在线呼叫中心的智能客服问答机器人

张心泽

2017 年 7 月 5 日

前言

呼叫中心（Call Center）又称为客户服务中心，是基于 CTI 技术[①]，利用互联网、电信通信网络和计算机网络的多像功能集成，并与企业连为一体的完整的综合信息服务系统。早期的呼叫中心依赖电话机或排队机，实现客户电话的纯人工接听。随着计算机技术、通信技术和电子商务模式的发展，呼叫中心呈现多媒体化、分布式的发展趋势。现今的呼叫中心架构于 IP 协议与计算机网络之上，将之前的电话呼叫转变为包括邮件、电话和基于 Web2.0 技术的在线文本交谈等多种形式，成为包含通信、计算、管理和业务支撑特性的全媒体交互中心[1]。

然而，新的技术和商业模式在给企业创造利润、引领企业进行销售与服务模式变革的同时，也给企业带来了新的挑战。以阿里巴巴淘系为例，包含淘宝、咸鱼等在内的电商平台具有亿量级的在线商品库，用户数同样在数亿规模。每天淘系平台会产生包含购物、物流和售后咨询在内的海量服务请求。即便具有现代的呼叫中心，简单但繁复的问题咨询仍给企业带来了庞大的人工成本。

因此本文从呼叫中心的核心功能即客服问答出发，结合机器学习等人工智能技术、方法、模型和算法，分析和设计一种智能客服问答机器人（以下简称智能客服），以帮助企业降低人工客服工作量和提高服务资源使用效率。

在此本文题目要求为：“呼叫中心的 AI 接线员的系统的设计与实现”，并有以下三点具体要求：

* 呼叫中心的 AI 接线员的系统的设计与实现

(1) 调研目前国内外最好的三家（款）系统。对比它们的功能、性能和可用性等相关指标，指出它们的优缺点；

(2) 完成总结报告和汇报 PPT，解释其主要工作原理和实现方式（如何训练模型和算法、如何实现应用），附演示视频；

(3) 提出具体研究和实施方案，即如何用机器学习等人工智能的技术、方法、模型和算法来实现面向特定领域的对话机器人（如替代电信 114 接线员）。

[①] CTI 技术传统的定义为“计算机电话集成”，即 Computer Telephony Integration。随着电信通信技术的发展，现在的定义为“计算机电信集成”，即 Computer Telecommunication Integration。

基于时间因素和论文篇幅，本文结合实际情况对要求进行了适当松弛，并对相关概念进行了界定。

基于 Web2.0 技术的在线文本交谈已成为呼叫中心客服与用户交互的主要形式。虽然电话和邮件等形式通过语音识别、分句等处理也可转化为文本交谈，但其属于另外相对独立的研究领域，不在本文研究范畴。因此本文并不考虑呼叫中心的电话、邮件等交互形式，而以在线文本交谈为情景，设计和实现智能客服问答机器人即“AI 接线员”。

呼叫中心的智能客服目前尚处发展阶段，缺乏统一和公认的评价体系。因此本文无法界定“最好”，在此通过经验确定了相关评价指标，挑选和对比了国内外三家具有代表性的此类系统。

由于市面上具有代表性的智能客服系统尚未开源，因此无法准确得知其原理细节和实现方式。在此，本文通过官网展示或演示视频，结合现有研究，对其工作原因和实现方式进行合理推测。

综上所述，本文以呼叫中心在线文本交谈为情景，调研了国内外三家具有代表性的客服智能问答系统，建立了相关评价指标进行对比，推断了其主要工作原理和实现方式，并具体提出了一种面向客服领域的智能问答机器人。

1 介绍

从早起的图书情报检索系统、电信 114 服务系统到现在的在线呼叫中心，快速并准确的获取与反馈信息一直是用户和企业的追求目标。伴随而生的智能问答系统已有近 70 年的发展历史。早期的智能问答系统大多数针对特定问题设计，且由于技术和环境的限制数据量十分有限，不易进行扩展和训练，如 Green Jr et al.[2] 和 Woods[3]。这些诞生在上世界六七十年代的智能问答系统仅接受特定形式的自然语言语句，且供系统训练的数据也很少，无法进行较广范围的问答从而未被广泛使用。

