多领域自然语言问句理解研究

冶忠林[1] 贾 真[1] 尹红风[2]
(西南交通大学信息科学与技术学院 成都 610031)[1] (DOCOMO Innovations 公司 帕罗奥图 94304)[2]

摘 要 问句理解是问答系统的主要任务之一。现有的问句理解方法大多是针对简单句的,且侧重于某种句式结构的理解。提出一种多领域问句理解研究方法,其涉及领域包括人物类、电影类、音乐类、图书类、游戏类、应用类。首先基于CRF算法对问句进行分类和主体识别,然后使用谓词词典和句法分析识别出问句的谓词,最后提出一种谓词消歧方法来解决相同问句具有不同表达方式的问题。实验结果表明,在封闭测试中,所提方法的问句分类和主体识别的平均F-measure值分别为93.88%和92.44%,谓词识别和问句理解的平均准确率分别为91.03%和81.78%。因此,所做的工作基本能满足问句理解的需求。

关键词 问答系统,问句理解,谓词消歧,问句分类,主体识别
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.06.036

Research on Multi-domain Natural Language Question Understanding

YE Zhong-lin[1] JIA Zhen[1] YIN Hong-feng[2]
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)[1]
(DOCOMO Innovations Incorporation, Palo Alto 94304, USA)[2]

Abstract Question understanding is one of the main tasks of question answering system. Current question understanding methods aim to solve semantic understanding of simple sentences or specific structure sentences. The method proposed in this paper addresses multi-domain question understanding which includes people, movie, music, book, game, and application domains. Firstly, the question classification based on CRF algorithm and the subject recognition based on CRF algorithm approach are presented. And then the prediction dictionary and semantic analysis are applied to recognize prediction. Finally, the prediction disambiguation method is proposed to deal with the problem that prediction in question has different ways of expression. Experimental results show that the average F-measure value is 93.88% and 92.44% in question classification and semantic analysis experiments. The average accuracy is 91.03% and 81.78% in the prediction recognition and question understanding. Thus, the works in this paper can meet the needs of question understanding.

Keywords QA system, Question understanding, Prediction disambiguation, Question classification, Subject recognition

1 引言

问答系统主要由3个部分组成:问句理解、信息检索、答案抽取[1-2]。用户使用自然语言方式提问,问答系统返回简洁准确的答案[3]。现有的问句理解框架主要是解决单一领域的简单问句的理解和分析问题,比如旅游类、地图类等单领域简单问句的理解。本文提出一种面向多领域的问题理解方法,所提方法的步骤主要包括问句分类、主体识别、谓词识别、谓词消歧。文本中的多领域问句理解不仅能够解决简单问句的理解问题,而且能够很好地理解跨领域的问句或较为复杂的问句。例如"《平凡的世界》的作者的生日是什么时候",该问句是跨领域的问句,需要在图书类和人物类知识库中联合查询才能找到答案。本文首先进行问句分类和主体识别,然后基于谓词词典和句法分析将问句转化为如下形式:[[平凡的世界,作者],生日]("[ ]"表示"问题元");再通过谓词消歧将问句中的谓词和知识库中的谓词对齐,并将其转化为:[[平凡的世界,作者],出生年月]。通过以上处理,问句可被转化为计算机可理解的结构化查询语句。在人物类、电影类、音乐类、图书类、游戏类和应用类的5组实验数据上(每组100个问句),本文提出的问句理解方法的平均准确率达到81.78%。

2 相关工作

问句理解中的问句分为事实型、定义型、列举型、复杂型

到稿日期:2016-04-02 返修日期:2016-09-26 本文受国家自然科学基金(61572407,61262058),国家科技支撑计划课题(2015BAH19F02,2016G04001),中央高校基本科研基金(2682015CX070)资助。

冶忠林(1989—),男,硕士生,CCF会员,主要研究方向为自然语言处理;贾 真(1975—),女,博士,讲师,主要研究方向为信息抽取、知识工程,E-mail:zjia@swjtu.edu.cn;尹红风(1967—),男,博士,教授,主要研究方向为语义搜索、大数据。

型[4-5]。魏楚元等[6]在理解句子时找出含有事件信息的句子中的疑问焦点块、问题主题块、问题事件块[6]，以辅助句子理解。黄沛杰等[7]将语法信息和语义特征相结合，实现了问答系统中的问题理解，并将其成功地应用到中文手机导购对话系统中。刘朝涛等[8]提出了针对疑问句的问句理解框架。陈永平等[9]提出了基于主题和焦点的问句理解方式。马莉等[10]提出了句子模式匹配与关键词相结合的方法，改进了以往仅用模式匹配问句理解的方法。侯永帅等[11]使用分类算法，得到对时间较敏感的句子，然后获取句子的时效区间，最后根据时效区间对检索之后的答案进行过滤，该方法能够对时效比较敏感的句子做出检索识别。赵东岩等[12]基于句子结构提出了一种启发式的命名实体识别方法，并且使用知识库对谓词进行消歧，然后将句子转化为查询语言。在国外，Lehmann 等[13]同时使用句子的依存关系和特征模板理解问句。Brown 等[14]将自然语言问句拆分为线索和子线索两部分，然后通过线索找到答案。Cimiano 等[15]首先对知识库中的本体构造出相关的本体特征，然后利用这些特征来理解问句和回答。在基于 Freebase 的问答系统测评中，Berant 等[16-17]使用基于语法分析的方法的准确率达到 39.9%，Bordes 等[18-19]使用信息回归的构建方法的准确率达到 39.2%，Fader 等[20]使用基于开放信息抽取的方法的准确率达到 35%，Bao 等[21]使用基于翻译的方法的准确率达到 44.19%。

