XANTHOUS

Bot Friday Talk

SMS CRM Chatbot w/ Rasa v2

梁皓然Simon
simon@x-tech.io

Introduction

XANTHOUS

梁皓然Simon，广州先思科技CEO，前美国亚马逊全栈工程师，Elasticsearch认证工程师，Wechaty、Rasa、OpenAI GPT-3/Codex开源社区贡献者，数字游民

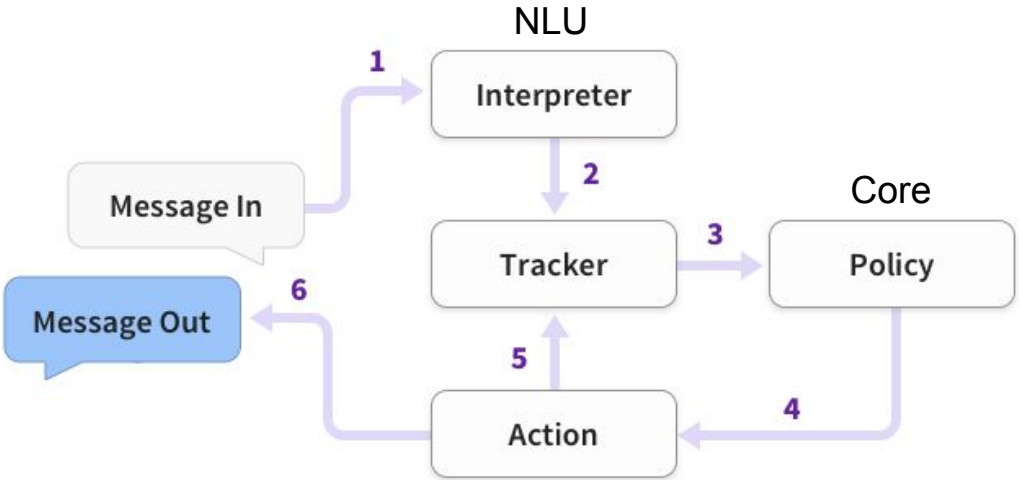
Bot Friday Talks: Rasa v1 New Features, Voice Chatbot w/ Snowboy

文晓新Clovin
先思科技全栈工程师
前网易云音乐工程师

赵春奇Alex
先思科技运维工程师
CNCF认证k8s管理员 (CKA)

Rasa Architecture

- https://legacy-docs-v1.rasa.com/user-guide/architecture/



XANTHOUS

Rasa Architecture Digest

- Interpreter - NLU - `(string) => { text, intent, entities }`
  - Dense Featurizers via Pre-trained Language Models (like BERT)
  - Dual Intent Entity Transformer (DIET) Classifier
- Tracker - `{ history: Array<Event>, slots: Map<string, any> }`
- Policy - Evaluate, predict then choose the best to determine next action
  - Transformer Embedding Dialogue (TED) Policy
- Action - `(Tracker) => Array<Event>`

arXiv.org > cs > arXiv:1706.03762

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5)]

Attention Is All You Need

客户需求介绍

- 客户在美国，是一个房屋中介SaaS软件
- 采集C端用户来源之后，通过Twilio群发短信息模板激活潜在用户
- 用户一旦回复，Chatbot启动，开始采集相关信息，判断用户是否近期有购房需求
- 如果购房需求强烈，C端用户成为hot lead，SaaS接入B端中介公司进行下一步沟通

Example

From UI

Drip: Daisy! How's the house search been lately? Sara's assistant w/ xxx.
Daisy: We still looking the hause but getting hard to find it
Bot: We're putting together some listings for you now and will get those sent to you soon! Are you just looking out of curiosity or are you considering making a move within the next year or so?