进入九十年代后，借助互联网技术的发展，大量可供训练的问答对在网上可被搜索和爬取，进而构建特定主题下的语料库[4]。这些语料库的出现极大促进了智能问答系统的发展，研究人员在这些语料库上训练和测试各种问答模型，先后提出了基于逻辑推理[5]、基于模式匹配[6] 和基于机器学习[7] 等许多方法。在此阶段，人们主要利用信息检索或浅层语义理解去候选应答集中寻找应答从而构建智能问答系统，故将此类系统归纳为检索式问答系统。但检索式问答存在固有缺陷。应答的准确与否很大程度上取决于当前问句的信息充分程度，一方面，不能很好的关联该问句的上下；另一方面，问答系统的训练依赖当前问句抽取后表述的标签或是关联规则。因此检索式问答系统对短问句的应答效果欠佳，且规则构建仍需要较高的人力或专家知识。

随着深度学习方法在学界和业界均取得了较好的效果，研究人员将端到端的思想应用在了智能问答系统，面向不同特定领域提出了基于端到端的统计机器翻译[8]、机器阅读理解[9] 和机器谈判代理[10] 等许多模型。在此阶段，人们利用词嵌入（Word Embedding）、编-解码

（Encode-Decode）和改进型循环神经网络（如 LSTM、GRUs）等方法生成应答文本，故将此类系统归纳为生成式问答系统。虽然生成式系统能够一定程度地解决长问句和问句的上下文理解难题，但生成式问答同样存在弊端。一方面，相较于基于规则或搜索的检索式系统，在处理已存在语义库和知识库的问答对时，可能需要上千次对话训练才能达到检索式问答系统几次简单设置（构建规则）的效果；另一方面，虽然生成式系统能够记忆上文甚至推演下文[10] 给用户一种在和人类对话的感觉，然而，这种模型很难训练，在进行长应答时很可能会犯语法错误。因此生成式系统相对检索式系统需要更多的对话训练数据。

目前，基于检索式或生成式的问答系统都可应用于对话机器人，但在面对开放领域和特定领域时，两者各有优劣。在面对开放领域时，典型如较纯粹聊天场景，此时用户的话题并不面向特定领域和具有特定目的或任务，希望得到类人的自然语言问答体验，因此生成式问答模型在面向开放领域即无任务驱动下具有天然的优势，此类对话机器人有微软“小冰”、Facebook“Messenger”和 Github“Hubot”等。而在面对特定领域，典型如呼叫中心，用户与对话机器人均面向特定领域和具有特定目的或任务，此时用户更希望对话机器人能够准确回答领域内问题。因此检索式问答模型在面向特定领域即任务驱动下更具有优势，此类对话机器人如面向报警的 Disrupt“911bot”[11]、面向在线客服的阿里“ALIME”[12]、京东“JIMI”[13] 和网易“七鱼”[14] 等。

近年来，基于 WordNet、HowNet 等词汇知识库和 Wikipedia 与电商数据这种动态更新的知识资源库，大规模知识图谱日益成熟；同时，基于统计机器学习的自然语言处理和基于深度学习的知识推理技术有了快速发展；此外，CUDA 加速计算的出现大大降低了机器学习与深度学习带来的庞大计算开销。这三方面的进步分别为智能问答系统的发展奠定了资源、技术和成本基础，给智能问答系统的发展带来了新的契机。

值得关注的是，随着深度学习的浪潮，使用深度学习完成任务驱动下的问答模型成为具有技术优势的学界主流；而在具有资源优势的业界，以智能客服为代表的任务驱动问答模型却几乎都采用更实用的检索式问答模型。如何通过深度学习对接包含词典、规则和知识图谱在内的知识，使检索式问答模型与生成式问答模型巧妙融合，达到学界和业界优势互补，是当前厄待解决的问题。

本文基于在线呼叫中心面向客服领域，首先对比国内外具有代表性的呼叫中心智能客服系统，合理推测其工作原理和实现方式，并进行对比与评价；随后融合统计机器学习与知识，总结了一套“生成-检索-生成”流程，提出一种混合式的端到端智能问答系统。

2 业界智能客服系统现状

本文基于在线呼叫中心面向客服领域，选择了当前具有代表性的三家（款）呼叫中心智能客服系统，阿里“ALIME”、京东“JIMI”和网易“七鱼”进行分析、对比和评价。由于包括这三家在内的业内主要智能客服问答系统均为开源，因此本文根据其网站介绍、演示 Demo 和使用说明对其工作原理和实现方式进行合理推测，并建立一套评价标准，对其进行

