

每天打电话提醒你拿快递的，也许是这个AI语音助手.....

原创：邱陆陆 机器之心 3天前

机器之心原创

作者：邱陆陆

新年刚过，身在老家的我就接到了这样一个电话。电话对方明显是一个对话机器人，我们在用时不到一分钟的时间里完成了下面这几轮对话：

我：我近期不在家。

它主动转为询问我是否有其他可投递的地址。

我没有直接回答，而是转为询问：是什么东西啊？

它告诉我是文具。

我确认了包裹内容后给出了投递地址：那麻烦帮我放到物业吧。

它重复我给出的地址，我确认，它礼貌地结束对话。

之后快递员没有再联系我，回到北京之后，我顺利地到物业拿到了快递。

作为（伪）AI 行业从业者，之心编辑部对这个 语音助手的多轮、多目标能力都表示震惊，于是顺藤摸瓜找到了菜鸟语音助手背后的算法研发团队，和阿里小蜜语音对话机器人的技术小二周伟（花名法一）聊了聊，这个能每天帮每位快递员打出数百个确认电话的语音助手，究竟是何方神圣。

机器之心：阿里小蜜为什么想要做语音对话机器人？

语音是阿里小蜜与用户接触的一个新的渠道，填补了过去机器人在通过电话直接与用户发生点对点交互的空白。语音对话机器人和基于文字的在线渠道、短信渠道、电话留言渠道等一起，建立了小蜜家族和用户接触的渠道闭环。

我们和菜鸟物流合作的这个帮快递员做派送前沟通的机器人，就是把可重复的、目的明确的、可以流程化的内容变成机器人的方式的交互。目前，机器人能够覆盖的业务包括信息/进度/流程通知、信息收集、问答咨询以及情感安抚。

从效果上来讲，对于快递员来说，它节省了打电话的时间和潜在的投递工作量；对于收件人来说，明确了投递的时间和地点可以提升体验；对于快递公司来说，终端服务质量可以得到保证。

机器之心：语音对话机器人还服务于其他哪些体系与场景？如何衡量系统整体完成任务的情况？

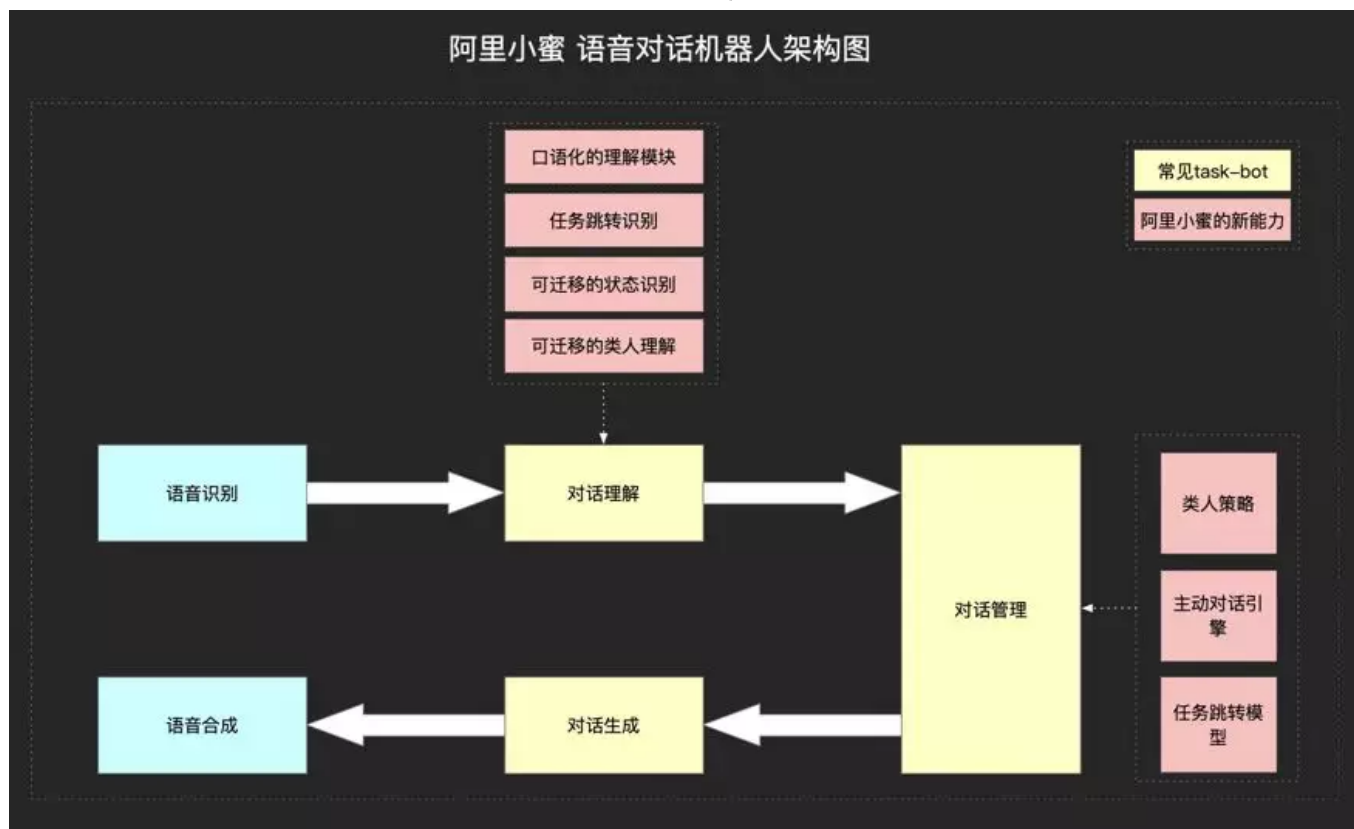
除了菜鸟物流之外，大家电派前的电联，盒马生鲜的派前电联，也都是基于阿里小蜜的语音对话机器人的产品。

