Computer Science and Application 计算机科学与应用, 2017, 7(4), 291-300
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74036

Key Technology of IQA Robot Based on Telecom Scene

Yaofeng Tu

Cloud Computing Institute, ZTE Corporation, Nanjing Jiangsu
Email: tu.yaofeng@zte.com.cn

Received: Apr. 1st, 2017; accepted: Apr. 14th, 2017; published: Apr. 19th, 2017

Abstract

Intelligent question answering system is a new type of information interaction for natural language understanding. With the development of intelligent question answering system, it will bring new human-computer interaction mode and new business pattern. Intelligent question answering system involves many fields such as Natural Language Processing, knowledge management, intelligent dialogue and so on. In this paper, take the scene of telecom service for instance, we discuss that the architecture of Intelligent Question Answering System based on natural language understanding is put forward, and the related technologies are analyzed deeply.

Keywords

Intelligent Question Answering, Natural Language Understanding, Machine Learning, Deep Earning

基于电信业务场景的智能问答机器人关键技术

屠要峰

中兴通讯，云计算研究院，江苏 南京
Email: tu.yaofeng@zte.com.cn

收稿日期：2017年4月1日；录用日期：2017年4月14日；发布日期：2017年4月19日

摘　要

智能问答系统[1]是一种针对自然语言理解的新型的信息交互方式。它的发展将带来新的人机交互模式，带来新的业务形态。智能问答系统涉及自然语言处理、知识管理、智能对话等多领域技术，本文以电信业务场景为例，提出了基于自然语言理解的智能问答系统架构，并对相关关键技术进行了深入分析。

文章引用: 屠要峰. 基于电信业务场景的智能问答机器人关键技术[J]. 计算机科学与应用, 2017, 7(4): 291-300.
https://doi.org/10.12677/csa.2017.74036

关键词

智能问答，自然语言理解，机器学习，深度学习

1. 引言

传统的客服中心以人工服务的呼叫中心为主，但随着移动互联网与智能手机的不断普及，社交渠道多元化和应用软件功能的不断丰富，传统客服面临着急剧增加的服务需求和更为碎片、多元化的客户服务场景，这导致传统客服中心陷入到现实的多方困境：第一，人工客服受时间影响，难以保证 7 * 24 小时全天候服务。第二，人工客服业务操作繁琐，响应速度和服务质量的一致性难以保证，容易影响客户满意度。第三，业务知识的更新速度加快，增加了对服务人员的培训成本。第四，为满足增大的业务量和提高并发度，需要增加服务人员或者加班时间，导致成本高昂，而且近年来人力成本不断的上升。同时，服务信息难以保存，难以进行及时有效的分析，导致知识无法共享，服务灵活性不足。

2. 智能问答系统发展机遇

艾媒咨询在 2015 年发布的报告显示：中国客服市场整体规模超过千亿人民币[2]。但在庞大的市场规模下，难掩的是多数传统企业客服部门正在陷入上述之高成本、低效率、需求碎片化、满意度低、订单转化率低等矛盾。传统客服面临的现实困境和巨大的市场容量，迫切需要智能化的客服产品来突破发展瓶颈。如何构建可商用的智能问答系统是运营商及客服厂家亟需解决的问题。

当前智能机器人技术不断发展和成熟，智能机器人被应用于金融、财务、客服工作等领域，其中，智能机器人在客服工作中的应用效果最为显著。它通过自动客服、智能营销、内容导航、智能语音控制等功能提高了企业客服服务水平[3]。

智能问答应用在客服工作中有着显而易见的优势。一是提高用户感知，为企业在线客服、新媒体客服等提供统一智能的自助服务支撑，减少了用户问题得到解决的难度和复杂度；二是提升服务效率，缩短咨询处理时限，分流传统人工客服压力，节省服务成本(据统计：智能机器人投入是人工座席成本的 10%)；三是收集用户诉求和行为数据，支撑产品迭代优化[3]。

目前国际上的智能问答技术主要采用检索技术、知识网络、深度学习这三大技术，代表性平台是苹果的 Siri、谷歌的 GoogleNow 和微软的 Cortana。国内的智能应答技术发展较晚，这和中文的语法、语义复杂性等多种因素有关,目前主要是以人工模板和智能检索技术为主，典型代表有小 i 机器人、百度度秘等。

