中文问句的形式分类和资源建设

黎江涛[1], **饶高琦**[1]
(1. 北京语言大学汉语国际教育研究院，北京100083)
eric_lijiangtao@163.com, raogaoqi@blcu.edu.cn

摘要

本文归纳了问句形式在问句语料筛选中的作用，探索了问句分类必需的形式特征，同时通过人工标注建设了中文问句分类语料库，并在此基础上进行了基于规则和统计的分类实验，通过多轮实验迭代优化特征组合形成特征规则集，为当前问答提供形式上的分类基础。实验中，基于优化特征规则集的有限状态自动机可实现宏平均F1值为0.94；统计机器学习中随机森林模型的分类效果较好，F1值宏平均达到0.98，表明问句形式分类具有相当可行性和准确性。

Formal classification and resource construction of Chinese question

Jiangtao Li[1], **Gaoqi Rao**[1]
(1. Research Institute of International Chinese Language Education,
Beijing Language and Culture University, Beijing 100083, China)
eric_lijiangtao@163.com, raogaoqi@blcu.edu.cn

Abstract

This paper summarized the role of question forms in question corpus screening, explored the necessary formal features of question classification, and constructed a Chinese question classification corpus by manual tagging. On this basis, this paper have conducted classification experiments based on rules and statistics, and optimized feature combinations to form feature rule sets through multiple rounds of experiments, which provides a formal classification basis for current questions and answers. In the experiment, the finite state machine based on the optimized feature rule set can achieve a macro average F1-score of 0.94; The classification effect of random forest model is better, and the average F1-score reaches 0.98, which indicates that the classification of question forms is feasible and accurate.

1 引言

问句分类的效果直接影响问句理解。传统的中文问题分类主要是根据答案对象的类型划分，如询问人物、地点、时间、数量等，曹志娟（2005）还在此基础上增加疑问词短语分类、问题标准型、特征词分词来增强计算机识别问题的能力的方法，刘朝涛（2008）则进一步将疑问词模式与问题类型对应起来，进行了基于疑问句句型识别的问题理解研究。在这些分类任务中，问句的形式只是作为分类的辅助特征。

第十九届中国计算语言学大会论文集，第108页-第117页，海南，中国，2020年10月30日至11月1日。

实际上，一定的问句形式下的问句类别可以对应一定的问句功能，但这方面的理论在问句理解实践中并没有得到重视；相反，随着数据集的增加，问句覆盖的范围越广，复杂的问句形式特征被当作解决新问题的补丁不断地添进，使得问句分类标准越来越复杂。如果能在问题分类中先提供一个形式分类接口，再按照不同问句形式下对应的问句功能对问句做进一步分类，那么就能在形式上不遗漏任何问句，同时也能在分类过程中根据问句形式定位问句的具体功能。所以在现有问句分类研究基础上，提倡问句的形式分类具有深刻意义。

2 问句的性质

2.1 问句的范围

傅惠钧曾根据“疑”和“问”的组合划分出“有疑有问、有疑无问、无疑有问、无疑无问”四类。很明显“有疑有问”和“无疑无问”均可以明显地判断句子是否为问句，问题就集中到了“有疑无问”和“无疑有问”这两类句子上。

先说“有疑无问”。吕叔湘给出过例句“也许会下雨吧”,表示有传疑但不发问。这个例句后面既可以加上问号标记也可以不加上问号标记，邵静敏根据这种对比指出，两种情况表达的疑问程度是一致的，区别仅仅在于是否发问，即是否要求对方表示态度。所以由此可见，从问答理解的角度来看，回答的前提是存在发问，所以将没有发问意图句子排除在分析目标之外是合理的，这也符合问句提出的预期，即发问-解答。本文也将根据是否有发问意图来区分疑问问句和非疑问问句。

