基于知识的自然语言问答发展综述

摘 要 问答系统中包括三个主要的部分：问题理解，信息检索和答案抽取。其中问题理解是问答系统的第一部分也是非常关键的一部分。首先对基于知识驱动的问答系统中的问句理解技术的发展进行了介绍，分别在问题分类、语义转化、关键词提取和查询扩展几个方面进行展开。此外，从基于语义分析的方法和基于信息检索的方法两大方面，介绍了近期基于知识库的自动问答系统的研究方法。最后总结了目前基于知识的自然语言问答研究存在的主要问题及未来发展方向。

关键词：问题理解，问答系统，知识库

1 引言

自然语言理解是研究如何让电脑读懂人类语言的一门技术，是自然语言处理技术中最困难的一项。问答系统（Question Answering System，QA）是新一代智能搜索引擎，它允许用户以自然语言提问，系统自动理解用户的问题并能够向用户返回准确的答案。通常，问答系统包含了问题理解、信息检索以及答案抽取这三个处理过程。[1] 问题理解是问答系统中的一个子过程，一个问答系统的处理流程大致如下图1所示：

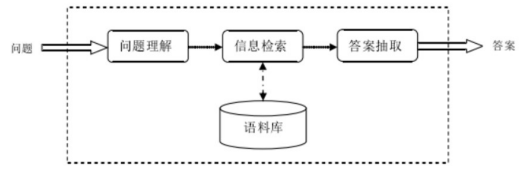

图 1

其中，信息检索过程所用到的语料库是人工建立起来的。近些年，随着维基百科，百度百科，互动百科这些互联网应用的兴起，越来越多的高质量数据被积累和得到，大量被精心设计以自动或半自动方式生成的知识库（例如Dbpedia、Freebase、Yago2）被建立起来。问题理解和知识库的建立是自然语言问答系统中的关键步骤。

2 知识驱动的自然语言问句理解

问题理解作为问答系统中的重要一部分，对问答系统的整体性能至关重要。据考证，现有问答系统回答用户提问的精准度普遍都不高，其中系统“理解”问句不够准确是一个相当重要的因素。总体问句理解流程如下图2所示：[2]

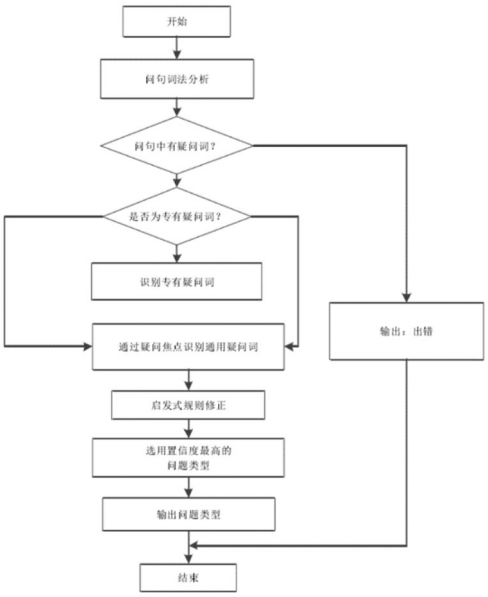

图 2

2.1 问题分类

首先，问题理解部分的问题分类，可以把待处理的问题分到已经定义好的某个类别里，这对于后续的检索和答案抽取工作都是很有帮助的，因为对于不同类型的问题，我们可以制定不同的规则来进行检索和答案抽取，这有助于提高运行效率和准确率；其次，对问题进行分析，提取出问题的关键信息，对后续的信息检索和答案抽取同样很重要，因为问题中的各个词的重要程度是不一样的，有些词甚至是可以直接去掉，不用考虑的，而有些词的重要程度却是非常高的，必须保留。冯等[1]通过基于规则的方法对问题进行分类，通过对问题关键信息的提取得到问题的关键词和扩展词。卢等[3]通过对基于事例的规则以及决策树来研究问句分析器，采用了多层次结构来描述问句，从理解问句的角度来进行问句处理。