为了促进传统客服形态向自动化、智能化、人性化、多渠道的方向演进，更好的支撑电信域的业务发展，推动智能客服在电信领域落地，本项目根据电信业务场景的特性与需求，提出了一种基于自然语言的交互模式，可通过引导式应答或者反问使得问题加以确认，分领域建立领域本体的知识库和问题应答库，采用推理技术进行问题的自动识别与应答，形成基于语义的智能问答机器人 IQA(下文简称 IQA)。

本文介绍了 IQA 系统的设计思想及架构，并对其中的关键技术进行了分析。IQA 已在电信、移动现网多个局点运营和实证，能快速准确识别用户意图，自组答案，大幅提升客服工作效率，在实践中检验

了技术路线的成效，并建立了丰富的电信领域语料库。

3. 智能问答机器人IQA设计思想

问答系统的目标是给定一个问题，能够得到简短、精确的答案。

IQA 是基于自然语言理解技术和知识图谱技术，并配合使用语音识别(ASR)、语音合成(TTS)等智能人机交互技术，通过微信、APP、网页、短信、电话等渠道，以文字、语音等方式提供智能问答交互服务的信息服务系统。

图 1 展示了 IQA 的设计和架构图，该系统分为应用层、接入层、分析层和数据层。各层之间相对独立，耦合度小，易于扩展。

智能应答系统首先要有数据，数据可以来自于互联网爬取，也可以是现有的知识库，或者特定的语料库，这就涉及到数据的来源、获取、挖掘、存储等设计。

- 数据源：大致可以分为三类：1、百度知道等社区问答对；2、电信领域内专业数据及网站语料；3、第三方提供的数据接口，比如天气、笑话等。
- 数据获取：针对以上三类数据源，对应的获取方式可以是垂直爬虫爬取、人工维护录入、数据厂家提供，以及第三方开放平台提供，比如中国气象网等。
- 数据挖掘：数据挖掘就是对所获取的数据进行挖掘成有用的信息或结构化信息，按一定的结构和规则来组织有用的信息和知识，最终形成各种语料库，比如问答对 FAQ、聊天对话库、各种分词、领域词等词典，以及通用和行业等知识库，以便于问答直接使用或进一步处理。
- 数据存储：存储方式至少包括四类：1、像各种词库等，可直接文本形式存储，为了管理方便也可以存在数据库中；2、对于问答对、训练数据等可采用关系数据库进行存储；3、对于知识库，可以采

