情报工程 ISSN 2095-915X
第2卷 第1期 32-42 2016年2月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.1 32-42 Feb 2016

doi:10.3772/j.issn.2095-915x.2016.01.005

中文问答系统问句分析研究综述

张宁，朱礼军

（中国科学技术信息研究所工程中心 北京 100038）

摘要：自动问答系统成为近年来自然语言处理领域的研究热点，问句分析作为问答系统的首要环节，在问答系统中起着关键的作用。简要介绍了中文问句分析的基本内容，主要包括分词、词性标注以及句法分析的发展；同时也对中文问句分析中问句分类和问句语义分析的研究内容进行了重点介绍；最后，提出中文问句分析面临的一些难点问题以及对未来可能研究方向的一个初步展望。

关键词：问句分析，问句预处理，问句分类，问句语义分析

中图分类号：G302

A Survey of Chinese QA System's Question Analysis

ZHANG Ning, ZHU Lijun

(Engineering Center , Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: The automatic question answering(QA) system has become a research focus in the field of natural language processing(NLP) in recent years. Question analysis plays a key role in QA system. The basic contents of Chinese question analysis are introduced,including segmentation, part-of speech tag and syntactic. The contents of Chinese questions such as question classification and semantic analysis questions are mainly introduced. Finally, we point out several difficult points in Chinese question analysis and a preliminary prospect of possible research directions in the future.

Keywords: Question analysis, question processing, question classification, question semantic analysis

---

基金项目：本研究得到中国工程科技知识中心建设项目“知识组织系统建设”（编号：CKCEST-2015-1-11）的资助。

作者简介：张宁，研究生，中国科学技术信息研究工程中心，研究方向：知识工程；朱礼军：男，博士，中国科学技术信息研究工程中心，研究方向：语义网、Web service 知识技术在科技信息服务、电子政务 / 商务中的应用以及知识组织系统相关研究。

1 引言

问答系统采用自然语言处理技术，一方面完成对用户疑问的理解；另一方面完成正确答案的生成。它能用准确、简洁的自然语言回答用户用自然语言提出的问题，是信息检索系统的一种高级形式。问答系统包括问句分析、信息检索和答案抽取三个部分。问句分析作为问答系统的首要环节，要想能够准确地回答问题首先要正确地理解问题。问句分析是问答系统的一项重要技术，准确深入的问句分析可以提高后续环节的处理质量与处理效率。问句分析主要完成对问句的语义理解，将问句从模糊的自然语言转化成清晰的逻辑语言，使问句得到预期地处理。中文的问句分析一般包括的基础工作有分词、词性标注、句法分析、命名实体识别、关键词提取与扩展，并在此基础上完成对问句的分类和语义分析等。

2 问句预处理研究

在对问句进行语义分析和分类之前，需要对问句进行预处理操作，主要包括了问句分词、词性标注、句法分析等工作。随着研究者对中文自然语言处理技术的不断完善，其中的词法分析和句法分析等研究内容已相对成熟，这些基础工作一般不作为问句分析中的重点进行研究。目前常用的中文分词和词性标注工具是中科院计算所研发的 ICTCALAS[1] 和哈工大社会计算与信息检索研究中心开发的语言技术平台（LTP）[2]。

2.1 分词和词性标注

问句分析中分词和词性标注工作是问句内容分析的基础。在 2002 年之前，中文分词基本上是基于词典的，并在此基础上分为基于规则的方法和基于统计的方法。2003 年，Xue[3] 利用最大熵模型实现了基于字的分词系统，在 AS2003 封闭测试项目中，获致最高 OOV 召回率（0.729）。不同于以往的分词方法，基于字的分词系统不依赖于词典，而是把分词过程等同于字在字串中进行标注的问题；该过程中所有的字根据预定义的特征进行词位特征学习并获得一个概率模型，然后在待分字串上根据字与字的结合程度得到词位标注结果，最后根据词位定义获取分词结果 [4]。基于字的分词系统提高了中文分词的性能，近年来借助学习算法，由字构词逐渐成为中文分词的主要方法。在此基础上，宋彦等 [5] 提出结合基于字的条件随机场模型（CRF）与基于词的 Bigram 模型的切分策略，发挥两个模型的长处，有效地改善了单一模型的性能，实现了字词联合解码的中文分词方法,可以更好地应用于中文信息处理。张桂平等 [6] 针对专利文献提出了一种基于统计和规则相结合的多策略分词方法。该方法利用文献中潜在的切分标记，结合切分文本的上下文信息进行最大概率分词，有效地解决了专利分词中未登录词难以识别的问题。

在中文的词法分析中，分词是词性标注的前提，和中文分词方法类似，词性标注也包括基于统计的方法、基于规则的方法以及规则和统计结合的方法。刘群等 [7] 提出了一种基于层叠隐马模型的词法分析系统，旨在将汉语分词、词性标注、切分排歧和未登录词识别集成到一个完整的理论框架中，实现了基于层叠隐马模型的汉语词法分析系统 ICTCLAS。朱聪慧等 [8] 基于无向图模型，将分词和词性标注有机地整合到一个序列标注模型中，以深层次的依赖关系为特征，该系统取得了较高的分词精确率和词性标注精确率。然而对于中文分词和词性标注联合模型存在两个问题：一是融合方式还需要进一步改进，二是模型性能受限于标注语料的规模。郭振等 [9] 依据序列标注的中文分词方法，重新设计了基于转移的中文分词的处理方案，使以往的中文分词研究成果融入联合模型；使用具有部分标注信息的语料，抽取