对话系统的整体评估是非常复杂的，根据系统不同的阶段可以分为对自然语言理解（NLU），对话状态追踪（dst），对话策略（dialogue policy），自然语言生成（NLG）的评估；根据粒度又可以按照整通对话、每轮对话等进行评估，这些指标好坏能够影响最终的效果，但是如果评估整体任务的完成情况，目前用的比较多的还是对话完成率，对话轮数，对话满意率等。

系统的呼出完成率平均可以达到 87%。也就是说，每拨打 100 通电话，平均有 87 通可以收集到我们想要的信息。剩下的 13 通可能由于种种原因，比如用户提前挂断，没有收集到时间、地点等所需信息。

另一些技术人员对系统的衡量指标包括垂域语音识别正确率（ASR）95% 等。此外，还有不同场景下的适配覆盖速度：将语音机器人部署到一个新领域可能只需要几个小时乃至几十分钟。

机器之心：语音对话机器人主要有哪些模块组成？



按照顺序主要有语音识别、对话理解、对话管理、对话生成和语音合成五个步骤。

语音识别的输入是语音，输出是文本；对话理解的根据用户的自然语言文本以及其他一些特征做多模态的用户意图识别；对话管理根据用户意图做对话状态追踪，然后确定对话策略；对话生成模块可以用问答模版/半检索/纯生成等做法进行话术生成和拼接；以及最后用 KAN TTS 和传统 TTS 并列进行的语音合成。其中 ASR 和 TTS 是由达摩院智能语音实验室提供技术，与阿里小蜜团队一起合作完成的。

机器之心：语音识别模块的效果有哪些衡量指标吗？

目前 ASR 在特定垂类的准确率可以做到 95% 以上。之所以强调垂类，是由于声学模型将声音翻译成特定的音节，会受到地域，特定领域的专业术语等影响，语言模型也会受到专业术语的影响。语音识别中负责解决同音字问题的语言模型需要学习在真实场景里能够经常遇到的词的组合。

机器之心：垂类如何划分？

垂类这个概念可大可小，划分是由领域之间的共享性决定的。两个应用场景用同样的垂类模型还是不同的垂类模型取决于二者的特性。例如，给菜鸟物流这个垂类训练的模型也可以用于大家电派送前电联。两个场景有很大的相似性，用物流领域的模型识别大家电配送的效果已经很好了，就没有必要再收集大家电领域的数据进行训练了。

机器之心：对话理解模块相比于其他的对话机器人有哪些特点？

从任务定位上来说，阿里小蜜的对话机器人与传统的任务驱动型对话机器人的一个主要差别在于，它更倾向于进行「主动对话」（Proactive dialogue）而非「被动回答」。这个特性是由我们设计语音对话机器人的目标，即构建和人比较类似的纯语音交互的机器人，所决定的。我们希望机器人不仅能正确地回答问题，也能和用户更好地交互下去。

举个例子来说，A 问 B「吃了没？」，B 如果只能回答「没吃」，对话就容易陷入沉默和尴尬，如果 B 可以反问 A「你吃了吗？」，再根据 A 的回答继续展开对话，相当于 B 掌握了对话的主动权。这样的主动对话是一种类人的交互能力，而这样的类人交互是建立在类人理解的基础上的，我们的对话理解模块除了任务型对话需要进行的槽位（slot）识别、意图理解之外，还会额外增加一些类人的显式或者隐式的意图理解。