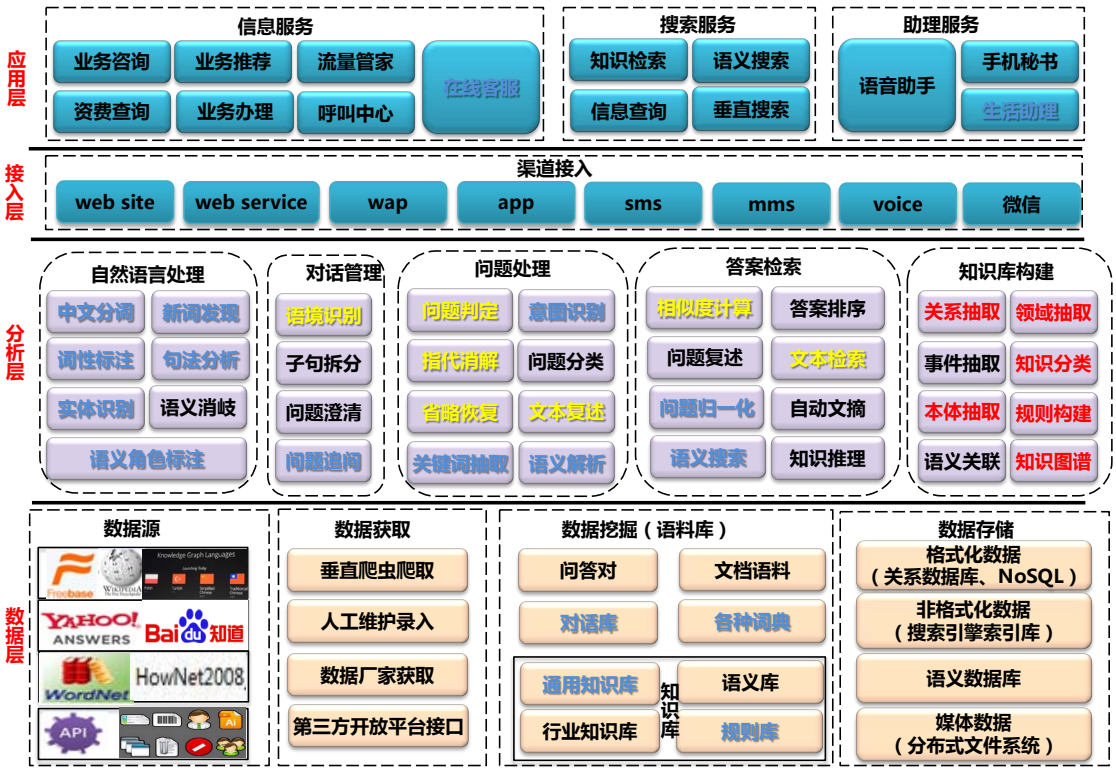

Figure 1. Robot IQA system architecture.
图 1. 智能应答机器人 IQA 系统架构

用语义数据库来存储；4、对于大数据量，可以采用分布式文件系统和 NoSQL 等。

有了结构化或者清洗后的数据支持，需要对数据进行分析，这里的分析主要分为两部分，一部分是对用户输入问题的分析，即问题语义理解，一部分是线下数据的分析，如知识库构建等。分析层是智能问答系统 IQA 的核心引擎，分为自然语言处理(预处理)、对话管理、问题语义理解、答案检索获取，以及知识库构建更新 5 大部分，在知识数据已具备的情况下，由这 5 个模块即可以组成一个问答系统。

- 自然语言处理 NLP：该模块属于预处理模块，主要对用户输入问题和知识库进行预处理，比如中文分词、词性标注，供后面的关键词提取使用；抽取语料库的实体和新词发现，获取领域词典；通过句法分析和语义角色标注，获取用户问题的主谓宾、施事受事等。
- 对话管理：该模块属于上下文的复杂问题处理：如问题一次输入多个问题，需要进行子句拆分；问句模糊无法理解，需要问题再次澄清；问题缺少必须元素，需要追问达到回答目的；问题涉及上下文语境的，需要语境识别等。
- 问题语义理解：首先对单个问题进行语义理解：如判定用户的是问题还是聊天；问的是哪类业务或类别问题，便于快速定位；识别用户问的意图，究竟是咨询还是购买等；对缺少的问题成分，根据上下文恢复成语义齐全的问题，便于检索答案。深层语义分析主要是理解问题的真正语义并处理复杂问题，将多个问题拆分，根据上下文进行缺省句恢复和意图理解，对于多种问法可以抽取语义规则，进行规则匹配，对检索的结果进行相似度计算，找出最佳答案。
- 答案检索获取：该模块主要为答案获取模块，如知识库为 FAQ,则需要对问题进行复述，或将问题归一化到 FAQ 库标准问题，在没有的情况下，可以直接根据关键词检索，返回的结果进行相似度计算，答案排序，最终返回答案；如为知识库，则需要进行语义检索，并进行一定的推理；如为文档，则需要进行自动文摘，找到答案。
- 知识库构建更新：构建知识库可以更好的组织知识，更快速准备检索答案，和 FAQ 相结合，使问答系统适用各种语料，而不仅仅局限于 FAQ，需要包括本体提取、领域词提取、关系提取、推理规则构建等功能。

IQA 为了方便用户使用，满足用户不同的使用习惯，需要有多种接入方式，主要包括：

- 语音：为了解放用户双手，支持语音接入交互，使问答更加智能。
- 微信：支持和微信平台接入，成为微信智能客服和问答机器人，便于用户低成本使用。
- 短信：在传统运营商有些业务依然会使用短信来提醒，为此，用户可以直接在此基础上进行回复，而无需记忆传统的短信代码，使用自然语言，提高用户使用体验。
- App：支持 app 接入，比如生活助理、吃喝玩乐等 app 应用中的智能问答。
- 平台：在设计时，需要考虑各个模块、算法和接口的松耦合性，做到可插可拔，黑盒复用，因此中间的分析层可以独立成问答平台，以便开放给第三方调用。

应用层是在智能问答核心技术的支撑下产生的各种应用和服务，不仅仅是问答，还可以是信息服务、搜索服务和助理等服务。

- 信息服务：可面向运营商或政企等在线客服场景，做业务咨询、业务办理、业务推荐等，也可和传统人工坐席相结合，在智能问答无法回答的情况下，再呼叫人工坐席。
- 搜索服务：我们的知识库支持专业的知识工程构建，因此可以提供知识检索、语义检索等，也可以作为智能检索，嵌入企业管理系统，作为垂直企业检索。
- 助理服务：生活助理、个人助理、语音助理也是智能问答的应用场景，为用户提供助理等服务。