字符串层面的 Ngram 特征融入联合模型，在宾州中文树库的实验结果中，该模型的中文分词和词性标注的 F1 值分别达到了 98.31% 和 94.84%，相对于单任务模型提升了 0.92% 与 1.77%。

从对中文分词的研究发展来看，其存在着从基于词典到基于字再到由字构词的演变特点，其对应的方法是从基于规则到基于统计最终发展为基于规则和统计相结合的方式。基于规则和统计的方法既利用了词典匹配切分效率高、速度快的特点，又能够结合上下文，统计字出现的频次来识别未登录词的特点，从而提高分词的准确率。在目前的中文分词中，还存在着歧义识别和新词识别两大尚未突破的难题，而目前国内对中文分词的研究大多是科研院校，如北大、清华、中科院、北京语言大学等一些研究所，真正专业研究中文分词的商业公司很少[10]，研究成果产品化的过程比较缓慢，因此中文分词技术想要更快更好地服务更多产品在现阶段还是面临一些困难。词性作为一个语法特征，在中文中表现出的作用较为不明显，而利用词性标注信息能够提高分词阶段中文分析的效果，研究者将这两个阶段合并到一个架构中，通过改进策略，获得了更高的分词和标注精度。

2.2　问句句法分析

中文句法分析是中文信息处理中的一个重要环节，也是难点之一。在中文问句分析中如果有精确的句法分析，则能使问句语义分析过程得到更加有效的处理，从而为问答系统的后续处理提供更好的基础平台。对句法分析的研究大体分为基于规则的方法和基于统计的方法；基于规则的方法以语言学理论为基础，是一种理性主义方法；基于统计的句法分析以某种方式对语法规则和语言形式进行描述，这种描述通过对已知句法分析结果进行训练获得[11]，即句法分析模型。基于统计的句法分析比基于规则的更为受到研究者的关注。中文问句句法分析常用算法包括基于上下文无关文法（CFG）规则的分析方法和基于概率上下文无关文法（PCFG）的分析方法。基于概率上下文无关文法的分析方法是较早使用的句法分析模型[12]。句法分析的主要任务是：（1）语块的识别和分析；（2）语块之间的依附关系分析；（3）构建句法分析树。

张亮等[13]探讨了中文问句的结构特征，在问句句法分析算法中，采用语料库句法处理技术，利用问句相关特点，如长度短、有疑问词和疑问结构句式等，实验结果表明基于问句结构特征的句法分析比 PCFG 有较大提高。袁里驰[14]提出利用依存关系、互信息对词聚类，同时考虑多种语义依存关系，将句法分析模型与分词、词性标注模型相结合，使 PCFG 模型中由概率上下文无关性假设和祖先结点无关性假设引起的问题得到有效解决。之后，依存句法分析受到越来越多的关注。陈永波等[15]提出简单边优先与 SVM 相结合的依存句法分析算法，实验证明该算法比单纯的优先边算法分析效率更高，且优于基于最大生成树算法的中文句法分析器。在句法体系中，作为底层核心技术，依存句法以其方便利用的特点，已被广泛应用到各项研究中，如词义消歧[16]、问句分类[17]、信息抽取[18]等的直接利用以及结合分词、词性标注技术实现一体化分析[14]的间接利用，均取得了良好的效果。

2.3　小结

中文自然语言处理中分词、词性标注和句法分析是三个基础任务。近年来联合模型成为研究热点，通过对字的分词系统和词典分词进行结合，实现字词联合解码的分词方法，更好的实现中文分词效果；通过将分词和词性标注融合到一个模型的联合模型，解决单任务模型存在的错误传递以及多层次特征无法获取等问题；较之于单一任务模型，不同的联合模型能使中文自然语言分析

的性能得到不同程度提高。作为问句分析的基础工作，高质量的词法和句法分析能够提高问句分析中问句分类和语义标注的工作效率。

3　问句分类研究

对问句进行分类能够对候选答案进行有效地筛选，根据不同的问句类型调整不同的答案选择策略，提高返回结果的准确率。现有的问句分类体系包括三种：基于问句语义信息的问句分类，基于答案类型的问句分类以及基于混合信息的问句分类。常用的是基于问句答案类型的问句分类。问句分类在问句处理过程中的作用主要是：一、有效减少相关片段和文档的候选答案空间，使检索更为高效；二、可以根据问句类型制定答案抽取策略，如果在系统处理过程中问句不能分到正确的类别，将直接导致答案抽取模块的选择失误。

3.1　问句分类研究概况

国外研究者对问句分类的研究起步较早。对问句分类的研究包括基于规则的方法、基于统计的方法以及基于规则和统计的方法。早期使用的是基于规则的方法，主要是通过人工分析句法结构来提取规则，进而对问句类型进行判定。该方法更多的是依赖专家，有很大的主观性，而且分类体系对专家的分类决策有很大的影响，因此使得其灵活性较差。随后基于统计学习的方法表现出良好的分类效果，但是其分类精度还是会受句法分析精度的影响。因此研究者将两者结合应用到问句分类中，取得了良好的分类效果。