2.2 语义转化

在中文信息检索系统中，用户用自然语言形式提问，这样的问句是不能作为查询问句而进行信息检索的。知识库通常都以 RDF 的格式存储数据。SPARQL 是目前已知的比较高效的查询 RDF 数据的语言，但是通常只有专业的程序员才能掌握 SPARQL 语法。大多数用户更愿意选择使用关键词或者自然语言问句来查找需要的信息。所以理想的解决办法是用户使用自然语言描述问题，而系统利用 SPARQL 语言在知识库查询。目前，面向知识库的问题理解主要针对英文的自然语言问题。许等[4]提出查询语义图的概念，并且利用图结构表示自然语言问题的语义，最终将其转化为结构化查询的方法。该方法从自然语言问句的句法结构入手，提出一套启发式识别实体与关系的方法，并利用语料库建立了从实体到知识库的映射，对谓词进行消歧，进而转化为计算机可理解的结构化查询语言。

2.3 关键词提取和查询扩展

对问句分析关键技术中的关键词提取和查询扩展进行研究工作的开展，将分类思想用在关键词提取上，将关

键词提取的过程看成是对一个词根据其经验对其分类的过程，在贝叶斯分类模型的基础上，提出了一种基于改进的贝叶斯分类模型的关键词提取算法，克服传统分词算法不准确的特点，使得到的关键词更准确，提高了检索系统的准确率。此外，作者还提出了一种基于潜在语义分析的问句查询扩展算法，有效地对关键词进行扩展，提高检索的准确率。[5]

3 基于知识的自然语言问答技术

基于知识库的自动问答系统在人工智能领域具有很长的发展历史。早期的研究主要针对小规模的专用知识库，使用的方法以语义解析为主。近期的研究方法主要分为基于语义分析的方法和基于信息检索的方法两大类。基于语义分析的方法首先将自然语言形式的问句转换为某种逻辑表达形式，常见的有 lambda 表达式和依存组合语义树等。

3.1 基于语义分析的方法

对于基于语义分析的方法，传统的语义分析解析方法存在两个问题：一方面是以来人工标注逻辑形式作为训练数据；另一方面是逻辑谓词的数量少，语义解析操作不能扩展到开放领域。Berant J[6]等人为解决以上问题，提出一种不依赖人工标注的逻辑形式、将可操作的谓词扩展到 Freebase 知识库内包含的所有谓词的语义解析方法，具体做法是：实现利用网络中的无结构化文本和 Freebase 知识库建立一个短语到谓词的映射，在将被影射到的词汇组合成逻辑表达式阶段，通过训练好的机器学习模型结合人工制定的少量规则自底向上合成逻辑表达式。Berant J 等人[7]在之前工作的基础上先通过一种简单的策略从问句生成出大量候选的逻辑表达式，并且对每个候选逻辑表达式都合成一个自然语言的问句，然后将解析逻辑表达式的任务转化成从大量候选逻辑表达式和生成问句对中找最优逻辑表达式的任务。在挑选最优候选逻辑表达式的过程中，会利用词向量表征方法对生成问句和原始问句计算相似度，然而传统词向量是通过与具体任务无关的无监督训练方法得到，用于复述问句评分时无法体现句子级的语义约束关系。因此，本文将改进词向量构建方法，使提取出的特征更适合于知识库问答任务。詹[8]提出了一种基于复述关系约束的词向量构建方法。针对传统词向量用于知识库问答中的复述问句评分时无法体现句子级的语义约束关系的问题，设计实现了基于问句对间相似度不等式的句子级语义约束信息表示方法，