此外，IQA 涉及大数据量的语料处理，需要支持离线分析。IQA 系统又是在线的高并发服务系统，需要支持实时访问、分布式扩展。

4. 智能问答系统关键技术

4.1. 知识图谱构建

智能问答的智能核心来自于强大的知识资源，这需要对大规模数据资源进行理解和抽取，转换成计算机可以处理的形式来表示和存储。

早期的知识库是把专家知识通过人工进行构建的，需要大量的人力和物力，这些知识库资源最典型的代表包括英文词汇知识库 WordNet、FrameNet、中文词汇知识库 HowNet 等。这些知识库基本属于通用领域的知识库。

通用领域知识库资源不能满足智能问答系统对知识资源的需求。针对电信业务场景，IQA 还需要构建专用领域知识库。

知识图谱构建主要分成三部分：核心信息抽取、知识库构建与更新、知识库可视化。其中核心信息抽取指从语料中抽取核心信息，如实体、实体类别、实体关系、实体属性等；知识库构建与更新指将核心信息抽取的结果以知识库的方式进行构建与更新；知识库可视化指知识库的展示，并提供一定的管理功能。

由于 IQA 的设计目标是电信场景的智能应答，所以使用了基于知识库和 FAQ 相融合的智能交互服务技术，将知识与知识的关联整合到系统和数据里面，使 IQA 更加智能。知识管理既包含了知识库，还有 FAQ 库等：

- 知识库：主要是进行知识库构建和更新，比如本体提取、本体关系提取，将领域知识构建成知识库专家系统，另外能进行一定的推理。
- FAQ 库：FAQ 主要是面向客服类语料，不能光靠检索和相似计算，也需要对 FAQ 库进行处理，如同一问题不同问答的语义规则提取、意图类别提取、问题类别提取等。
- 对话库：任何问答都离不开正常的对话交流，因此需要提供对话寒暄库来提高问答的用户体验性。
- 各种词典：存放问答需要的各种词典，如分词、同义词、领域词等。

4.1.1. 实体识别

命名实体识别包括对实体的识别及属性的抽取。实体识别是把文本中的实体划为某一语义类型。主要有三种方法，基于字典、基于统计与基于规则的方法。基于字典的方法这一方法较为简单，主要是通过字符串匹配找寻词库中命名实体，但是通常没有一个全面的实体库，而且比对费时。基于规则算法主要在实体识别过程中加入词法规则、语法规则、语义规则，然后通过规则匹配的方法识别各种类型的命名实体。基于规则方法受限于人工添加规则。基于统计的方法先建立语言模型，利用人工标注或原始语料进行训练，然后在训练数据上估算模型参数，这有利于移植到不同的语言及新领域。基于统计的方法主要利用一些统计模型如隐马尔可夫模型、最大熵模型、支持向量机、条件随机场(Conditional Random Fields，CRFs)等。

在电信领域，NER(命名实体)识别的目的为从语料中抽取出电信领域相关实体。例：“如何办理酒店**留言**优惠**套餐**？答：…”，其中加粗部分即为电信领域相关实体。

IQA 的实体识别结合了基于字典和基于统计的方法。一方面查找字典，一方面采用基于统计的方法，通过标注语料获得一定数量的已标注的 NER 数据，用于训练 NER 模型。然后对于给定的生文本语料，先进行文本预处理(分词、词性标注等)，然后使用训练好的 NER 模型进行 NER 识别，最终得到 NER 识别结果，如图 2 所示。

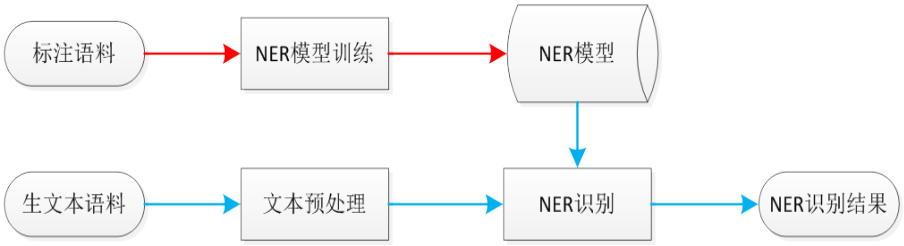

Figure 2. The process of Named Entity Recognition
图 2. 命名实体识别流程图

构建电信领域命名实体库是一个动态的过程。开始时针对电信业务手动建立较小规模的电信实体及属性词典，然后从电信文本及期刊网上获得电信领域的文献，通过建立的词典信息提取特征，使用统计模型进行训练，从而使模型在样本有限的情况下学习到新知识，将筛选出的元素加入词表中。随着项目的深入，数据的增多，则通过对大量数据的学习自动识别电信实体从而扩大命名实体库的规模。