再说到“无疑而问”，学界对这类句子众说纷纭，普遍认同的一个观点是反问句（也叫反诘问句）可以作为“无疑而问”的典型代表，马氏文通中将这类句子的功能称为“传信”，与“传疑”相对。判断这一类句子必须要明确一点：“信疑”皆是从说话人的意图中推断出来的，而不是站在对话的全知视角或是听话人视角。如果“信疑”脱离了说话人意图，那么问句就可能会随着不同的回答而有不同的定性，在疑问句和反问句之间摇摆不定。例如，“谁欠你钱？”，说话人如若想表达“我不欠你钱”的意思该句则是反问句，但如果不考虑说话人的意图，仅考虑该问句的可回答性，也可以说“某某欠了钱”，但这明显已经脱离了说话人想表达的意图。所以“无疑而问”本质上是不含发问意图的句子。对于问句理解来说，如果是在问答系统中，“无疑而问”的问句显然不能成为分析的对象，因为句子本身不存在疑问点，也就无法对问题做出回答；但如果从人机对话的角度来说，“无疑而问”更偏向是一种套着疑问形式的表达方式，这样的句子往往承载着说话人的某些观点、意图，计算机要做的就是要在遵守语用交际原则的情况下回应这些句子，此时的“无疑而问”类句子无疑应该纳入该研究的分析对象。

而本文讨论分析的对象以含有说话者发问意图的问句为主，对不含发问意图的问句只做简单的功能探讨。

2.2 问句的分类

含有说话者发问意图的问句通常又叫疑问句，按照形式上的不同，它们又可以分为四类：是非问、特指问、选择问、正反问。

是非问。结构类似陈述句，一般用升调，句尾一般有“？”，句尾有时兼有语气助词“吗”显化疑问语调，也可以用“啊、哇”，但不可用“呢”。例如：“21世纪人类将要开发月球吗？”

特指问。用疑问代词代替未知部分，常用的疑问代词有“谁、什么、哪儿、怎么、多少”等，句尾有时用“呢”或“啊”，不用“吗”。例如：“这是哪里啊？”

选择问。有并列的若干分句，前后分句常用“是”“还是”相呼应，有时用语气助词“呢”或“啊”，但不用“吗”；另外，选择问中语词助词和连词可以兼有。例如：“是吃西餐还是吃中餐？”

正反问。通常包含否定词“不”“没有”，不采取复句的形式，在谓语中心或补语中用肯定和否定并列形式来提问。具体情况如下表所示：

形式	例句
V/Adj+不+V/Adj	你饿不饿？
V+不+V+X	你吃不吃饭？
V+X+不+V	你吃饭不吃？
V+X+不+V+X	你吃饭不吃饭？
V+否定词（不/不成/否）	你吃饭不？
V+补+V+否定词+补	这饭你吃得了吃不了？

表 1　正反问形式及例句

2.3　问句形式概述

问句形式是判断问句的依据，主要包括语音语调、标点形式、句法格式、特征词。语音语调主要指句子的句调，一般问句的句调均以声调为主。标点形式主要指问号，这是问句的主要形式标记。句法格式指不同问句类型由特定句法单位构成的格式，按照问句类别可以分为是非问句法格式、特指问句法格式、选择问句法格式和正反问句法格式。而特征词是指能够帮助判断问句类别的典型词语，比如特指问的疑问代词，选择问中的“还是”等。

根据承载问句的介质不同，可以从两个方面来说明问句形式的作用和特点。

1.语音问句识别中，本该使用标点停顿的地方用语音停顿替换，表达疑问的标点形式用相应的语音语调替换，因此主要是语音语调、句法格式和特征词等在语音问句识别中起作用。

2.文本问句识别中，标点完全代替语音信息起到停顿、疑问语气的作用，所以标点形式、句法格式和特征词在识别中占据主要地位，其中标点形式尤以问号“？”为主。

所以在问句判别的领域中，语音语调信息与标点信息形成对立，句法格式和特征词两者相互补充，甚至两者还互有交叉，一定情况下还可以相互转换。问号往往就是问句的标志。

3　问句形式在问句分类中的作用

问答系统一般由问题分类、查询扩展、搜索引擎、答案抽取以及答案排序选择多部分组成。问题分类是建构问答系统的重中之重。而对于问题分类而言，目标问句语料的筛选又是问题分类的前提条件。质量高的问句语料可以提高问题分类及后续工作的效率。

通常提取的问句对象都是文章中的对话内容，即引号内的问句，这样做有两个好处：一，可以保证问句提取的自然度，能够最大限度地模拟日常问答；二，为判定问句的意图提供了条件，可以通过问句的上下文来推测说话人的意图从而判别句子是“有疑而问”还是“无疑而问”。而文本问句的形式在上文已提到包括标点形式、句法格式、特征词三类，下面将围绕这三点说明问句形式在问句语料筛选中的作用。