中文自然语言处理的研究起步较晚。对中文问答系统的研究比较著名的有中科院计算所研发的知识问答系统 NKI[19]、北京理工大学开发的银行领域自动问答系统 BAQS[20] 以及台湾国防管理学院开发的中文问答系统 CQAS[21] 等；对中文问答系统的研究机构主要还有复旦大学、哈尔滨工业大学、北京大学等以及台湾中央研究院、台湾大学等。国内研究者对中文问句的分类最初是借鉴于英文问句分类。但是中文和英文还是有诸多差异的，中文自然语言有表达形式多样、语法结构复杂等特点，相比英文自然语言处理起来也更难。因此对中文问句的处理还需结合中文的特点来进行研究分析。

张宇等 [22] 利用词与词之间的无关性，并且引入贝叶斯模型来处理中文问句，简化了问句的分类，在一定程度上取得了良好的效果。这种方法容易实现，但是侧重于词频而忽略了句法结构和语义信息，在问题分类算法方面仍有需要改进的地方。文勖等 [17] 通过使用句法分析的结果，提取出问题的疑问词和主干信息及其他附属成分作为分类特征。该方法有效地降低了噪音，突出问题分类的主要特征，利用贝叶斯分类器大幅度提高了分类精度。试验结果证明该方法在大类和小类的分类精度分别达到了 86.62% 和 71.92%，取得了良好的效果。该方法的分类精度很大程度上受句法分析精度的影响，因此如何更好地利用句法分析技术有待进一步研究。孙景广等 [23] 提出了使用知网作为语义资源来选取分类特征，并且利用最大熵模型来对问句进行分类。该方法抽取问句的疑问词、句法结构、疑问意向词及其在知网中的首义原作为分类特征，实验结果显示在知网中选取的首义原能很好地表达问题焦点词的语义信息，作为分类的一个主要特征，能显著得提高问题分类的精度，在大类和小类的分类精度分别达到了 92.18% 和 83.86%。牛彦清等 [24] 针对问句分类特征对问句分类的影响不同，提取和处理特征的负责度不同的问题，对问题疑问词和核心关键词的主要义原、核心关键词的首义原、问句主谓宾的主要义原、命名实体、名词单复数等六种特征，利用支持向量机分类算法，对问句不同特征进行组合实验证明，采用词义消歧技术提取的主要义原对分类的准确率影响较大，而且可以大幅度的

降低特征向量的维数。在问句分类中对问句特征进行有效组合可以辅助构造问句分类器。杨思春等[25]针对当前问句分类中特征组合问题，提出了基于重要性和差异性的特征组合。通过计算获取候选特征中优化的特征组合，在哈工大中文问句集进行实验结果表明，这种特征组合的方法更加灵活高效，准确率也更高。王小林等[26]针对固定训练集生成的分类器不能跟踪用户需求的问题，提出了将增量式贝叶斯思想用于问句分类的方法。该方法采用遗传算法来选取最优特征子集对分类器进行优化，使分类器在学习过程中动态扩大并调整参数，实验结果表明增量式贝叶斯分类器与朴素贝叶斯分类器相比有更高的分类精度和运行效率。

3.2 问句分类研究内容

问句分类指的是在事先确定的分类体系下将用户提出的问句分到与其最接近的类别从而判定问句类型。问句分类是问句分析部分的一个主要内容。对其进行形式化地表示如下：

F：Q {$C_1,C_2,\cdots C_n$}

其中，Q 表示给定的问句样本集合，{$C_1,C_2,\cdots C_n$} 表示问句类别集合，F 表示映射规则，即系统根据每类样本的先验信息建立的映射函数，负责将未知类别问句 $q\,\varepsilon Q$ 根据建立的规则映射到问句类别集合 C_i 中。问句分类通常包括问句特征提取、特征权重计算、构建分类器、输出问句类别标签。大多数问句分类的研究重点都是如何合理地抽取问句特征来提高分类算法地准确率。问句特征抽取是指在问句预处理的基础上将问句表示成特征向量，即选用向量空间模型 (VSM：Vector Space Model) 将非结构化的问句表示成计算机可识别的形式。向量空间模型是由 Salton 等研究者在 20 世纪 70 年代提出[20]，成功地用于 SMART 文本检索系统，目前也是自然语言处理常用模型之一。利用向量空间模型表示问句包括两个部分：一是根据问句特征选取方法将问句表示为特征项序列 $S=\{t_1,t_2,\cdots t_n\}$，其中 S 表示问句本身，t_i 表示特征项；二是根据问句特征项序列对问句进行特征权重计算，将其转化为分类特征向量。问句特征抽取包括基于句法特征的方法和基于语义特征的方法：句法特征包括对问句中词和词性、句法结构信息；语义特征包括命名实体和词义特征。

当前的问句分类方法中基于统计的机器学习方法因其诸多优点而被广泛使用，其所使用的统计模型主要有：支持向量机模型 (SVM) [27]、贝叶斯分类模型 [17、22]、K- 最近邻模型 (KNN) [28]、最大熵模型 (ME) [29]。