属性抽取的任务是为每个实体语义类构造属性表并抽取出属性值。一方面对于一些数据如电信套餐业务类表格可以通过解析原始数据中的半结构化信息来获得属性和属性值，这种方法绕开了自然语言理解的阅读句子。而更多的知识隐藏在自然语言句子中，当前从句子中提取属性和属性值的基本手段是模式匹配[4]和对自然语言的浅层处理。

4.1.2. 关系抽取

由于知识库采用实体－属性－值的方式进行构建，知识库中的实体之间的关系有以下三种：

- 顺序关系：

在电信领域的问答中，用户对于实体的提问具有先后的次序关系，比如先问手机话费业务，再问短信业务。实体之间的顺序关系抓住用户提问的这种先后关系，建立实体间的顺序关系。利用这种顺序关系，解决自动问答系统中当用户提到一个实体时，为其自动推荐下一个实体。

- 共现关系：

在电信领域的文本中，不同的实体经常一起出现，具有比较强的关联关系，我们称之为共现关系。与顺序关系不同的是，共现关系不仅在用户提问中获取，也可以在其它文本中获取(比如业务说明文档等)，没有先后顺序关系，从某种意义上来讲，顺序关系是共现关系的一种。利用这种共现关系，当用户提到某一实体时，为其推荐关联度较大的若干个实体。

- 父子关系：

从电信领域的目录结构中获取实体之间的上下级关系。

关系抽取一般通过人工定义各类关系模型，然后通过模式匹配或者机器学习的方法进行抽取。一般领域知识内，关系模式相对比较固化，可以通过统计算法找到关系模式，再基于关系模板进行抽取。基于深度学习的抽取算法近来比较火热，不仅能减少人工标定的工作，还能够返现尚未人工定义的规则。

4.2. 问题理解

智能问答系统只有知识远远不够，还需要理解人提出的问题，将自然语言表述的问题转化为计算机可以理解的形式化语言。让计算机理解自然语言是非常困难的，也是人工智能领域最难处理的问题之一。

4.2.1. 领域分类

领域分类主要解决的是把用户提出的问题对应到相应的领域中，减少问题答案的范围，提升对问题的理解力，这是智能问答系统要解决的一个关键问题。在实际的问答系统中，领域知识可能会包含很多

种，如果分类错误，答案基本就是错误的。领域分类有词匹配法、规则方法、统计学方法等几种常见方法。

最早的词匹配法仅仅根据文档中是否出现了与类名相同的词来判断文档是否属于某个类别。很显然，这种过于简单的方法无法带来良好的分类效果。

规则方法是为每个类别定义大量的推理规则，如果一篇文档能满足这些推理规则，则可以判定属于该类别。由于在系统中加入了人为判断的因素，准确度比词匹配法大为提高。但是分类的质量严重依赖于规则的好坏，也就是依赖于制定规则的“人”的水平。规则方法可推广性差，一个针对金融领域构建的分类系统，则无法扩充到医疗或社会保险等相关领域，造成巨大的知识和资金浪费。

统计学习方法进行文本分类的一个重要前提是认为文档的内容与其中所包含的词有着必然的联系，同一类文档之间总存在多个共同的词，而不同类的文档所包含的词之间差异很大。而且不光是包含哪些词很重要，这些词出现的次数对分类也很重要。这一前提使得向量模型(VSM)成了适合文本分类问题的文档表示模型。在这种模型中，一篇文章被看作特征项集合来看，利用加权特征项构成向量进行文本表示，利用词频信息对文本特征进行加权。它实现起来比较简单，并且分类准确度也高，能够满足一般应用的要求。

然后根据 VSM 生成的描述文本特征向量，用例如朴素贝叶斯算法、kNN 算法或者 SVM 算法进行分类。上述方法各有各的优势[5]。kNN 算法具有简单、稳定、有效的特点，但时间复杂度较高，适用于文本训练集合规模较小的文本分类系统。朴素贝叶斯算法可应用到大规模文本集合中，具有方法简单、速度快、分类准确率高等优点，但由于朴素贝叶斯算法所基于的假设太过于严格，在实际应用并不完全符合理论中假设条件的情况下，其准确率会有一定程度的下降。因此，在类别数目较多或者类别之间相关性较小的情况下，该算法模型的分类性能才能达到最佳。SVM 是以 VC 维理论和结构风险最小化原则为基础的，克服了特征空间中的维数灾难问题，解决了小样本学习问题，可得到小样本条件下的全局最优解。相对于其它分类算法，稀疏、高维的数据对 SVM 算法基本没有影响，能够更好地体现文本数据的类别特征。