3.1　标点形式

问号是问句的主要标志，根据问句中问号的多少可以把问句大致分为以下两类。

(1)问句中存在多个问号

一般包括两种情况：其一，问句是个连续问句群，例如：“你是谁？你来自哪里？”，此时问句能被分解为若干个单独的问句；其二，问句是选择问句的一种形式变体，如：“你要喝果汁？还是牛奶？”，此时每一个以问号成句的句子不能单独理解，必须将问句群看作一个整体，因为从语义上来说，单独的问句语义并不完整，只有问句群才能够表达完整的意义。

连续问句往往不能成为问句分类分析的典型语料，但它作为问句的组合形式一种，能拆解成若干个问句来理解。而选择问句的形式变体实际上是标点的一种误用，在形式上与连续问句相同，但它在问句语料中也占据一定数量，应算作问句分类分析中的典型语料，否则会使选择问在自然语言中的比例不能得到正确的反映。

(2)问句中只存在一个问号

又可根据问句内部是否存在标点分为两类：一类是组合问句群，另一类是常规问句。汉语中的连续问句可以用逗号连接，以问号煞尾。此时句子并不是单一的问句，而是一个组合式的问句群，不能成为问句分类分析的典型语料。例如：“我是谁，来自哪里，又将会去何处？”

3.2 句法格式

问句中存在一些包含特殊句法格式的句子，这类句子如若按照形式去分析，其问句理解的复杂程度相较其余典型问句要大得多，可细分为以下几类。

(1)“W+呢”类

“W+呢”类又可细分为“NP+呢？”和“VP+呢？”两类。

“NP+呢”在形式上没有明显的问句形式特征，但可以根据其前行句在深层语义上对其进行不同的扩展。例如：

清少爷，你这一向好啊？—好，您老人家呢？（曹286）

“您老人家呢？”可以作“您老人家好不好”、“您老人家怎么样”、“您老人家好吗？”等三种语义理解，且这三种理解分别属于正反问句、特指问句、是非问句。所以可以看出，理解这类问句在语义上需要借助语用信息，在形式上做进一步分类也容易出现分歧。

“VP+呢？”，邵敬敏（1997）将这类问句分成了三种类型：

甲（要是）VP呢

乙（要是）VP，怎么办呢

丙（要是）VP呢、（要是）VP，怎么办呢

形式上来看，“VP+呢”类问句中，甲句型最简洁，乙句型最完整，丙句型为兼备甲乙句型的特点，三种类型都能表达相同的语法意义。另外从功能上来看，“VP+呢”类问句既能表示假设也能表示询问，但无论作何种功能，这类问句的理解同样需要语用信息，且问句往往以甲句型出现。当然，如果考虑到根据深层语义补足原有形式的话，这类问句应是特指问，即根据完整句型乙推出。所以，在问句语料的筛选中，这类问句往往因为其功能的复杂性排除在典型问句的筛选范围之外。

(2)省略疑问成分的问句

一些问句还存在一些缺省疑问成分，但在一定语境下仍旧可以表达疑问。例如两人初次见面时，一方可以用“您是？”提问，意为“您是哪位/您是谁”；对对方的变化感到疑问，可以用“您这是？”提问，意为“您这是怎么了？”。这类句子在省略了疑问词的情况下，以是非问句的形式的存在，但如果根据深层语义补足原有形式，这类句子大多属于特指问，且要理解句子省略了何种疑问词也需要结合语用信息才能说明。所以，在问句语料的筛选中，这类问句往往排除在典型问句的筛选范围之外。

(3)回声问句

回声问是“对话的问题”，具有更多的交际价值，但对于问题本身来说它需要依托于一定的语境才能理解它的含义或补全它的完整问句形式。所以，在问句语料的筛选中，这类问句往往排除在典型问句的筛选范围之外。如下例。