（1）支持向量机模型

V. Vapnik 等[30]研究者在 20 世纪 90 年代中期基于统计学习理论提出支持向量机分类模型。其原理为通过已选择的非线性映射将输入的向量 x 映射到一个高维向量空间，在该空间寻找一个最优切分两类数据的超平面，使两类模式向量分类间隔最大，以保证经验风险以及分类器的结构风险最小。这个超平面可以表示为分类函数

$f(x)=w^Tx+b$，

其中 x 为训练样本集的特征向量，w 为权重向量，b 为偏移量。当 f(x) 等于 0 的时候，x 便是位于超平面上的点，而 f(x) 大于 0 的点对应 y=1 的数据点，f(x) 小于 0 的点对应 y=−1 的点，如图 3.1：

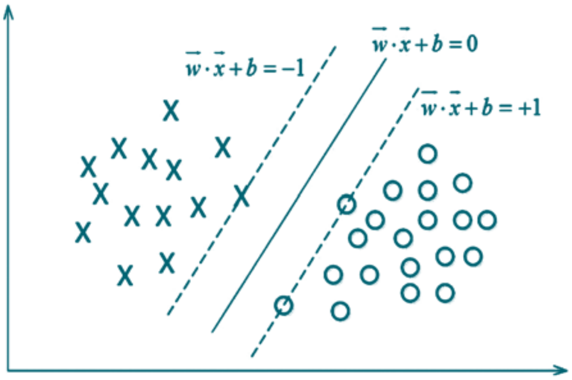

图 3.1 支持向量机（一般分界）

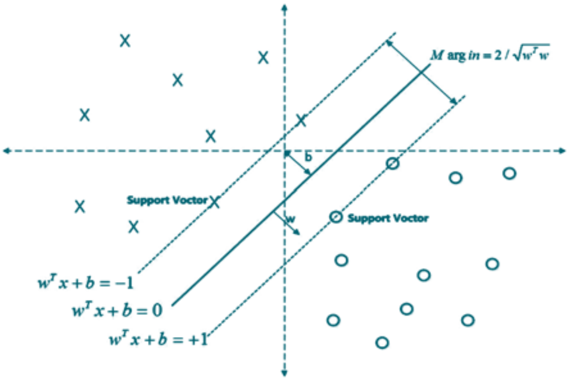

图 3.2 支持向量机（最优分界）

图 3.2 表示支持向量机的最优分界，其中中间的实线是寻找到的最优超平面，其到两边虚线距离相等为 $y(w^T x+b)$，虚线上的点表示支持向量，满足 $y(w^T x+b)=1$，对于不是支持向量的点，则有 $y(w^T x+b) >1$。

支持向量机构造的是二值分类器，对于多类模式识别需要建立多个二值分类器，其处理结果依赖于掌握的模式样本集的构造。对大规模训练样本实施起来较为困难，解决多分类问题时时间开销大，适合小样本学习。

（2）贝叶斯分类模型

贝叶斯模型的分类原理是通过类别的先验概率和特征项分布，利用贝叶斯公式计算出该对象属于某一类的后验概率，选择具有最大后验概率的类作为文本类别。贝叶斯分类器常用于文本分类，文献 [14] 设计了一个改进的贝叶斯分类器对问句进行分类，利用词袋模型，数学表达式为

$$\arg\max_{qc} P(qc|Q_1,Q_2,\dots,Q_n)=\arg\max_{qc}(P(qc,Q_1,Q_2,\dots,Q_n))/(P(Q_1,Q_2,\dots,Q_n))$$

其中 qc 为问句类型的变量，Q_1 至 Q_n 为对问句进行分词后的特征项，对上式做简化后，由于分母不变，所以只需要处理

$$\arg\max_{qc} P(qc,Q_1,Q_2,\dots,Q_n)$$

根据上面提到的词袋模型，可以简化为

$$\arg\max_{qc} P(qc,Q_1,Q_2,\dots,Q_n)= \arg\max_{qc}(P(qc,Q_1)\times\dots\times P(qc\ ,Q_n))$$

对于某个 $P(qc,Q_i)$，采用近似算法并使用 TF−IDF 进行权值处理

$$P_2(qc,Q_i)=P_1(qc,Q_i)\times\log((N+0.1)/(M+0.1))$$

其中 N 表示问句分词后词项个数，M 表示 Qi 在问句类型中出现，加 0.1 为调零因子。

贝叶斯分类模型由一种数学概率运算演化而来，其特点为算法简单，能够处理大规模和多类别的样本，对缺失数据不敏感。贝叶斯模型处理分类问题较为高效但分类精度较低，而且无法满足特征的独立性。

（3）K− 最邻近模型

K 最近邻 (KNN) 分类算法是在理论较为成熟的方法，其核心思路是：在特征空间中一个样本的 k 个最相似，即在特征空间中的最邻近样本中的大多数属于某个类别，那么该样本也属于该类别。可以表示为

$$p(Q)=\max_{1\le j\le n}\sum_{q_i\in KNN} y(q_i,c_i)$$

其中 Q 为样本测试问句，q_i 为特征空间中的 k 个最相似， 为类别属性判定函数，若 q_i c_j，则 =1，否则 =0。

该方法在分类上的主要不足是：一个类的样本容量很大且数据不平衡而其他样本容量很小，有可能导致输入新样本时样本中 K 个最邻近中大容量样本占多数，但有可能并不是最接近目标样本；另外还有一个不足之处是该方法计算量较大，对每一个分类文本都要计算其到全部已知样本的距离才能求得 K 最邻近点。目前 KNN 分类算法多与其他模型或算法结合进行分类运算或处理其他问题。