统计学习模型在实现文本分类中也存在明显的缺陷。统计学习模型没有利用上下文的关系来理解文本信息进行分类。VSM 中生成的描述文本特征向量，需要人为的统计并人为的做特征提取工作。统计学习模型泛化能力薄弱，在不同的场景下要进行人为的调整模型的特征。

为了解决上述统计学习模型的缺陷，我们设计了深度学习中的 RNN [6]模型来进行分类模型的实现，RNN 很好的利用了整个文本结构的信息并减少了人为固定特征提取的工作。加强了泛化能力的同时，在应用在类似领域的分类只需增加几个参数而已，可以大大减少人为的工作。

使用 RNN 模型实现我们要做的领域分类过程如图 3 所示，首先用对大量广泛语料和特定领域的语料进行分词处理，将分词后的语料代入 word2vec [7]中，得到词向量的训练模型。然后将我们要验证的特定领域语料代入词向量的训练模型，继而生成特定领域中每个句子所有词的词向量。最后将每个句子的词向量作为输入数据输入到 RNN 模型当中，根据我们要求的分类情况给出输出向量，进而对整个模型进行训练得到 RNN 的分类模型。将我们要验证的语句代入到 RNN 分类模型当中，判断出该语句是否属于特定领域。其中，word2vec 开源工具直接从网上下载获得。大量广泛的训练语料从网站上爬虫获得。

我们以聊天语料作为特定领域为例，根据设计的 RNN 神经网络，对聊天语料进行训练，然后用训练好的 RNN 分类模型对验证语料进行测试，判断出该语句是否属于聊天类别。

整个问题的难点在于如何用词向量来训练 RNN 模型，如何确定 RNN 模型中各个参数，如何根据分类的结果对句子是否为聊天语料进行判别。

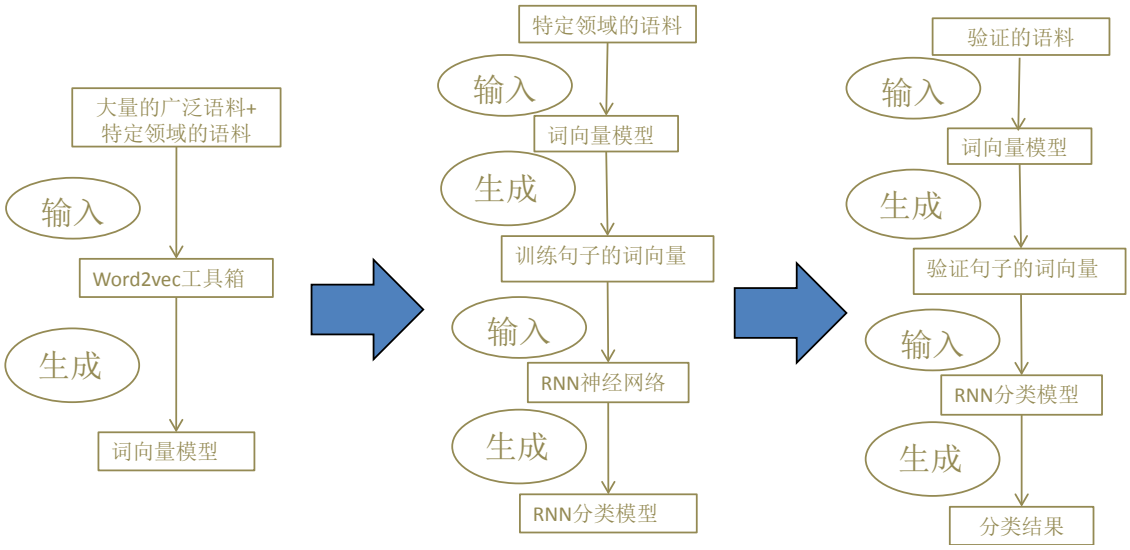

Figure 3. The process of domain classification with RNN
图 3. RNN 实现领域分类的过程

下面给出 RNN 的整个 BPTT 算法[8]：

输入：训练集 $D=\left\{\left(x_i^t,z_j\right)\right\},i=1,2,\cdots,I,j=1,2,\cdots,J,t=1,2,\cdots,T$

