多轮问答调研

1. **检索式多轮**
2. ***使用场景***

用户不会严格按照一问一答的形式去提问，问句中会缺失一些他前面提到过的信息。

示例：

Transformer rewriter数据集

|  |  |  |  |
| --- | --- | --- | --- |
| User | bot | user | 改写 |
| 保险能养老好吗 | 养老有医疗保险啊 | 你觉得好吗 | 你觉得医疗保险好吗 |
| 百年康惠保 | 什么保险 | 重疾类 | 百年康惠保是重疾类的保险 |
| 是在新华保险公司 | 工作没几年的上班族适合买什么保险 | 亲爱的我不知道去这个单位的选择对不对 | 亲爱的我不知道去在新华保险公司的选择对不对 |
| 人工智能是啥 | 就是大概有脑子的计算机 | 可以说是有学习能力的软件吗 | 人工智能可以说是有学习能力的软件吗 |

1. ***技术方案***
2. 多轮对话改写

将多轮对话改写为一句话

Transformer ReWriter: <https://arxiv.org/abs/1906.07004>

1. 多轮对话检索

基于历史对话进行检索

DAM: <https://www.aclweb.org/anthology/P18-1103.pdf>

MRFN:<https://www.semanticscholar.org/paper/Multi-Representation-Fusion-Network-for-Multi-Turn-Tao-Wu/1556fcd90e60218066505b0d3cd842788b8cb929>

1. **任务式多轮**
2. ***使用场景***

解决用户某一实际任务进行的多轮交互

数据集示例：

CrossWOZ（2019ACL） [<https://github.com/thu-coai/CrossWOZ>]

|  |  |
| --- | --- |
| User | 你好，麻烦帮我推荐一个 门票 免费 的景点。 |
| bot | 你好，您可以选择 故宫，八达岭长城，颐和园 或者 红砖美术馆。 |
| user | 这几个景点中哪个评分为5分？ |
| bot | 你好，红砖美术馆的评分为5分。 |
| user | 好的，那红砖美术馆的地址在哪？ |
| Bot | 地址为：北京市朝阳区崔各庄乡何各庄村顺白路马泉营西路路口西100米路北。 |
| user | 好的，那游玩时间大概是多长时间？ |
| Bot | 游玩时间在1小时 - 2小时。 |
| user | 好的，麻烦再帮我看一下拿渡麻辣香锅(万达广场店)的营业时间。 |
| bot | 营业时间为：周一至周日 09:00-21:00。 |
| User | 那这家店评分是多少？ |
| Bot | 拿渡麻辣香锅（万达广场店）评分为4.5分。 |

说明：上面的多轮对话完成了订票或者订酒店的任务，需要对用户话术中的实体以及意图进行识别。

1. ***技术方案***

**涉及的技术：**

1）实体/意图识别

NER识别，意图分类；或者意图和实体联合训练。

1. 对话状态追踪（DST）

基于规则的状态追踪，基于模型的

1. 对话策略学习

基于有限状态机规则；基于强化学习

1. 自然语言生成

基于模板；基于模型（SC-LSTM [<https://arxiv.org/pdf/1508.01745.pdf>]）

**对接的答案数据来源：**

1. 知识图谱
2. 业务API或者数据库
3. **IM场景**

IM小程序感觉适合知识图谱以及基于检索的多问问答，首先问题中有大量的关于保险信息的问题，eg”xxxx保险的价格”，“XXX能不能投保XXX”。这种问题目前用一问一答来做需要扩充大量的问题库，对检索召回以及相似度匹配的准确率都是问题。在前期没有知识图谱的情况，可以采取实体+目的拼接的方案或者句子融合方案近似解决上述问题。另外这种保险知识图谱对保险推荐也有一定帮助，可以发现一些互补保险或者相关保险进行推荐。

IM的任务型问答目前是依靠前端点选来实现的，而且目前也没有什么比较好的任务场景，所以暂时可以先这样。

1. **难点&问题**
2. 对于多轮需要用到的模型，例如：实体识别，意图识别，多轮对话改写，实体链接等等都需要标注数据, 前期无标注数据只能使用一些规则的方法。
3. 对于知识图谱，需要大量数据来构建知识图谱。对非结构化数据还涉及到实体关系属性的提取。
4. 对于基于槽位的任务型多轮，需要设计出明确的多轮应用场景，以及每个场景的意图还有槽位。比如【查询保费】场景，有哪些槽位，哪些是必填槽位，哪些是可选择的。
5. 多轮对话的对话管理模块也是个难点，目前落地的大部分都是规则写的有限状态机，用序列模型也可以做，难点就是需要大量的历史聊天标注数据。要是用强化学习也可以，但是需要建立一个用户模拟器，弄这个就又需要数据了。