…

SMS vs IM

- Text only (no buttons)
- 短信信息含量比较高，往往一条消息有多种含义
- 短信模板接入转化率低，需要流畅的chat flow和人性化的语料提高转化率

客户遇到的挑战

XANTHOUS

- 目前使用第三方提供商现成chatbot服务，费用高，每个月需要处理~300k会话，每月session数量持续增多
- 花一年以上时间基于Rasa V1尝试制作chatbot，但是一直停留在意图识别阶段，用于处理70%以上无效sessions（用户不愿意接收短信的情况）

我看到的问题

- 客户现有AI工程师不清楚如何正确使用Rasa制作Chatbot，并对Story Training不买账，也没有尝试寻找替代方案，如Rasa NLU + BotFramework
- 客户期望使用现成数据，有标注团队，但是使用的工具非常原始（Tableau + Google Sheets），数据量大，效率低下
- 客户对Chatbot的成功标准没有明确定义，盲目追求意图识别准确率，迟迟无法推进Chatbot构建流程

Main Chatbot Flow

- 群发消息模板(Drips)
  - 含有一定的context，用户回复的内容通常和drip相关，需要提供一定个性化回复
- 如果用户开始回复短信，进入Slot Filling流程
  - 如果用户的回复没有包含填充槽位的关 键信息，不重复提问
  - 允许中途跳出或暂停(通过后续群发消息重新激活)
- 所有关键Slots填充完毕，进入简单对话模式(Chitchat)
  - 如果遇到不能回答的 问题，统一回复“会让中介解答”

Our Solution V1

- Drips
  - Story Training
- Slot Filling
  - Rasa Form Action
- Chitchat
  - Story Training

Annotation Tool

See UI

Solution V1 Issues

- 现成数据有随机出现的闲聊，story需要明确intent输入和明确action输出，不允许同样的intent输入对应不同action输出，无法100%还原对标Chatbot
- Rasa默认Form Action不够灵活，需要用自定义Form Validation Action提取Slots，原厂Python SDK不利于工程化

XANTHOUS

We Broke Rasa!

- 使用了multi-intent功能，intent组合达到了2k+
- YAML配置模板不能胜任大模型，debug非常困难
- Rasa X无法正常工作
- 修改任何Domain配置需要重新训练（使用GPU需要6小时以上），特别是response语料，对大模型调优不友好

Our Solution V2

- 使用SQL数据库管理Rasa Domain配置
  - 确保intent，action，response能在YAML里正确关联，防止误删操作导致训练出错
  - 制作了Rasa Model Builder前端项目，通过UI管理整个Rasa Domain配置，保存到SQLite
- NLU
  - ConveRT (by PolyAI) - https://arxiv.org/abs/1911.03688
  - NER - facebook/duckling + Microsoft/Recognizers-Text + AWS Comprehend + pyap + Rasa Lookup Tables
  - DIETClassifier (Rasa)
- 全套基于AWS的自动训练Pipeline
  - 标注团队完成标注后，可以提交训练任务
  - 训练使用AWS Batch自动申请GPU机器训练，训练后自动销毁机器，节省成本
  - AWS CodePipeline构建带模型的Docker Fat Image，方便版本回滚

Our Solution V2 (cont.)