对比和评价。

2.1 ALIME

阿里“ALIME”是阿里巴巴推出的在线呼叫中心客服对话机器人平台，其核心场景为：在商家促销活动时（如“年中 6 • 18”、“双 11”等）人工客服资源紧张，无法顾及到所有顾客，此时，若顾客无法及时得到关于商品信息的应答，则销售机会转瞬即逝，给商家带来损失；在平常时段，大型商家具有固定的工作时间，往往面临夜间无人或缺人值班的情况，而小商家则一般身兼数职，分身乏术。因此阿里针对此类场景提出了检索式“ALIME”客服对话机器人平台。

“ALIME”平台系统由“千牛店小蜜”和“阿里小蜜”两部分组成，完整商业版于 2017 年 6 月正式上线。其中“千牛店小蜜”面向淘宝商家，支持所有淘宝和天猫店铺，是阿里商家版智能客服机器人；“阿里小蜜”面向淘宝顾客，支持所有淘宝顾客，是阿里顾客版智能客服机器人。

“ALIME-千牛店小蜜”（以下简称“店小蜜”）目前主要的功能有：识别顾客购物意图、判断顾客购物缺失信息、查询顾客购物信息和一定时限内上下文识别等。具体产品功能和知识库见表1和表2[②] 。

表 1: “ALIME-千牛店小蜜”产品能力

售前服务	订单服务	售后服务
回复询价	回复发货时间	回复退货流程
回复询单	确认发货快递	回复退货事项
关联商品回复	确认订单修改	—
解决包邮议价	—	—
回复商品优惠	—	—
回复活动内容	—	—

表 2: “ALIME-千牛店小蜜”电商知识库

通用知识点	行业知识点
基础商品问题	手机行业
活动优惠问题	服务行业
下单付款问题	鞋类行业
商品物流问题	零食行业
售后退款问题	...
店铺服务问题	...
聊天互动问题	...

本小节以图1为例详细阐述“ALIME”对话过程。图1中展示了“ALIME”的两个对话案例。其中，图1a中为“店小蜜”在服装类淘宝商家“森马”中的应用，即“森小蜜”。由图可知，顾客：“这件衣服身高 160 选多大码合适”[③]，“森小蜜”：“请问体重是多少公斤？”，顾客：“50”，“森小蜜”：“身高 160.0 公分，体重 50.0 公斤，..., 推荐 [修身 M，宽松 L] 码...”；图1b为“阿里小蜜”在顾客查询中的应用。

- 问答系统首先接收到顾客的输入“这件衣服身高 160 选多大码合适”。

执行步骤①：结合上下文根据语义分析模型对该语句进行分词和关键词抽取。

[②] 其中，“—”表示无，“...”表示省略。

[③] 结合后文，这里的“衣服”应该在之前对话中进行过明确。

(a) “千牛店小蜜”案例 (b) “阿里小蜜”案例

图 1: 阿里“ALIME”案例

从该句中抽取关键词“衣服”、“身高”-“160”、“多大”、“码”。其中，根据“多大”判定该语句为疑问句；根据“衣服”和上文关联，确定该语句中“衣服”的目标；根据“键-值”关系“身高-160”和“码”，确定“衣服”的部分商品信息。综合上述抽取和分析过程，确定该顾客的意图，即：连接该商家电商服装知识库，查询在“衣服”等于当前衣服、“身高”等于 160 时的码数，完成识别购物意图功能。

执行步骤②：执行顾客意图，进行检索操作。

此时，根据电商知识库的返回为“衣服”、“身高”、“体重”和“码数”的组合，返回值不唯一，体现在“码数”和“体重”的返回数量上，因此判定缺失“体重”信息。

执行步骤③：询问顾客购物缺失信息。

向顾客提问“体重”信息，以补全购物信息，以获取缺失查询条件“体重”。

- 问答系统抛出应答“请问体重是多少公斤”。
- 问答系统接受顾客输入“50”。

执行步骤①：结合上下文根据语义分析模型对该语句进行分词和关键词抽取。

获取到体重信息“键-值”关系“体重:公斤-50”[④]，得到完整顾客信息查询条件。

执行步骤②：执行顾客意图，进行检索操作。

此时，根据电商知识库的返回为“衣服”、“身高”、“体重”和“码数”的组合，返回值

[④] “键-值”关系 $pair\{key-value\}$ 实际是 $pair\{index-value\}$ 关系，即索引-值关系。因此键值有多种形式，如上文中身高信息的“键-值”关系为“身高-160”，此处体重信息的“键-值”关系为“体重:公斤-50”等。