鲁侍萍　　老爷那种绸衬衣不是一共有五件？您要哪一件？

周朴园　　要哪一件？（曹63）

3.3 特征词

不同的问句类型有自己的特征词，这些特征词是判定句子类别的标志。如果特征词出现了错误，就可能影响问句的分类，进而影响问句的理解。主要表现为疑问代词，例如：“在中国有好多人在看摇滚”、“浮云是神马意思”。前者的“好多”带有地域方言色彩，应属疑问词，对应标准式“多少”；后者的“神马”是网络词汇，属于疑问词“什么”一种语言变体。如果在问句理解中，不能对这些形式的问句加以区分，容易在语法结构和语义的分析上造成偏差，最后影响问句的理解。由此可知，在问句语料的筛选中，还需要主要特征词的错写对语料筛选的影响。

所以，标点形式、句法格式、特征词在问答系统的任务中具有举足轻重的作用，规范的问句形式和正确信息同等重要，规范的问句形式是保障问句语料正确性、完整性的基础。

3.4 问句特征选取与特征集构建

根据语言学对是非问、特指问、选择问和正反问的定义，可以进一步将句法格式和特征词细化为疑问格式、语气词、语气副词以及疑问代词四大类，这四大类在具体语料中又可以细分为七个小类：语气词“呢”、语气词“吗”类、疑问代词、语气副词、是非问疑问格式、正反问疑问格式以及选择问疑问格式。注意到在是非问句中，一些句子的显性问句标记过少，不含七小类特征中的任一特征，如是非问“他走了？”，所以为避免无特征匹配是非问句的情况，我们将

增加一类补充特征，即当问句不存在疑问代词、正反问疑问格式和选择问疑问格式任一特征时，默认该句有补充特征，否则没有。

4 问句语料库建设

4.1 数据标注

为测试问句形式对语料的筛选的有效性，同时也为问句数据做进一步的分类，我们从一批小说语料中选取了2400个问句并将这些句子分成三组，每组800句，交由6位语言学专业的研究生两两标注，问句的分类标准主要参照上文的问句定义。是非问、特指问、选择问和正反问分别以数字1、2、3、4表示。一个完整标注的问句如下所示，问句前的数字代表问句的类别。

1：还有其他异常情况吗？（问句标注示例）

经统计，三组在没有对抽取句子进行形式上的筛选之前，一致率分别为：0.855，0.82，0.845，平均一致率达0.84；而经过对抽取的句子按照常规问句形式的筛选，剔除句意理解与语用信息相关的句子后，一致率分别为：0.965，0.943，0.894，平均一致率达到0.934。可见，问句形式有助于提高问句标注的一致率。同时，以上实验也表明，根据问句的语言学特征来判定问句种类并不是一件过于复杂的任务，在此基础上可以继续扩大问句标注规模。

4.2 问句分布情况

经标注及筛选后，我们得到1679句问句。在此基础上，我们还标注了一批形式上较为规整，不依赖语境且可以自足分析的百度知道问句数据集，共2621句。各数据集的问句分布如下所示：

	小说	百度问答	总和
是非	651（38.8%）	527（20.1%）	1178（27.4%）
特指	749（44.6%）	1857（70.9%）	2606（60.6%）
选择	23（1.4%）	40（1.5%）	63（1.5%）
正反	256（15.2%）	197（7.5%）	453（10.5%）
总和	1679	2621	4300

表 2　问句数据分类分布情况

从上表可以看出，特指问在问句中数量与占比均为最高，其次是是非问、正反问以及选择问，这一定程度上也反映了这四类问句在自然语言中的大致分布情况。

此外，在不同数据集上，四类问句的分布也稍有差异。在小说问句中，是非问与特指问占比相当，特指问略高于是非问；而在百度问答问句中，特指问占比超过70%，远远超过是非问的20.1%，这一定程度上呈现了小说问句与百度问答问句的特点，两者既有联系又有区别。百度知道问句是属于百科问答式问句，对概念的提问、事件发生的原因等问句比例较大，致使包含疑问代词的问句较多，也就造成了特指问句在百度问答数据集上分布较多。而小说问句中并没有这种明显的倾向性，使得是非问句与特质问分布较为均匀，同时小说问句的语境也更接近于日常生活场景的问句使用情况。

本资源将向学术界开放使用[0]