学习速率 η

过程：

1. 在(0, 1)内随机初始化网络中的所有连接权和偏移量。
2. Repeat。
3. for all $\left\{\left(x_i^t,z_j\right)\right\}$ do。
4. 求解中间变量 $\dfrac{\partial L}{\partial y_j^t}=\dfrac{\sum_{t=1}^{T}\dfrac{y_j^t}{T}-z_j}{T}$。
5. 更新输出层偏移量 $\Delta\theta_j=\eta\sum_{z=0}^{T}\dfrac{\partial L}{\partial y_j^{t-z}}$。
6. 更新隐藏层到输出层的权值 $\Delta w_{hj}=-\eta\sum_{z=0}^{T}\dfrac{\partial L}{\partial y_j^{t-z}}b_h^t$。
7. 更新隐藏层到隐藏层的权值 $\Delta p_{ih}=-\eta\sum_{z=0}^{T}b_i^{t-z-1}\left[1-\left(b_h^{t-z}\right)^2\right]\sum_{h=1}^{H}w_{hj}\dfrac{\partial L}{\partial y_j^{t-z}}$。
8. 更新隐藏层的偏移量 $\Delta r_h=\eta\sum_{z=0}^{T}\left[1-\left(b_h^t\right)^2\right]\sum_{h=1}^{H}w_{hj}\dfrac{\partial L}{\partial y_j^{t-z}}$。
9. 更新输入层到隐藏层的偏移量 $\Delta v_{ih}=-\eta\sum_{z=0}^{T}x_i^{t-z}\left[1-\left(b_h^{t-z}\right)^2\right]\sum_{h=1}^{H}w_{hj}\dfrac{\partial L}{\partial y_j^{t-z}}$。
10. until 停止条件。
11. end for。
12. 输出：各层的连接权和偏移量。

一般的停止条件设定为确定误差 $err=\sum_{z=0}^{T}\left(\frac{e^{y_j^t}}{\sum_{j=1}^{J}e^{y_j^t}}-z_j\right)^2<\varepsilon$，其中 ε 根据要求去人为给定。

当然上述给出的 BPTT 算法是基于 SGD(随机梯度下降法)去实现的，即每输入一个句子语料的词向量进行一次权值的更新。这样在以后对模型进行泛化能力的加强有更好的作用，也可以作为增强模型去训练。

4.2.2. 语境识别

交互式问答中问题语境检测可以看作一个二元分类问题，即判别用户提出的问题是与之前问题相关还是不相关。如果相关就可以认为是属于同一个语境，否则认为其属于不同语境。构造二元分类器主要问题就是找出能够有效进行问题相关检测的特征。目前研究来看，对分类有效的特征有：指代成分特征、线索词特征、最长公共词序列特征、名词特征等。利用以上特征就可以利用二元分类算法对当前的问题进行语境识别了。

指代成分特征：如果在问题中包含第三人称代词和指示代词，例如“他”“它的”“这些”等，并且这些指代成分不是指代本问题中出现的实体，那么指代成分只能指代之前问题或答案中出现的词语，所以问题是后继问题。这个特征通过构造的指代词表过滤分词后的问题获得，是布尔值特征，即当问句中包含指代成分时，特征值为 True，否则为 False。

线索词特征：线索词是指问题句中包含的如“其他的”“综上所述”“总之”这样的词汇，这些词汇提示了问题是与之前问题相关的。我们发现，尽管包含线索词的后继问题数量不是很多，但是线索词可以很准确地识别出后继问题。这个特征是通过线索词表过滤分词后的问题获得的，是布尔值特征，即当问句中包含线索词成分时，特征值为 True，否则为 False。

最长公共词序列特征：如果当前问题与之前问题中，包含有相同顺序出现的内容词，这样的问题通常为后继问题。例如：“问题 1：布什和戈尔的第一场辩论在哪所大学举行？问题 2：辩论是在什么时候举行？”，在问题 2 和问题 1 有相同的公共词序列“辩论举行”。最长公共词序列特征通过计算两个问题分词之后序列的最长公共字串获得，为减少计算复杂度，可以采用动态规划算法。最长公共词序列特征值是一个自然数集合。