（4）最大熵模型

最大熵模型 (ME) 的原理是：对一个随即事件的概率分布进行预测应当满足全部已知条件，对未知的情况不做任何主观假设；这种情况下概率分布最均匀，预测风险最小，概率分布的信息熵最大。在 20 世纪 90 年代，IBM 的研究员系统地描述了最大熵地框架和实线算法，在自然语言

处理任务上取得了良好的效果。

假设问句训练样本集 $Q=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中任意 $x_i(0<i<n+1)$ 表示问句的特征向量，$y_i(0<i<n+1)$ 表示样本中问句地类别。在给定样本 Q 和其相关约束条件下，存在唯一一个概率模型 $p(y|x)$ 使其熵分布最大，来证明 $p(y|x)$ 的取值符合如下指数模型：

$$z(x)=1/(\sum_{y}\exp(\sum_{i=1}\lambda_i f_i(x,y)))$$

$$p(y|x)=1/(z(x))\exp(\sum_{i=1}\lambda_i f_i(x,y))$$

其中，f_i 是一个特征函数，其值只能是 0 或 1，λ_i 是模型参数，第二个算式将模型由求概率值转化为求参数值 λ_i。一般常用的 λ_i 估计方法是 Darroch 和 Ratcliff 的通用迭代算法 (Generalized Iterative Scaling,GIS)[32] 和 Pietra 等人提出的改进迭代算法 (Irnproved Iterative Scalin&IIS)[33]。

3.3　小结

问句分类是问句分析中的重要模块，其对问答系统中答案抽取部分有着间接的决定性作用。从目前的问句分类研究进展来看，仍有一些问题需要改进：一是问句训练规模有限，问句分类受到语料库规模和质量的约束；二是用于学习的分类模型不够完善，对问句分类还达不到更好的效果；三是目前还没有一个统一的分类体系标准，各个系统根据自身领域的特点自行进行类别的定义，使得训练数据难以共享。因此，今后的问句分类可以在问句语料规模、分类训练模型、问句处理平台等方面进行进一步研究，为更高性能的问句分析奠定基础。

4　问句语义分析研究

问句分析主要是完成对问句的语义理解，只有正确地理解问句的意思，才能够找到正确的答案。深入的问句分析需要明确问句的主题信息，问句的类型，问句的条件约束等。问句的语义分析涉及到复杂的语言学知识，而加上中文属于弱语法结构的语言，因此对中文进行语义分析变得更加困难。对问句进行语义分析是在问句预处理的基础上对其进行序列化标注以及潜在语义分析等工作，最终分析出问句完整的语义信息。

4.1　问句中语义块序列化标注

对问句进行序列化的标注，可以将问句从模糊的自然语言文字空间，映射到逻辑清晰的结构化的语义空间，更加利于计算机对问句的理解。在自然语言处理领域，基于机器学习的方法在序列化标注问题中受到了广泛的关注，主要适用的模型有：隐马尔可夫模型 (HMM)、最大熵马尔可夫模型 (MEMM)、条件随即场模型 (CRF) 等。

隐马尔可夫模型是非常经典的数学统计模型，现已广泛地应用于查询扩展[34]、语音识别[35]、信息抽取[36]等研究中。HMM 是一种产生式模型，它包含了一个可观察层和一个隐藏层，其任务是从可观察层的参数中确定隐藏层隐含的参数，利用这些参数来做进一步分析。一个 HMM 可以用一个五元组 $\{N,M,\pi,A,B\}$ 表示，其中 N 表示隐藏状态的数量，M 表示可观测状态的数量，可以通过训练集获得，$\pi=\{\pi_i\}$ 为初始状态概率，$A=\{a_{ij}\}$ 为隐藏状态的转移矩阵，$B=\{b_{ik}\}$ 表示某个时刻因隐藏状态而可观察的状态的概率即混淆矩阵。状态转移矩阵和混淆矩阵中的每个概率都是与时间无关的，当系统演化时，其不随时间改变。

对于一个 N 和 M 固定的 HMM，用 $=\{\pi,A,B\}$ 表示 HMM 参数。在序列标注问题中，对于给定参数 和观察序列 $O=\{o_1,o_2,\cdots,o_T\}$，找出状态序列 $q=\{q_1,q_2,\cdots,q_T\}$，使

$$q=\arg\max_q P(Q=q|O=o)$$

这个问题通常使用动态规划的 Viterbi 算法来解决。HMM 考虑了标记之间的关系，可以得到

整体的最优解。HMM 引入的两条独立性假设使 HMM 的发射概率只是局限于部分上下文特征，这是 HMM 模型的一大缺点，而且为定义观察值和状态值的联合概率，产生式模型必须列出所有可能的观察序列，这在实际操作中是很难实现的。