一个显式意图的例子是，用户明确地打断了机器人：表达了不想继续听下去。一个隐式意图的例子是，用户说「这个快递怎么还没到」，那么在给用户提供快递的预计送达时间等信息前，我们首先判断出用户在表达愤怒，因此选择先致歉，「不好意思，没能及时送达给您带来了困扰」，再沟通信息，「您的快递预计会在两点到达」。这样的用户体验就会比单纯提供信息好很多。

机器之心：对话理解模块应用到了哪些模型？如何进行模型选型？

这里面在不同的任务上，我们用到了不同模型，包括分类、序列标注、排序、相似性计算 等等。

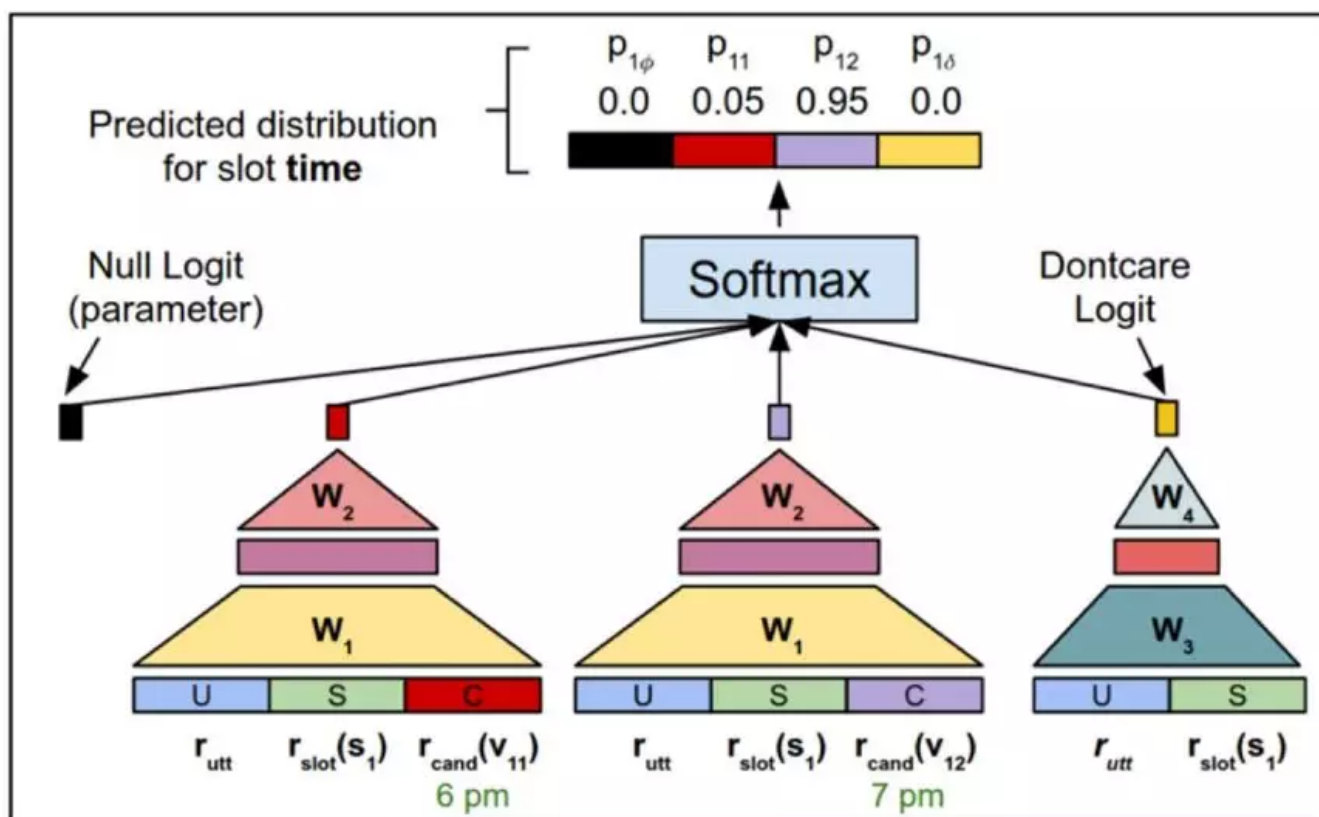
基础模型在进行模型选型时，主要考虑到模型的可迁移性、可扩展性，以及根据电话语言口语化的特点，在模型设计上进行了有针对性的调整。

比方说，多分类模型的问题是每增加一个类别就要重新训练一次，二分类模型的问题是每增加一个类别就要补充大量新语料，且不同分类之间量纲并不可比。两种做法在业务扩展较快，流程变化大的场景下都存在流程过长的的问题。所以我们格外关注迁移学习和多任务学习，让识别不同意图的模型共享底层结构，以便共享过去学到的语言基础表征，极大减少了新增意图对语料的需求。另一方面，将语言统一表征到同一个度量空间中，就可以通过在度量空间中定义的相似性函数实现不同意图之间的可比。