唯一，获得“码数”信息。

执行步骤④：返回顾客购物缺失信息。

- 问答系统抛出应答“身高 160.0 公分，体重 50.0 公斤，..., 推荐 [修身 M，宽松 L] 码...”。

这里的问答系统应带内容中虽然带有“推荐”字样，但实质将查询结果 [修身 M，宽松 L] 带入到应答模版“身高 * 公分，体重 * 公斤，..., 推荐 [*] 码...”，并非推荐系统的推荐结果。

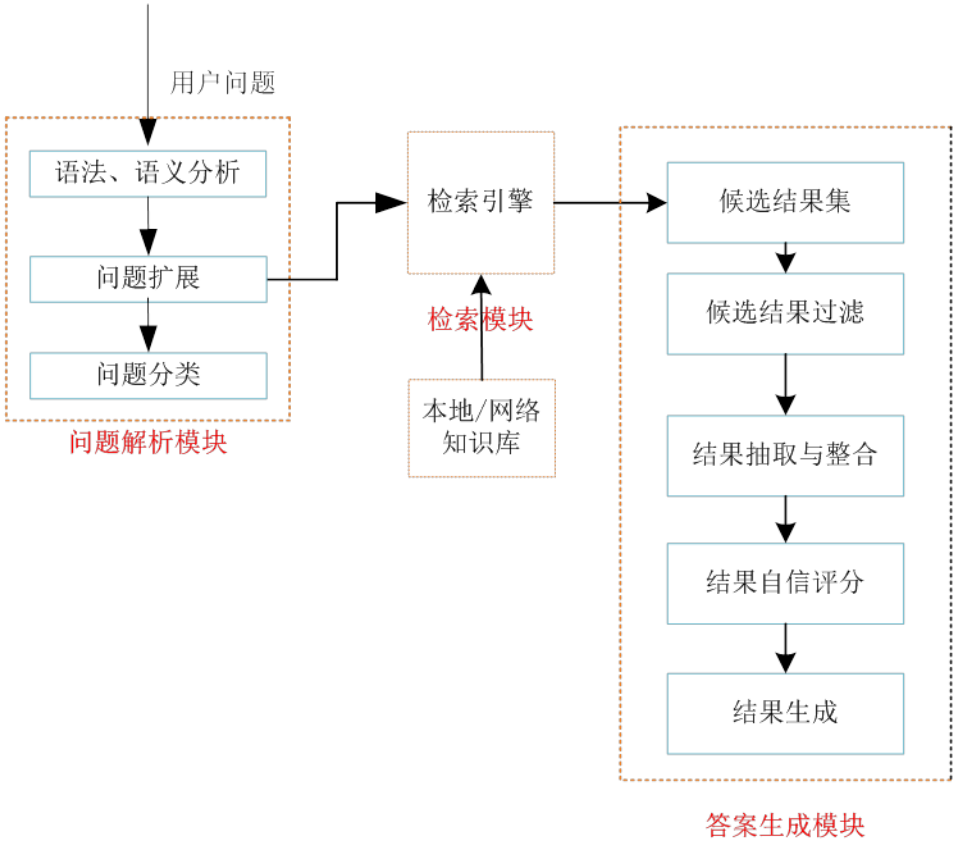

图 2: 问答系统流程

现阶段业界普遍采用的问答系统流程见图2。

2.2 JIMI

京东“JIMI”（JD Instant Messaging intelligence）是京东推出的检索式智能客服机器人，旨在为京东顾客提供更好的购物和咨询体验。其应用场景与阿里“ALIME”相同，但比“ALIME”上线更早，于 2013 年 3 月通过测试上线，具体过程见表3，功能见表4。

“JIMI”和“ALIME”均为检索式智能客服机器人，通过表5和图3 可推测出“JIMI”的工作原理和实现方式与“ALIME”类似。根据表5中“提问完整”技巧可知，“JIMI”相比“ALIME”缺失了一定的上文关联能力，且对输入要求更加苛刻。