以及加入不等式约束项的词向量训练方法，提高了问句间复述关系评价和知识库问答系统中问题回答的准确度。另外，作者还设计实现了一种结合神经网络问句生成的知识库的问答方法。针对知识库问答模型以来训练数据而人工生成大规模问答对较为困难的问题，设计实现了基于编码器-解码器神经网络模型的问句生成方法，通过使用神经网络生出问句取代传统规则生成问句用于知识库问答模型训练，有效提升了知识库问答系统的准确率。

3.2 基于信息检索的方法

基于信息检索的方法侧重于抽取有效特征，对候选进行排序。Yao 等[9]使用依存分析技术，获得问题的依存分析数，进而找到问句中涉及的主要实体，从知识库中找到该实体的 Topic Graph，之后从问题的依存树和 Topic Graph 中抽取多种特征，将其送入逻辑回归模型中进行分类。随着神经网络技术的不断进步，研究人员开始尝试将神经网络应用于自动问答领域。Borders 等[10]使用简单的前馈神经网络，对候选答案的各方面信息进行语义编码，将问句和候选答案分别转换为相同维度的向量，最后以两个向量的点积作为候选的得分。Dong 等[11]使用 MCCNN（multi-column CNN）网络，分别对问句中隐含地答案类型、关系和上下文信息进行语义编码，对候选答案的这 3 类信息同样进行语义编码，最后对这 3 类信息的语义向量的点积进行加权，得到最后的得分。Zhang 等[12]在 Dong 等[11]的基础上，使用注意力机制，对候选答案的不同内容，训练不同的问句表示。周等[13]针对实体识别和属性映射两个问题进行研究工作，在进行实体识别时，首先采用别名词典获取候选实体，然后使用 LSTM 语言模型结合简单的文本特征进行打分。在进行属性映射时，考虑到属性与问句中的词对应关系，结合两种不同的注意力机制，使用双向 LSTM 结合简单文本特征获取正确的属性。

结束语

问答系统是目前自然语言处理的研究热点，具有非常广泛的应用前景。其中问句理解是其研究基础和难点。中文处理和英文处理有着很大的不同，所以研究人员在使用新技术的同时，也要会回归到不同语言本身的特点，从本质上分析问题。目前问答系统的相关研究依旧是热点问题，随着研究的逐步深入，问答系统的研究将会取得重大的突破。

参考文献：

[1]冯晓波，李蕾，刘冬雪.中文问答系统中问题理解的研究[A].全国网络与信息安全学术会议论文.2010

[2]黄振兴.中文问句理解技术研究及在 IT 领域问答系统中的应用.华南理工大学.2015

[3]卢志坚，张冬荣.中文问答系统中的问句理解[A].计算机工程.2004

[4]许坤，冯岩松，赵东岩，陈立伟，邹磊.面向知识库的中文自然语言问句的语义理解.北京大学学报.2014

[5]姚健.问答系统中文问句分析关键问题研究. 哈尔滨工业大学.2009

[6]Berant J,Chou A,Frostig R,Liang P.Semantic Parsing on Freebase from Question-AnswerPairs[C].The 2013 Conference on Empirical Methods on Natural Language Processing,2013

[7] Berant J,Liang P,Semantic parsing via paraphrasing.In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,2014

[8]詹晨迪.基于知识库的自然语言问答方法研究.中国科学技术大学.2017

[9]Yao X,Van Durmc B.Information extraction over structured data:question answering with freebase ACL.Baltimore,2014

[10]Bordes A,Chopra S,Weston J,et al.Question answering with subgragh embeddings // EMNLP.Doha,2014

[11]Dong L，Wei F,Zhou M, et al.Question answering over freebase with multi-column convolutional neural networks // ACL.Beijing,2015

[12]Zhang Y,Liu K,He S,et al.Question answering over knowledge base with neural attention combining global knowledge information.arXiv preprint arXiv:1606.00979,2016

[13]周博通，孙承杰，林磊，刘秉全.基于 LSTM 的大规模知识库自动问答.北京大学学报.2018