最大熵马尔可夫模型是一种判别式模型，它不需要 HMM 那样严格的独立性假设。MEMM 模型将观察序列看作是条件事件，而不是由状态生成的。它结合了 MEM 和 HMM 的优点，允许状态转移可以基于输入序列中的非独立性特征，使得 MEMM 在处理自然语言处理的任务时，性能优于 HMM[37]。MEMM 是通过求局部最优的条件概率来获得最终的条件概率。

MEMM 模型可以用一个四元组 {N,M, π ,C} 表示，其中 N 为状态数量，M 为观测事件数量，C 为概率模型集合 , π 为初始状态概率。MEMM 中序列标注问题算法也可以用 Viterbi 算法处理，与 HMM 的 Viterbi 算法十分相似。MEMM 在计算转移概率时仅对局部求解条件概率，取其概率最大的标记作为最终的输出标记，导致存在一些标记偏置的问题。

条件随机场模型是对最大熵马尔可夫模型的进一步发展，解决了后者标记偏置的问题。CRF 也是一种判别式模型，其核心思想是利用无向图理论使序列标注的结果达到在整个序列上全局最优。理论上，无向图的结构是任意的，但是当用于序列标注任务时，一般假设图是最简单和最通用的图结构，将其称为线性链条件随机场 (Linear-chain CRF)。CRF 在自然语言处理领域的应用包括命名实体识别 [38]、词义消歧、答案抽取 [39]、机器翻译 [40] 等。

三种模型都可以通过 Viterbi 等动态规划算法求得最优值。HMM 模型是对转移概率和表现概率直接建模，统计共现概率。MEMM 模型是对转移概率和表现概率建立联合概率，统计的是条件概率，容易陷入局部最优。CRF 模型统计了全局概率，考虑了数据在全局的分布，而不是仅仅在局部归一化，解决了 MEMM 中的标记偏置问题。

4.2 潜在语义分析

在中文中，词的同义和多义问题给文本处理带来了诸多麻烦。解决词的同义现象常用的方法是查询扩展，词的多义现象是利用受控词表以及人工处理作为转换机制，或者利用布尔交集对多义词进行消歧。但是上述这些方法取得的效果并不理想。潜在语义分析 (LSA) 在解决词的同义和多义问题上效果较为显著。

潜在语义分析是在自然语言处理过程中用到的方法，其通过“矢量语义空间”来提取文档与词中的“概念”，进而分析文档与词之间的关系。潜在语义分析是在大量文本的基础上构建一个矩阵，这个矩阵中一行代表一个词，一列代表一个文档，矩阵元素代表该词在该文档中出现的次数，然后再此矩阵上使用奇异值分解 (SVD) 来保留列信息的情况下减少矩阵行数，之后每两个词语的相似性则可以通过其行向量的 cos 值来进行标示，此值越接近于 1 则说明两个词语越相似，越接近于 0 则说明越不相似。下面是潜在语义分析的一个具体理论表达式：

$$\mathrm{F}(\mathrm{m}\times\mathrm{n})=[f_{ij}]=(\mathrm{d}_1,\mathrm{d}_2,\ldots,\mathrm{d}_n)=(\mathrm{t}_1,\mathrm{t}_2,\ldots,\mathrm{t}_m)^{\mathrm{T}}$$

式中 f_{ij} 表示词汇 i 在文档 j 中出现的频次，$\mathrm{t_i}$ 表示词汇 i 在文档中的列向量。一般情况下， 也可以是该词汇的 TF-IDF。$\mathrm{F_{mxn}}$ 表示的是该词 - 文档矩阵，任何一个文档均是由有限个词汇构成，而不是所有 m 个词汇构成，因此 $\mathrm{F_{mxn}}$ 是一个稀疏矩阵。在构建好词 - 文档矩阵之后，LSA 将对该矩阵进行降维，来找到词 - 文档矩阵的一个低阶近似。这里需要将 $\mathrm{F_{mxn}}$ 分解：

$$\mathrm{F}_{\mathrm{m}\times\mathrm{n}}=(\mathrm{U}_{\mathrm{m}\times l}\ \mathrm{S}_{l\times l}\ \mathrm{V}_{\mathrm{n}\times l})^{\mathrm{T}}$$

$$\mathrm{U}_{\mathrm{m}\times l}=(\mathrm{u}_1,\mathrm{u}_2,\ldots,\mathrm{u}_l)$$

$$\mathrm{S}_{l\times l}=\mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_l)$$

$$\mathrm{V}_{\mathrm{n}\times l}=(\mathrm{v}_1,\mathrm{v}_2,\ldots,\mathrm{v}_l)$$

其中 $\sigma_1,\sigma_2,\ldots,\sigma_l$ 被称作是奇异值，同时也是 $F_{m\times n}F^T_{m\times n}$ 和 $F^T_{m\times n}F_{m\times n}$ 所有特征值的平方根，满足 $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_l$；而 $u_1,u_2,\cdots,u_l$ 和 $v_1,v_2,\cdots,v_l$ 则叫做左奇异向量和右奇异向量，并且分别是 $F_{m\times n}F^T_{m\times n}$ 和 $F^T_{m\times n}F_{m\times n}$ 的特征向量。因此词－文档矩阵可以用下面的表达式表示：

$$F_{m\times n}=\sigma_1 u_1 v_1{}^T+\sigma_2 u_2 v_2{}^T+\sigma_3 u_3 v_3{}^T$$

选择 k 个最大的奇异值和它们对应的 U 和 V 中的空间向量相乘，则能够得到一个 F 矩阵的 k 阶近似，这样做可以将词向量和文档向量映射到语义空间。向量 t_i 与 k 个奇异值矩阵相乘，实质是从高维空间转换到低维空间，k 就是低维空间的维数。