4.3 问句特征在语料的分布情况

上文我们整理出了问句的八个小类特征，分别用F1–F8来表示，在语料库中，这些形式特征的计量统计如下：

[0]链接：https://pan.baidu.com/s/1R9se1GPUcQcPLpkzZaGaiw　　提取码：ppux

特征	类别	数量/占比	说明
F1	语气词	253/5.88%	是否有语气词“呢”
F2	语气词	802/18.65%	是否有语气词“吗、么、嘛、吧”
F3	疑问代词	2790/64.88%	什么、如何、哪、哪里、几、谁、啥、为啥、何、何不、为何、为什么、怎么、咋、干吗、多 X
F4	疑问格式	1292/30.05%	是非问：能愿动词+语气词
F5	疑问格式	66/1.53%	选择问：X 还是 X
F6	疑问格式	479/11.14%	正反问：X 不 X、X 不、X 没有、X 不成
F7	语气副词	96/2.23%	莫非、莫不是、难道、难不成、到底、何必、何须、何妨、何曾、何尝、何不、何苦、究竟、岂
F8	补充特征	1053/24.49%	F3、F5、F6 的补充特征

表 3　问句特征分布

表3中，各特征多寡是和不同类型问句占比有关的，部分特征分布情况甚至可以直接反应问句整体的分布情况。如特征F3、F4、F5、F6的占比与四类问句在数据集中的分布情况相当，反映出特指问和是非问在问句中占比较大，选择问相较正反问数量更少。另一方面，疑问格式与疑问代词特征的占比相加大于100%，说明问句分类的结果不是仅由疑问格式决定的，至少存在一个问句包含多个疑问格式或疑问代词的情况，问句分类的复杂性也体现于此。

5　基于问句形式的自动分类

5.1　基于统计机器学习的多特征分类

从问句特征到问句种类的识别实际上是一个从特征到分类的问题。其过程就是把每个问句中能匹配的问句特征转化为可量化的特征向量，最终将特征向量映射到该问句所对应的类别。根据表3的问句特征我们对语料中的问句进行向量化处理，含有指定特征即将特征所在维度的向量值记为1，反之记为0；是非问、特指问、选择问、正反问分别用1、2、3、4表示。如表4。

例：《亮剑》中李云龙求婚对白是第几集啊？								
F1	F2	F3	F4	F5	F6	F7	F8	CLASS
0	0	1	0	0	0	0	0	2

表 4　特征转换示例

在获得多维度向量和其对应的分类标签后就已经进入了根据特征分布进行问句分类的任务。根据以往分类任务经验，本文拟用支持向量机、逻辑回归分类器、贝叶斯分类器、K近邻、决策树以及随机森林等六种机器学习方法来验证问句特征对问句的分类效果。

此外，不同特征数量的选择对问句分类的结果也会有影响。F1至F8等特征近似于从语言学角度对问句形式进行列举，但哪些特征组合能够使得问句分类效果最佳需要进一步实验证明，所以本文将对F1至F8等8个特征做排列组合，共计225种组合结果。

我们再将人工标注的1679句小说问句作为训练语料，后续标注的2621句百度知道问句作为测试语料，将机器学习方法与特征组合结果结合后，下文将从多角度来分析模型的分类效果。

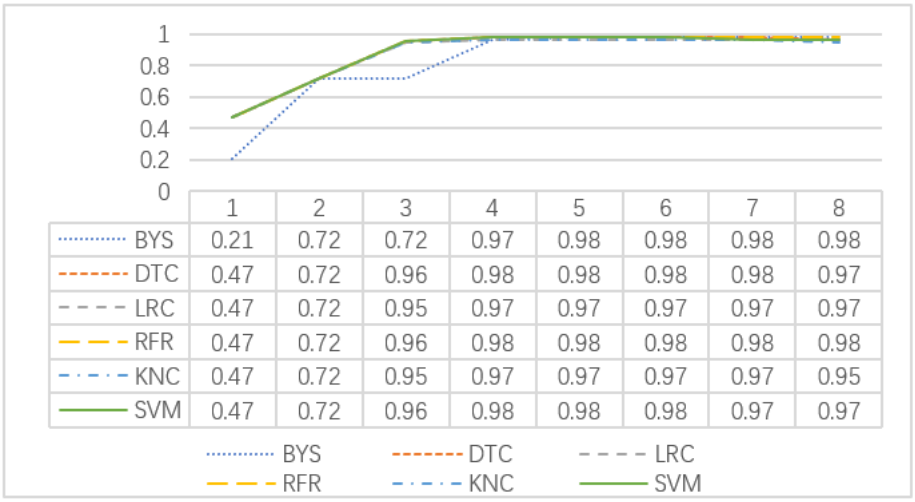
5.2　基于形式特征集的有限状态自动机

不同问句特征对于问句分类判定的贡献不同。根据表3我们可以把特征的覆盖率作为问句特征对问句分类的贡献程度，便有如下排序：语气词=疑问格式¿疑问代词¿其他。那么基于此，