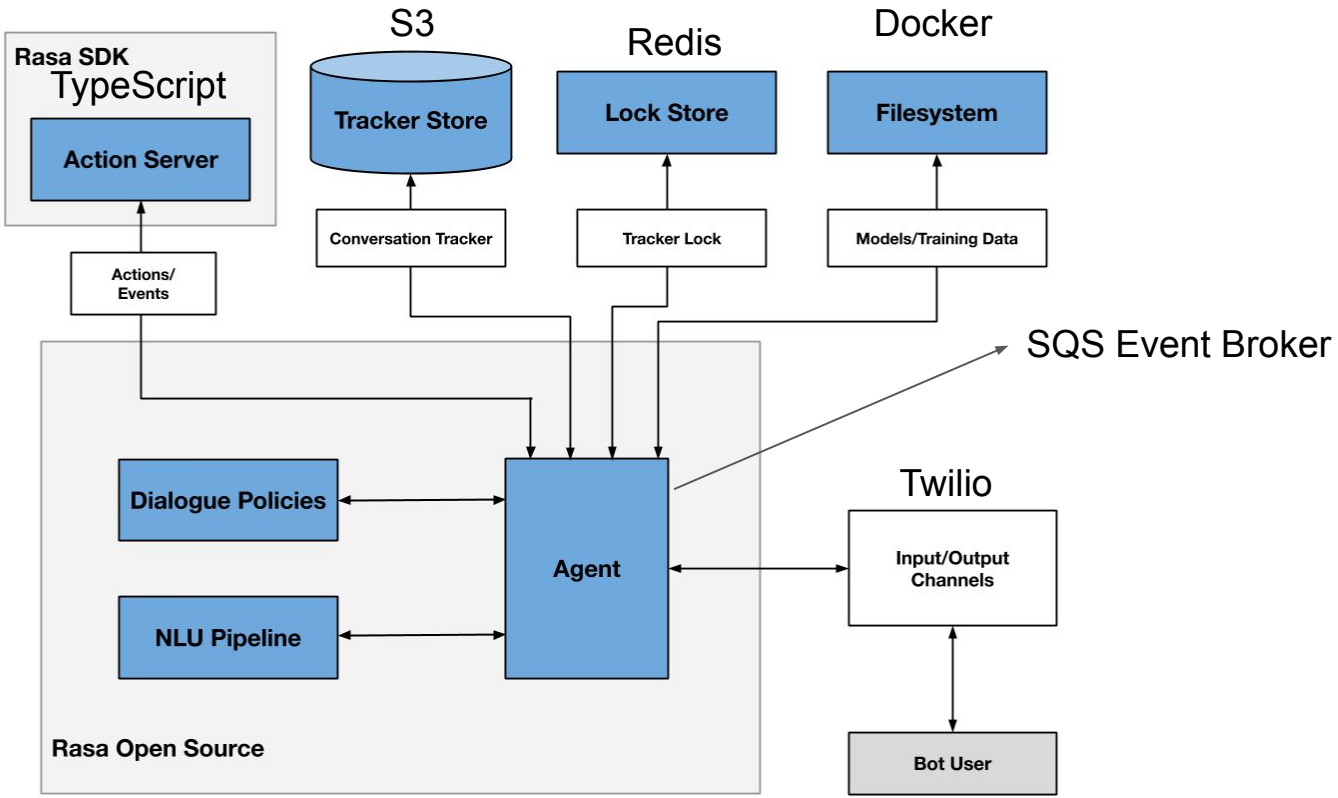
- 用TypeScript编写Rasa SDK，制作Action Server和NLG Server
  - @xanthous/rasa-sdk , @xanthous/rasa-action-server , github
  - 全面拥抱工程化 (NestJS, Dependency Injection, Prisma)
  - 通过SQLite读取规则配置，为Form Validation服务
  - 快速制作可灵活配置的 规则编辑器UI
  - NLG Server避免语料修改需要重新 训练Rasa模型

TODO: screenshots / screen share

Our Solution V2 (cont.)

- Main Chatbot
  - 不带任何stories，只有少量rules
  - 使用自定义的 `action_default_fallback` 作为入口 (router)，通过判断slots状态和当前轮intents组合做出了state machine的效果 - HACK
  - Drips
    - 根据不同的群发模板，和对应的规则配置，做出应对和chitchat，并进入选定的Form
  - Slot Filling (Forms)
    - 每一轮根据规则配置从intent或者entity提取slot的值，并做出应对和chitchat
    - 全部required slots填充之后，往 `form_completed` slot填充状态为True
- Completed Chitchat Chatbot
  - Form completed之后chitchat采用story training
  - 当`form_completed` slot为True，`action_default_fallback` 把消息从main bot转发到chitchat bot

Production Architecture



XANTHOUS

Takeaways

- 造chatbot !== 造火箭，现有框架和算法已经能够快速制作聊天机器人
- 造chatbot需要造UI，把标注团队和设计团队效率最大化
- 造chatbot最好从最小最简单的domain做起，慢慢扩充应对的语料和分支，能更快上线，获取更贴合实际的真实数据
- Rasa是个好框架，但是Python不一定能完美工程化
- ConveRT是个极好的预训练模型，期待有开源实现和中文预训练模型

XANTHOUS

Shameless Plugs

- 开源项目
  - xanthous-tech/rasa-chinese-paddlenlp - Rasa通过PaddleNLP提供中文支持（感谢段清华）
  - xanthous-tech/chat-operator - 基于NestJS的chat中间件（Rasa和微信客服/Wechaty快速桥接）
  - lhr0909/jupyterlab-codex - 在JupyterLab对OpenAI Codex模型进行测试
  - More around Rasa & OpenAI soon!
- Bilibili视频 - 西门良
  - Rasa + Siri 点麦当劳 - github
  - GPT-3中文介绍
  - OpenAI Codex中文介绍
- 承接大中小型自定义chatbot的搭建项目！我们是专业的！

Questions?

Thank you!