LSA 最早在 1988 年由 Dumais 等 [41] 人提出，随后 LSA 方法被证明是对传统空间向量技术的改进方法，其在自然语言处理方面开始被广泛使用，包括答案抽取 [42]、问句相似度计算 [43]、文本聚类 [44]、关键词查询扩展 [45] 等方面。

4.3　小结

问句语义分析的目的就是为了使自然语言能够合理地转化为机器语言，从而使计算机得到正确的理解。研究者对自然语言进行语义序列标注和潜在语义分析的最终目标就是更深层次的语义分析。序列标注也属于句法结构部分内容。中文问句中语义分析和句法分析通常是结合在一起的，对句法结构进行分析时往往也需要对句中的语义关系、语义特征进行分析。因此，我们应该建立性能更加完善的句法分析和语义标注融合系统，来解决深层语义分析的难题。

5　总结

本文对中文问答系统问句分析中的问句预处理、问句分类以及问句语义分析的研究内容与研究方法进行了系统地阐述，发现好的问句分析需要用到多方面的处理技术，需要考虑不同技术的融合问题。作为一个热门的研究课题，目前国内对中文问句分析的研究还不是很成熟，主要表现在：

（1）语料库数据稀疏

在中文问答系统中，面向开放域的问答系统现已取得了一些成果，而在受限域问答系统的研究方面，领域语料库资源还比较缺失，比如在医药领域、金融领域、旅游领域等，通用的问答语料库并不能很好地满足这些领域的需求。在对问句进行分析时，由于可供学习的数据有限，导致分词和标注错误以及问句分类错误，从而使一些新兴领域问答系统中问句分析的性能下降。

（2）中文语义分析不足

中文的语法灵活，句子中没有明显的时态和语态变化，各成分之间的联系大多是靠词序来表现，这种关系是“虚实相间”的；同时中文表达的语义灵活，同形歧义现象十分普遍，这些都使计算机在处理时受到很大的阻碍，而现有的语义标注系统还尚未成熟，因此，对完全进行语义分析的工作还是任重而道远。

今后还需要在以下方面继续深入研究：

（1）改善分析模型

中文句子的语义角色之间存在一定的约束关系，进行语义标注时规定不能出现重复的语义角色，目前的序列标注模型并没有考虑到这些内容。可以考虑的解决方案是使用语义序列标注模型来处理语义角色标注的问题，将存在的约束关系加入到模型中，建立新的分析模型。这种方法现已在句法分析方面取得了良好的效果 [46]。

（2）改进语言处理平台

目前分词和词性标注普遍是利用通用的词法和句法分析器进行处理，但对这些平台的使用还没有达到理想的效果。在中文问句中，分词和词

性标注的错误导致句法分析和问句分类也进行错误的处理。因此可以考虑利用传统中文自然语言分析中的研究方法，如基于条件随机场[5][8]就是当前比较成熟的研究方法，在此基础上来提高分词和词性标注的正确率。

问句分析作为问答系统的重要组成部分，对提高问答系统的性能有着至关重要的作用。本文从三个方面综述了近年来中文问答系统中问句分析的研究进展，分析了中文问句分析中存在的难点，提出了今后在研究方法和研究模型方面应作出的改善，希望能为进一步研究奠定坚实的基础。

参考文献

[1] 中科院分词系统 ICTCLAS[EB/OL] [2015-12-26]. http://ictclas.org/.

[2] 哈尔滨工业大学信息检索研究室 [EB/OL] [2015-12-26]. http://ir.hit.edu.cn/.

[3]XUE N, CONVERSE S P. Combining classifiers for Chinese word segmentation[J]. First SIGHAN Workshop, 2002:57-63.

[4] 黄昌宁，赵海．中文分词十年回顾 [J]. 中文信息学报，2007(03):8-19.

[5] 宋彦，蔡东风，张桂平，等．一种基于字词联合解码的中文分词方法 [J]. 软件学报，2009, 20(9):2366-2375.

[6] 张桂平，刘东生，尹宝生，等．面向专利文献的中文分词技术的研究 [J]. 中文信息学报 ,2010,24(3):112-116.

[7] 刘群，张华平，俞鸿魁，等．基于层叠隐马模型的汉语词法分析 [J]. 计算机研究与发展 ,2004,41(8):1421-1429.

[8] 朱聪慧，赵铁军，郑德权．基于无向图序列标注模型的中文分词词性标注一体化系统 [J]. 电子与信息学报，2010 (03):700-704.

[9] 郭振，张玉洁，苏晨，等．基于字符的中文分词、词性标注和依存句法分析联合模型 [J]. 中文信息学报，2014, 28(06):1-8.

[10] 何莘，王琬芜．自然语言检索中的中文分词技术研究进展及应用 [J]. 情报科学 ,2008,26(5):787-791.

[11]Kantor P. Foundations of statistical natural language processing[J]. Natural Language Engineering, 1999, 26(2): 91-92.