我们可以让贡献大的问句特征优先参与问句判定，而问句特征无法覆盖的问句可以归入形式最多样的是非问，这样问句分类就是在一个有限规则内进行，只要输入一个问句，必定可以输出问句所属的类别。这样就完成了基于形式特征集的有限状态自动机构建准备。

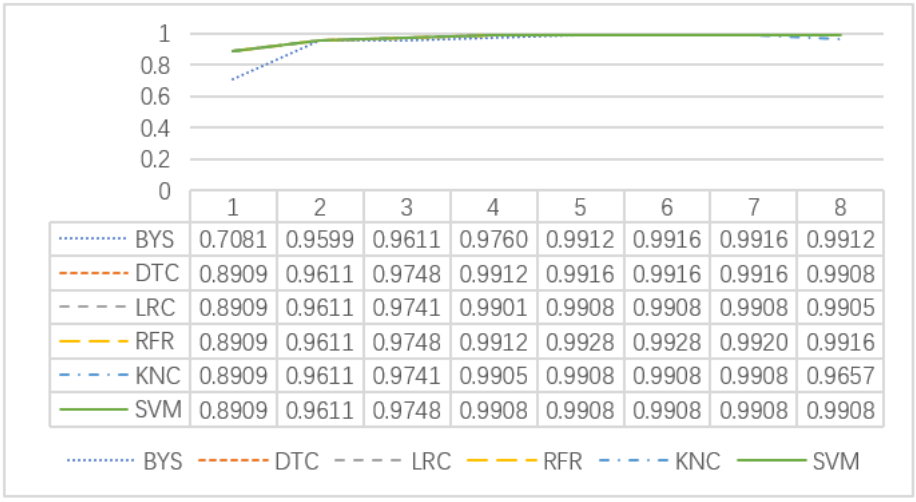
5.3 实验结果

由于问句类别包含四类，我们主要从宏观的角度来分析模型随特征数量变化的情况,即通过不同模型分类的F1值宏平均和微平均分析问句分类整体的优劣(如下两图)。考虑到在某一特征数量下，存在不同特征组合影响分类结果准确性的情况，我们只选取某一特征数量下最好的结果作为比较对象。



	1	2	3	4	5	6	7	8
BYS	0.21	0.72	0.72	0.97	0.98	0.98	0.98	0.98
DTC	0.47	0.72	0.96	0.98	0.98	0.98	0.98	0.97
LRC	0.47	0.72	0.95	0.97	0.97	0.97	0.97	0.97
RFR	0.47	0.72	0.96	0.98	0.98	0.98	0.98	0.98
KNC	0.47	0.72	0.95	0.97	0.97	0.97	0.97	0.95
SVM	0.47	0.72	0.96	0.98	0.98	0.98	0.97	0.97

图 1　各模型F1值在特征数量上的宏平均



	1	2	3	4	5	6	7	8
BYS	0.7081	0.9599	0.9611	0.9760	0.9912	0.9916	0.9916	0.9912
DTC	0.8909	0.9611	0.9748	0.9912	0.9916	0.9916	0.9916	0.9908
LRC	0.8909	0.9611	0.9741	0.9901	0.9908	0.9908	0.9908	0.9905
RFR	0.8909	0.9611	0.9748	0.9912	0.9928	0.9928	0.9920	0.9916
KNC	0.8909	0.9611	0.9741	0.9905	0.9908	0.9908	0.9908	0.9657
SVM	0.8909	0.9611	0.9748	0.9908	0.9908	0.9908	0.9908	0.9908