此外，语音端的聊天机器人接收到的信息都非常口语化。口语化意味着片段化，一句话会分成好几段说完，还会包含大量显式或隐式的指代关系。因此需要模型具有跨句子的指代关系识别和歧义消解。

机器之心：对话管理模块由哪些模型组成？模型选型过程中都有哪些考量？

对话管理由对话状态追踪（DST）和对话策略（Dialogue Policy）模块组成，在电话场景中主要由触发模型和组合模型构成。



上图是一个多领域可迁移的对话状态追踪（DST）模型中的例子。上图中的输入里，U 是用户说的一句话，S 是槽位，在这里是时间，C 是不同的时间点，例如六点、七点。

模型最后的目的是，给定用户说的一句话，判断派送时间是六点（ P_{11} ）、七点（ P_{12} ）以及句中没提到准确时间（ $P_{1\delta}$ ）、或者此句与时间无关（ $P_{1\phi}$ ）的概率。

这个模型具有非常好的可扩展性：无论 C 的取值是什么，模型都用相同的参数（ W_1 , W_2 ）来在底层进行相似度度量。

对话策略模型的触发模型决定是否要说、说什么，组合模型决定如何将说的不同内容按照先后顺序组合在一起。

传统的对话管理模块通常有两种模式，一是根据规则，规则匹配到了什么就说什么；二是根据槽位完成情况，缺少什么状态就继续对该槽位提问。我们的对话管理模块要完成的任务更多：是否进行安

抚、是否进行引导，什么时候该沉默，什么时候要强化等等，都要组合在一起，对每一个状态都随时进行检测，但不一定每回都会选择说出来。

对话策略部分的最终目标是让对话更好地进行下去，监督学习是常见的做法，对单个策略进行预测和对策略组合进行预测都是常见的方法。我们还尝试了用强化学习的做法，因为「任务效果」是一个很难在一句话结束时立刻得到反馈的指标，我们只能在一段对话结束后，把用户整体的满意程度作为奖励进行强化学习。

我们的系统从过去的单目标导向的对话策略转化为多目标导向的对话策略：除了要完成任务，还要把任务完成「好」，系统对于「好」有多个定义，对应多个指标，在每一个任务上进行精调。

机器之心：对话生成模块的主要职责是什么？

对话生成模块和经典的机器人对话生成模块基本一致：将系统的决策变成和用户交互的口语化描述。模块把上一轮选择出来的话术进行拼接和修改。拼接主要考虑先后顺序。修改主要兼顾连贯性以及一定的多样性。

机器之心：语音合成模块使用了哪种模型？

目前市面上主流的商用语音合成产品和服务，绝大多数都使用传统 TTS 框架构建，传统框架的问题是，用户往往很容易听出合成语音的机械感。达摩院智能语音实验室的提出的 KAN TTS 在传统语音合成系统的基础上，充分利用了领域知识，构建了表现力、稳定性都更高的在线中文实时语音合成系统。

目前线上语音合成模块采用了传统的 TTS 和 KAN TTS 的结合。相比于传统做法，KAN TTS 在准确度上有一个很大的提升，和真人语音的相似度由 88% 左右提高到 95% 以上。接下来我们会将 KAN TTS 全面应用到线上。


机器之心：是否有部分计算可以在和用户打电话之前完成？

是的，为了节省线上的计算性能，并尽可能优化响应时间，我们会把能提前算好的部分都提前计算。这部分包括一些通用的知识的表示以及对用户的表示。而现场出现的、之前没说过的内容，必须要现场经过编码器模型编码。

机器之心：大部分用户对语音机器人的了解还停留在菜单式的层面。在引入如此多新技术之后，机器人已经能做到「以假乱真」，那么是否有必要明确向用户传达「现在进行对话的是机器人而不是真人」？

人」这样的信息呢？

在其他一些国家，有法律明确规定机器人在打电话时必须率先声明自己的机器人身份的要求。

从技术角度出发，我们的目标是为了给用户更好的交互体验和类人体体验，因此，我们希望让机器人的声音和交互过程尽可能像人。当然，我们认为一个语音机器人，主动的表明自己的身份是一个机器人也是必要的。

本文为机器之心原创，转载请联系本公众号获得授权。

✂-----

加入机器之心（全职记者 / 实习生）：hr@jiqizhixin.com

投稿或寻求报道：content@jiqizhixin.com

广告 & 商务合作：bd@jiqizhixin.com