[12]Wu C, Su H, Chiu Y, et al. Transfer-based statistical translation of Taiwanese sign language using PCFG[J]. ACM Transactions on Asian Language Information Processing, 2007, 6(1):1-20.

[13] 张亮，王树梅，黄河燕，等．面向中文问答系统的问句句法分析 [C]//. 第四届全国搜索引擎和网上信息挖掘学术研讨会 (SEWM2006), 2006:85-88.

[14] 袁里驰．基于依存关系的句法分析统计模型 [J]. 中南大学学报：自然科学版，2009 (06):1630-1635.

[15] 陈永波，汤昂昂，姬东鸿．中文复杂名词短语依存句法分析 [J]. 计算机应用研究，2015, 32(6):1617-1620.

[16] 卢志茂，刘挺，李生．统计词义消歧的研究进展 [J]. 电子学报，2006, 34(2):143-153.

[17] 文勖，张宇，刘挺，等．基于句法结构分析的中文问题分类 [J]. 中文信息学报 ,2006,20(2):33-39.

[18]SURDEANU M, HARABAGIU S, WILLIAMS J, et al. Using predicate-argument structures for information extraction[J]. ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, 2003(1):8-15.

[19] 王树西，刘群，白硕，等．基于动态知识库的问答系统研究 [C]// 全国第七届计算语言学联合学术会议，2003:587-592.

[20] 樊孝忠，李宏科，李良富，等．银行领域汉语自动问答系统 BAQS 的研究与实现 [J]. 北京理工大学学报，2004(6):528-532.

[21]Huang G, Yao H. Chinese question-answering system[J]. 计算机科学技术学报：英文版，2004, 19(4):479-488.

[22] 张宇，刘挺，文勖．基于改进贝叶斯模型的问题分类 [J]. 中文信息学报，2005, 19(02):100-105.

[23] 孙景广，蔡东风，吕德新，等．基于知网的中文问题自动分类 [J]. 中文信息学报，2007 (01):90-95.

[24] 牛彦清，陈俊杰，段利国，等．中文问句分类特征的研究 [J]. 计算机应用与软件，2012, 29(3):108-111.

[25] 杨思春，高超，戴新宇，等．基于差异性和重要性的问句特征组合 [J]. 电子学报，2014(05):918-924.

[26] 王小林，镇丽华，杨思春，等．基于增量式贝叶斯模型的中文问句分类研究 [J]. 计算机工程，2014(09):238-242.

[27]Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.

[28]Cover T M, Hart P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1):21-27.

[29]Ratnaparkhi A. Maximum entropy models for natural language ambiguity resolution[J]. Proc of the IEEE, 1998:1557 - 1560.

[30]Salton G, Wong A, Yang C S. A vector space model for automatic indexing[M]//. Morgan Kaufmann Publishers Inc., 1997:273-280.

[31]Cortes C, Vapnik V. Support-vector networks[C]// Machine Learning, 1995:273-297.

[32]Darroch J N, Ratcliff D. Generalized iterative scaling for log-linear models[J]. Annals of Mathematical Statistics, 1972, 43(5):1470-1480.

[33]Pietra S D, PIETRA V D, LAFFERTY J. Inducing features of random fields[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 1997, 19(4):380-393.

[34] 矫健，张仰森．基于隐马尔可夫模型的查询扩展方法 [J]. 计算机科学 ,2014,41(12):168-171.

[35] 吕勇，吴镇扬．基于隐马尔可夫模型与并行模型组合的特征补偿算法 [J]. 东南大学学报（自然科学版）,2009,39(5):889-893.

[36] 刘亚清，陈荣．基于隐马尔可夫模型的 Web 信息抽取 [J]. 计算机工程，2009, 35(18):25-27.

[37] 林亚平，刘云中，周顺先，等．基于最大熵的隐马尔可夫模型文本信息抽取 [J]. 电子学报，2005, 33(2):236-240.

[38] 燕杨，文敦伟，王云吉，等．基于层叠条件随机场的中文病历命名实体识别 [J]. 吉林大学学报：工学版，2014, 44(6):1843-1848.

[39] 齐保元，史忠植．基于维基百科和条件随机场的领域主题词抽取方法 [J]. 高技术通讯，2014 (6):602-608.

[40] 马萌，唐卓，李仁发，等．基于条件随机场的改进型 BLP 访问控制模型 [J]. 计算机科学，2015, 42(08).138-144

[41]Scott D, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6):391-407.

[42] 余正涛，樊孝忠，郭剑毅，等．基于潜在语义分析的汉语问答系统答案提取 [J]. 计算机学报，2006, 29(10):1889-1893.

[43] 李珀瀚，何震瀛，向河林．一种基于链接聚类的查询扩展算法 [J]. 计算机研究与发展，2011,48(z2):197-204.

[44] 曹春萍，崔海船．基于 LSA 和结构特性的微博话题检测 [J]. 计算机应用研究，2015(9):2720-2723.

[45] 张世博，刘博爱，柳朝阳，等．基于潜在语义分析的文档检索设计方法 [J]. 北京石油化工学院学报，2015(02):37-42.

[46] 刘挺，车万翔，李生，等．基于最大熵分类器的语义角色标注 [J]. 软件学报，2007, 18(3): 565-573.