图 2　各模型F1值在特征数量上的微平均

通过F1值宏平均以及微平均的筛选，得出随机森林模型在特征数量为5时，分类模型的F1值宏平均和微平均达到最高值，分别是0.99和0.98。此时选取的特征分别是F2、F3、F4、F5、F6，即语气词“吗、么、嘛”，是非问疑问格式，选择问疑问格式，正反问疑问格式以及疑问代词。随后，我们将百度知道的2621条问句作为实验对象，采用有限状态自动机分类和随机森林模型分类的效果如表5所示:

分类方法		精确率	召回率	F1 值
有限状态自动机	是非	0.96	0.87	0.91
	特指	0.95	0.99	0.97
	选择	0.93	0.95	0.94
	正反	0.99	0.88	0.93
	F1 值宏平均	0.94		
	F1 值微平均	0.96		
随机森林	是非	0.99	0.99	0.99
	特指	1.00	0.99	1.00
	选择	0.97	0.95	0.96
	正反	0.96	0.99	0.98
	F1 值宏平均	0.98		
	F1 值微平均	0.99		

表 5　随机森林模型和有限状态自动机分类结果

从模型整体效果来看，随机森林的F1值宏平均和微平均相较有限状态自动机的分类结果提高了0.04和0.03。这一方面说明了有限状态自动机分类的方法对问句分类也有较好的效果，通过特定的问句规则可以有效覆盖大多数问句，但这种方法往往会出现召回率偏低的情况，无法处理一些组合特征；另一方面也说明了随机森林模型在进行问句分类过程中具有更好的分类效果。

从各个问句小类的分类结果来看，特指问的F1值在两种分类方法中均为最佳，但在是非问、正反问中，有限状态自动机的F1值却偏差随机森林颇多，体现了是非问句、正反问句的判定受形式特征的多样性明显，单一的问句特征不足以覆盖大多数此类问句；而对于正反问句来说，有限状态自动机的方法在精确率上高于随机森林模型，说明正反问的问句形式特征对正反问的判定具有较强的作用，但在召回率上低于随机森林模型，与是非问情况相同，也体现了正反问形式特征的多样性。

5.4　随机森林模型错例分析

按照错判的类别分为以下典型几类：

例句：有谁能帮忙解释一下，吴尊拍这张照片的这时候在干吗?

上述句子是特指问句，却被错判为是非问句。究其原因是“干吗”作为疑问代词，词中含有“吗”字，使得模型误以为含有是非问特征词，加之语气词“吗”属于强形式特征，模型会倾向于将问句分为是非问。

例句：听说有位明星自杀了真的假的?

上述句子是选择问句，却被错判为是非问句。这是由于选择问句的形式不能覆盖原问句形式所致。选择问句中最典型的疑问格式是“X还是X”，但也存在选择并列的情况，如上句。并列的成分可以是谓词性成分也可以是体词性成分，但不论何种，并列成分在结构上总存在一定的相似性。也正由于这个原因，这类问句形式上难以量化，本实验的模型无法对此类问句的识别效果有限。

例句：韩庚什么的，没上09央视春晚吧?

上述句子是是非问句，却被错判为特指问句。这是由于原是非问句缺少明显的是非问形式特征，但却存在次强形式特征疑问代词，使得模型倾向于将原问句判断为特指问。从另一个角度来说，上述问句的疑问代词“什么”并不是疑问点，而是表示虚指，要正确对此类问句分类必须分清句中的疑问代词是否表示疑问。

例句：可最近心情又是不好，吃药都没作用啦，难道说还是抑郁症?

上述句子是是非问句，却被错判为选择问句。这是由于句中出现了选择问的强形式特征，但“还是”前后连接的并不是选择的对象。结合前文中选择问句错判的例句，可以得出对于选择问问句，精确率较其他分类低，是由于连词“还是”作为选择问的典型特征易与状中结构“还是”混淆，召回率低则是选择问存在不易归纳的问句形式所致。

例句：如何判断经营者决策是否正确?

例句：怎么看哈士奇纯不纯?