名词特征：如果问题在去除了停用词、疑问词以及问题常用词后，没有包含任何名词，这样问题中没有提及任何具体信息，所以之前的问题一定提供了相关信息，则问题应与之前问题相关。通过对问题分词和词性标注处理，我们就可以获得这个特征值。这个特征值是一个布尔值，即当问句包含名词时，特征值为 True，否则为 False。

利用构建的基于以上 4 个特征二元分类器来进行交互式问答的问题相关检测，分类器采用标注了相关和不相关的 TRECQA 任务翻译成中文的问题集训练获得。训练二元分类器的方法是对于每个问题获取 7 维向量 $v_i=\left\langle v_i^1,v_i^2,v_i^3,v_i^4\right\rangle$，其中每个向量分量对应于指代成分特征、线索词特征、最长公共词序列特征、名词特征、相同实体特征、句法结构特征和内容词相关性特征。首先对问题集中每个问题进行分词、词性标注、句法分析以及命名实体识别的预处理，然后根据问题句预处理的结果，就可以获得指代成分特征、线索词特征、名词特征以及句法结构特征。而对于最长公共词序列特征、实体相同特征以及内容词相关性特征，采用经过预处理后的问题 q_i' 和之前的 n 个问题 $Q_n'=\{q_{i=1}',q_{i=2}',q_{i=3}',\cdots,q_{i=n}'\}$ 分别按照特征中描述的方法计算，取其最大值作为问题向量的相应分量，即问题 q_i 的 m 特征分量：$v_i^m=\max f\left(q_i',q_{i-1}'\right),1\le i\le n,m\in\{3,5,7\}$。根据训练集问题的问题向量以及问题相关和不相关标注，采用

C4.5 决策树的方法通过训练就可以获得问题相关检测的二元分类器。

智能问答系统涉及的关键技术点不限于以上所述，还有省略恢复[9]与指代消解[10]、相似度计算、智能对话管理等等，鉴于篇幅有限，不再展述。

5. 总结与展望

智能问答系统是目前人工智能和自然语言处理领域中一个备受关注并具有广泛发展前景的研究方向[1]。从行业发展方向来看，融入了人工智能技术的智能客服机器人成为了推动传统客服转型升级的变革力量。从技术现状和我们项目的实际应用经验来看，目前智能问答机器人 IQA 还需要借助于工程化的方法达到商用要求，例如关键词、等价句等配置使用等。通用的、高质量的智能应答系统，仍然是较长期的努力目标。随着大数据和深度学习技术的发展，基于知识和推理的深层方法和基于统计等“浅层”方法有机结合，未来智能问答系统将向着更智能化和人性化的方向发展。随着项目语料库的扩充和丰富，IQA 将在省略恢复技术点上使用深度学习的 LSTM 网络进行训练，尝试通过自动学习得到隐藏信息的网络结构；在领域分类技术点上，当 LSTM 训练模型的准确率积累到一定的程度时，摒弃线上分类规则，直接使用 LSTM 进行领域分类处理；问题归一化能力的强弱决定问答系统智能化程度的强弱，现在研究解决通用领域的问题归一化还有很多困难，但研究解决垂直领域的归一化还有很大的可能性。2016 年 9 月，谷歌提出了一种使用单个神经机器翻译(NMT)模型在多种语言之间进行翻译的简洁优雅的解决方案[11]，实现了机器翻译领域的重大突破。而问题归一化本质上也是一种机器翻译，可以尝试使用神经网络训练得到口语化到标准问题转换的模型网络。

参考文献 (References)

[1] 史忠植. 人工智能[M]. 北京: 机械工业出版社, 2001。

[2] http://money.163.com/16/1130/10/C745HHAB002580S6.html#from=keyscan

[3] http://finance.qq.com/a/20160302/004820.html

[4] 杜永萍, 黄萱菁, 吴立德. 模式学习在 QA 系统中的有效实现[J]. 计算机研究与发展, 2015, 43(3): 449-455.

[5] Russell, S.J., Norvig, P. 人工智能, 一种现代的方法[M]. 第 3 版. 北京: 清华大学出版社, 2013.

[6] http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns

[7] https://code.google.com/p/word2vec

[8] Graves, A. (2012) Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin.

[9] 李旺, 李绍滋. 基于 DRT 理论的汉语省略恢复研究[J]. 计算机工程, 2004, 30(17): 39-41.

[10] 宋洋, 王厚峰. 基于马尔可夫逻辑的中文零指代消解[J]. 计算机研究与发展,2015, 52(9): 2114-2122.

[11] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., *et al.* (2016) Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.