\_snippet\_id:该次话语片段的唯一标识序列号
* utterance\_1, utterance\_2, utterance\_3为本次话语片段中的连续三次用户的话语。其中每个话语分为了三列，第一列为用户话语的文本内容，第二列为话语的请求日期，第三列为该话语的请求时间。
	1. 训练集、验证集、测试集
		1. 数据来源

本次测评数据是主要来自人机对话系统中音乐领域以及非音乐领域的真实用户utterance请求记录。

* + 1. 数据标注过程说明

本次测评发布的数据为人机对话系统中，对于utterance的解析结果。之后，在通过标注人员对于系统给出的utterance解析结果进行校正。对于系统解析结果与人工标注结果不同的utterance再进行交叉验证，从而得到此次测评的发布数据

* + 1. 训练集、验证集、测试集的说明：

**训练集&验证集**：

在训练及验证数据发布阶段，我们会发布1.2w对左右的标注训练问句对数据集和3000对左右的验证问句对数据集。

**测试集**：

在测试数据发布阶段，我们将会再发布2000对左右的问句对数据集，不含标注结果，作为测试集。

# 评价指标

* 1. 音乐领域意图判断

音乐领域意图判断为分类任务，对此，我们采用精确率（Precision, P）、召回率（Recall, R）、F1值（F1-measure, F1）作为评价指标。

$$正例意图判断精准率P\_{p}=\frac{正例标注正确的话语数}{预测为正例的话语总数}$$

$$负例意图判断精准率P\_{n}=\frac{负例标注正确的话语数}{预测为负例的话语总数}$$

$$意图判断精准率P\_{I}=\frac{P\_{p}×N\_{p}+P\_{n}×N\_{n}}{N\_{p}+N\_{n}}$$

$$正例意图判断召回率R\_{p}=\frac{正例标注正确的话语数}{标注为正例的话语总数}$$

$$负例意图判断召回率R\_{n}=\frac{负例标注正确的话语数}{标注为负例的话语总数}$$

$$意图判断召回率R\_{I}=\frac{R\_{p}×N\_{p}+R\_{n}×N\_{n}}{N\_{p}+N\_{n}}$$

$$意图判断F1\_{I}值=\frac{P\_{I}×R\_{I}}{P\_{I}+R\_{I}}$$

* 1. 音乐领域槽填充

音乐领域槽填充为序列标注任务，对此，我们分两种情况讨论：

1. 对于包含实体的话语

$$包含实体精准率P\_{E}=\frac{预测正确的实体数}{预测出的实体数}$$

$$包含实体召回率R\_{E}=\frac{预测正确的实体数}{标注出的实体数}$$

$$包含实体F1\_{E}值=\frac{P\_{e}×R\_{e}}{P\_{e}+R\_{e}}$$

1. 对于无实体的话语

$$无实体准确率Acc\_{NE}=\frac{预测为无实体的话语数}{标注为无实体的话语数}$$

* 1. 整体数据集

整体而言，我们采用如下方式评价一个方法的好坏：

$$Score=\frac{意图判断话语总数×F1\_{I}+包含实体话语总数×F1\_{E}+无实体话语数×ACC\_{NE}}{数据集所有话语总数}$$

若两个队伍Score得分相同，则$F1\_{E}$更高的队伍排名更高；若Score与$F1\_{E}$得分相同，则$F1\_{I}$更高的队伍排名更高；若这三者也相同，则$Acc\_{NE}$更高的队伍排名更高。

# 任务提交

* 1. 提交格式说明

本次任务将采取刷榜的方式，验证集发布后，允许参赛队伍多次向平台提交结果，文件命名为“参赛队名称\_valid\_result.txt”，格式与结果文件result.txt相同，排名随时更新。参赛队伍可在评测集发布之前随时上传验证集的计算结果，管理系统会及时更新各队伍的最新排名情况；

测试集发布后，允许参赛队伍多次提交验证集结果文件，结果文件提交格式如下。

每一行的数据包含两列，每一列之间以制表符\t分割，内容如下所示:

**utterance\_snippet\_id’\t’semantic\_result**

* semantic\_result为根据该话语片段对其中的utterance\_3进行意图解析的结果，其基本内容为json格式。
* 若该话语片段中utterance\_3未表达音乐意图则输出应为：{"code":"No-music"}。
* 若该话语片段中utterance\_3表达了音乐意图则输出因为具体的字段和对应的值，其中存在该话语中存在哪些字段需要参赛系统自行判断。比如，若utterance\_3为“我想听周杰伦的稻香”，则输出应该为“{"song":"稻香","artist":"周杰伦"}”；若utterance\_3为“我想听周杰伦的歌”，则输出应该为“{"artist":"周杰伦"}”；若utterance\_3为“我想听稻香”，则输出应该为“{"song":"稻香"}”；若utterance\_3为“我想听首歌”，则输出应该为“{"code":"RANDOM"}”。

以下为音乐领域的意图解析与槽填充的意图解析结果，即semantic\_result字段说明

|  |  |
| --- | --- |
| 字段 | 解释 |
| album | 专辑名 |
| artist | 歌手人名 |
| genre | 曲风字段 |
| lyrics | 歌词 |
| mood | 心情字段 |
| code | 表明为RANDOM或者No-music |
| scene | 场景 |
| tag | 音乐标签 |

备注：

code字段的取值有两个,

* RANDOM:表明该utterance的意图为播放任意歌曲

No-music表明为非音乐字段

* 1. 评测平台

本次评测将依托biendata平台（https://biendata.com/）展开，请有意向的参赛队伍关注平台上的竞赛列表。

# 时间安排

评测任务发布：4月1日

报名时间：4月1日—7月15日

训练及验证数据发布：4月20日

测试数据发布：7月15日

提交测试结果：7月20日

评测论文提交：8月5日

# 组织者

中国科学院自动化研究所，刘康

中国科学院自动化研究所，郭尚敏

北京云知声信息技术有限公司，刘升平

北京云知声信息技术有限公司，张勇

关于本评测有任何问题，请联系：ccks2018@unisoud.com