表 3: “JIMI”主要上线事件

日期	事件
2013 年 3 月 1 日	JIMI 通过测试正式上线
2013 年 11 月 1 日	JIMI 登上促销活动页面
2013 年 8 月 7 日	服装类商家店铺 JIMI 上线
2013 年 8 月 21 日	3C 全部单品页 JIMI 上线

表 4: “JIMI”产品能力

售前服务	售后服务	闲聊服务
查询订单	回复退换货流程	笑话
取消订单	回复退换货事项	对诗
回复商品优惠	—	天气查询
回复活动内容	—	—

表 5: “JIMI”提问技巧

技巧	错误举例	描述
省略问候语	“您好”、“请问一下”、“在吗”等	不需要添加问候语 直接陈述问题
问题简洁	“我的订单不想要了，可以吗”	只需问“我要取消订单” 避免冗长
提问完整	“怎么解决呢”	“地址错了，怎么解决”或“地址错了，怎么办” 避免一个问题两次发送
单次提问	“您好，我京东的注册密码及支付密码忘记了，请问该怎么找回呢”	“如何找回京东注册密码”、“支付密码忘记，怎么办” 避免一个问题包含两个提问
减少错字	“优汇券”、“那个商品有活动”	“优惠卷”、“哪个商品有活动” 避免错别字
问题引导	-	输入关键词后，系统会匹配一系列相关的问题 可以直接点击相似或同类问题，直接获取答案

功能上，“JIMI”比“ALIME”多出了包括“笑话”、“对诗”和“天气查询”的闲聊服务功能。首先“天气查询”本质上并不属于聊天的范畴而属于较为简单不涉及复杂语义分析的检索功能；其次，通过实际测试，如图3可知，“JIMI”所提供的“笑话”和“对诗”能力亦不属于生成式对话系统。“前者”需要顾客输入关键词“笑话”，系统随后从笑话知识库中随机抽取一段笑话抛出；“后者”甚至不能识别一些经典诗句，以致系统完全误解顾客意图。因此“JIMI”的闲聊服务功能并不出彩，甚至给顾客造成了预期落差，产生了消极的对话体验。

虽然“JIMI”在对话窗体底部提供了“评价晒单”、“京东会员”、“交易纠纷”和“京东售后”等查询入口，相比“ALIME”给顾客带来了一定程度的便利。但这些入口严格意义上并不属于本文所研究的客服问答系统，因此不归为京东“JIMI”的优势功能。

图 3: 京东“JIMI”案例

2.3 七鱼

网易“七鱼”是网易研发服务企业的检索式智能问答系统。其应用场景亦与“ALIME”和“JIMI”类似，于 2016 年 4 月上线。“七鱼”与“ALIME”、“JIMI”同为检索式问答系统。通过企业自建的知识库和相似词库完成关键词抽取和语义识别，如图4所示。

图 4: 网易“七鱼”知识库

网易“七鱼”作为国内智能客服问答领域新生产品的代表，同样采取问题-知识检索的

方式构建智能问答系统。相较于“ALIME”、“JIMI”，“七鱼”采用开放网站、APP 和微信接口的方式获取企业用户[⑤]，然而这种渠道优势如同“JIMI”查询入口亦不属于问答系统的原理范畴，因此同样不归为其优势功能。

2.4 评价指标

目前由美国国家标准技术句主持的文本检索会议（TREC）针对 TREC/QA 任务下的问答系统提出了一套评价指标[15]。该套评价指标针对不同类别问题略有不同，总体由 3 个部分组成，实例查询准确率（Instance Precision）、实例查询召回率（Instance Recall）和 F 值（F1 Measures）。其中，实例查询准确率指的是问答系统给出的正确答案占给出的全部答案的比例，实例查询召回率指的是问答系统给出的正确答案的数量占所有正确答案的比例，F 值则是实例查询准确率和实例查询召回率的调和平均值。然而这套评价指标并不完全符合客服问答系统。

企业作为客服问答系统的使用者，核心需求为保持和扩展顾客群体，进而提高盈利。当然，反馈给顾客正确的应答毫无疑问为非常重要的问答系统评价指标，然而在企业的需求下，在无法给予顾客准确应答的情况时，给予顾客以适当的情感抚慰，也有可能挽留顾客并满足顾客的情感需求。顾客作为客服问答系统的参与主体，不局限于答案准确率，同时希望在快速获取准确答案的同时感受类人的交互。因此应当考虑建立时间分数和智能客服类人程度的评价标准。