上述句子是特指问句，却被错判为正反问句。这是由于句中同时存在正反问的强形式特征和特指问的次强形式特征所致，正反问强形式特征对问句分类的直接增益更大，所以原句分为正反问句。实际上，上句中的“经营者决策是否正确”和“哈士奇纯不纯”并不是原问句的疑问焦点，“经营者决策是否正确”等价于“经营者决策的正确性”，“哈士奇纯不纯”等价于“哈士奇的纯度”，要解决这一个问题，需要介入问句焦点信息的识别工作。

6 结论

本文详细分析了问句形式在问句语料筛选和问句分类中的作用，并以此筛选、标注了4300句问句，构建了目前最大的中文问句分类语料库；此外，还借鉴语言学上的问句形式特征，利用多种机器学习方法构建问句分类模型。根据问句的类别进行分布情况统计，得出特指问在目标数据集中数量与占比均为最高，其次是是非问、正反问以及选择问，这一定程度上也反映了这四类问句在自然语言中的大致分布情况。

问句形式自动分类实验表明，当形式特征集为语气词“吗、吧、么、嘛”、是非问疑问格式、疑问代词、选择疑问格式、正反疑问格式时，对于问句分类具有较高的准确度，表明句尾语气词不仅区分问句，也是问句内类型分类的最有力特征。最终在随机森林模型分类下F1值宏平均达到0.98，F1值微平均达到0.99，特指问分类的F1值最高可达1，是非问分类的F1值达到0.99，正反问分类的F1值达到0.98，选择问分类的F1值达到0.96。本研究当中的问句数据一定程度上可以反映自然语言中问句句类的分布情况，对具体领域中的问句分布研究有一定的参考价值。

同时，可以看出，问句的形式分类本身是一个特征较为明确，规则性较强的问题，使用规则系统也可以获得不差的效果。因此我们认为，在为问句分类时可以增加一个问句形式分类的接口，一方面问句形式自动分类的精度有一定的保障，另一方面可以在这个问句形式分类接口可以集中处理所有问句形式的问题，为问题进一步分类提供分类基础。

文中也存在以下不足：第一，在资源建设方面，本文采用的数据集规模仍需要扩大来进一步考察问句形式特征的效果，届时大规模的数据集可以给深度学习提供充足的泛化空间，将深度学习的方法用于问句分类，以此来与现有分类效果做比较；第二，本文的着重研究的是问句形式对含有说话人发问意图的疑问句语料的筛选和分类问题，而问句形式对于反问句的分析有何作用尚进一步分析。不过已知的是，反问句与疑问句在问句形式上差异不大，只存在有无发问意图的区别，所以通过问句形式来识别反问句效果可能不理想。但可以通过研究一般陈述句变为反问句所需要的问句形式条件，建构反问句形式的意图表达机制，来完成反问句自动生成，从而达到机器表达具有“拟人性”的目的。

参考文献

曹志娟,李祖枢,刘朝涛. 2005. 自动问答系统中的问题理解研究. 计算机科学,2005(11):158–160+230.

傅惠钧. 2008. 关于疑问句的性质与范围. 浙江师范大报(社会科学版),2008(5):77–82.

范继淹. 1982. 是非问句的句法形式. 中国语文,1982(6):426–434.

郭婷婷. 2005. 现代汉语疑问句的信息结构与功能类型. 武汉大学.

黄伯荣. 2017. 现代汉语. 北京:高等教育出版社, 101–107.

刘朝涛. 2010. 中文问答系统中的句型理论及其应用研究. 重庆大学.

刘朝涛,李祖枢. 2008. 基于疑问句句型识别的问题理解研究. 计算机科学,35(12):151–153+189.

吕叔湘. 1985. 疑问•否定•肯定. 中国语文,1985(4).

李宇明. 1997. 疑问标记的复用及标记功能的衰变. 中国语文,1997(02):97–103.

陆俭明. 1982. 由“非疑问形式+呢”造成的疑问句. 中国语文,1982(6).

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

邵敬敏. 1996. 现代汉语疑问句研究. 上海:华东师范大学出版社.

文勖,张宇,刘挺,马金山. 2006. 基于句法结构分析的中文问题分类. 中文信息学报,2006(02):33–39.

袁毓林. 1994. 正反问句及相关的类型学参考. 北京:北京语言学院出版社.

镇丽华,王小林,杨思春. 2015. 自动问答系统中问句分类研究综述. 安徽工业大学学报(自然科学版),32(01):48–54+66.

朱德熙. 1982. 语法讲义. 北京:商务印书馆, 202–205.