时间分数是客服回答机器人与顾客对话时间的分数。顾客和企业都希望能够以尽量短的时间得到（给出）正确的答案。此处的时间可定义为对话时长或交互次数。对话时长目前主要受客户打字速度或其他因素的影响，若采用对话时长作为时间的评价标准，则会产生相比交互次数更大的误差，因此以在顾客最终得到正确答案时客户回答机器人与顾客对话的交互次数作为时间分数的影响因素。在顾客得到正确答案时，交互次数越少，时间分数越高，反之，交互次数越多，时间分数越低；在顾客无法得到正确答案时，时间分数为 0。

知名的机器类人程度评价有“图灵测试”。若在智能客服问答系统中建立图灵测试，则需要取消问答系统中转接人工客服的功能并对顾客进行一定的误导以使顾客无法先验客服身份，但此方法代价颇高。由于自然语言的天然复杂性，现阶段的客服问答系统均无法达到较高的人类自然语言相似度。机器无法理解自然语言表述导致无法抛出正确应答，顾客又无法转接人工客服，极易使顾客失去耐心和信任，造成顾客流失，而这正是企业最不愿发生的。因此图灵测试不适合直接应用与智能客服的类人评价。具体如何建立一套合理和有效的类人评价方法，值得学界和业界的继续探讨。

综上所述，本文借鉴 TREC/QA 评价体系，根据智能客服问答系统情景综合考虑了问答系统的应答质量、时间和类人程度因素，提出了准确率、召回率、F 值、时间分数和类人程度五元的综合评价体系。其中，准确率为问答系统给出的正确答案占给出的全部答案的比例；召回率为问答系统给出的正确答案的数量占所有正确答案的比例；F 值为准确率和召回

[⑤] https://github.com/qiyukf

率的调和平均值；时间分数为问答系统与顾客交互次数的函数；类人程度是问答系统应答与人工客服应答的相似程度。

2.5 系统对比

由于缺乏相应训练集、测试集、知识库和完整评价指标，故此小节未完成...

3 混合智能问答系统

3.1 意图识别

3.2 端到端学习

参考文献

[1] 马晓军. 全媒体交互中心及其大数据分析的研究. 电信科学, 30(Z2):82–89, 2017.

[2] Bert F. Green Jr, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: an automatic question-answerer. pages 219–224. ACM, 1961.

[3] William A. Woods. Progress in natural language understanding: an application to lunar geology. pages 441–450. ACM, 1973.

[4] Hoa Trang Dang, Diane Kelly, and Jimmy J Lin. Overview of the trec 2007 question answering track. In Trec, volume 7, page 63, 2007.

[5] Dan I Moldovan and Vasile Rus. Logic form transformation of wordnet and its applicability to question answering. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pages 402–409. Association for Computational Linguistics, 2001.

[6] Martin M Soubbotin and Sergei M Soubbotin. Patterns of potential answer expressions as clues to the right answers. In TREC, 2001.

[7] Hui Yang and Tat-Seng Chua. The integration of lexical knowledge and external resources for question answering. In In the Proceedings of the Eleventh Text REtrieval Conference, 2002.

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[9] Chuanqi Tan, Furu Wei, Nan Yang, Weifeng Lv, and Ming Zhou. S-net: From answer extraction to answer generation for machine reading comprehension. arXiv preprint arXiv:1706.04815, 2017.

[10] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. arXiv preprint arXiv:1706.05125, 2017.

[11] Omri Klinger, Tomer Coreanu, and Hadar Landao. Facebook 911bot. https://techcrunch.com/2016/05/08/911bot-is-a-chat-bot-that-could-save-your-life/, 2016. (Accessed On 1/7/2017 20:43).

[12] 阿里巴巴. Alime. http://alixiaomi.com/. (Accessed On 1/7/2017 20:35).

[13] 京东尚科. Jd instant messaging intelligence. http://help.jd.com/o/help/question-1024.html, 2015. (Accessed On 1/7/2017 20:30).

[14] 网易. Qiyukf. https://qiyukf.com/, 2015. (Accessed On 1/7/2017 20:39).

[15] Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 200–207. ACM, 2000.