3　方法流程

3.1　方法概述

本文方法的流程如图 1 所示。

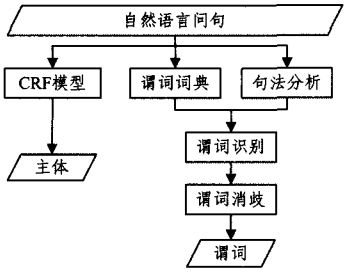

图 1　问句语义理解过程

如图 1 所示，当接收到自然语言问句之后，首先使用 CRF 模型识别出句子中的主体以及该问句的类型；然后进行谓词识别；之后将问句中的谓词进行消歧。进行谓词识别时，先在谓词词典中查找句子的谓词，如果句子中存在谓词，则直接返回谓词；否则，按照问句类别，使用句法分析工具进行句法分析，得到句法分析后的可能的谓词，然后用该谓词与谓词词典中的词计算相似度以确定问句准确的谓词。当得到句子中的主体和谓词之后，句子的语义基本上可以被计算机所理解。但有时识别出的谓词并不一定存在于知识库中，此时需要将该谓词和知识库中的谓词对齐，比如将“妻子”对齐为“配偶”。

本文对人物类、电影类、音乐类、图书类、游戏类、应用类的问句进行分析。将问句属性分为以下类型：

(1)直接属性类：查询主体的直接属性；

(2)一次间接属性类：查询主体的属性的属性，属性之间需要两次跳转；

(3)二次简介属性类：查询主体的属性的间接属性，属性之间需要 3 次跳转；

(4)简介类：查询主体的所有属性类问句；

(5)操作类：对主体做打开、下载、播放等操作。

本文使用问题元将自然语言转化为结构化查询语言，问题元是理解问句的最基本的单位，由问句的主体和属性组成，用“[ ]”表示。在简单的问句中，仅存在单个问句元，但在复杂问句中，通常需要在问题元中套嵌问题元。比如，问句“刘德华的老婆的生日是什么时候”转化为问题元后为“[[刘德华，配偶]，出生日期]”，将问句转化为问句元可方便观察问句理解是否准确。

3.2　问句分类与主体识别

本文使用 CRF 模型同时进行问句分类和主体识别，即使用 CRF 模型建立主体识别模型。在命名实体标注集上加上类别信息，当用模型预测出句子中的主体时，同时返回该主体的类别信息。为了能得到问句的类别，我们定义了 19 个标注集，将“BIE”标注加上类别信息，如“BPEOPLE”表示人物类主体首部。问句分类有利于按照领域知识库查询问题答案或者跨领域查询问题答案。

3.2.1　分类方法和 CRF 特征集的选择

在人物类、电影类、音乐类等六大类句子主体识别的算法的训练过程中，需要定义标注集。标注集主要定义了主体识别时主体的开始、内部、结尾和无关词语，如表 1 所列。

表 1　标注集描述表

标注	含义
B+catagory	当前词为主体的首部
I+catagory	当前词为主体的内部
E+catagory	当前词为主体的尾部
O	当前词不是主体或者组成部分

category 在分类时具有重要作用，当经过 CRF 识别后，可以通过 category 直接获取类别信息。文中 category 的值有 PEOPLE，MOVIE，MUSIC，BOOK，GAME，APP 等 6 个类别，因此，本文定义了 19 个标注标记，例如 BPEOPEL，IPEOPLE，EPEOPLE 等，经过 CRF 主体识别之后，可以得到带有标注集合的主体列表。通过 BIE 标注集可以确定主体的边界，通过 category 可以确定问句的类别。例如，CRF 主体识别后通过 BPEOPEL，IPEOPLE，EPEOPLE 标注的边界可以确定主体，然后通过 PEOPLE 信息可以确定问句为“人物类”问句。该方法可以避免对问句分类再次建立模型的过程。通过对搜索引擎日志中常用问句的分析，发现问句中第一个出现的主体往往是问句最为主要和关键的主体，因此在问句中识别出两个主体时，根据经验值，设置第一个主体为问句的主体。

3.2.2　CRF 特征模板

特征模板是 CRF 模型在训练和测试时选择特征的方式，良好的特征模板能增强边界的组块功能，即中心词前后的特征选择影响组块能力，但过量的特征又会使训练的效果变差，同时会增大训练和预测的开销。

原子特征：获取当前中心词的前后几个词或词性作为特征。本文使用的原子特征模板如表 2 所列。

表 2 CRF 原子特征模板

编号	模版	编号	模版
1	U00:%x[-3,0]	8	U07:%x[-3,1]
2	U01:%x[-2,0]	9	U08:%x[-2,1]
3	U02:%x[-1,0]	10	U09:%x[-1,1]
4	U03:%x[0,0]	11	U10:%x[0,1]
5	U04:%x[1,0]	12	U11:%x[1,1]
6	U05:%x[2,0]	13	U12:%x[2,1]
7	U06:%x[3,0]	14	U13:%x[3,1]

组合特征：相对于原子特征，组合特征加入了词性和词以及多个词或者词性的组合等方式，本文使用的组合特征如表 3 所列。

表 3 CRF 组合特征模板

编号	模板	编号	模板
1	U00:%x[-3,0]/%x[0,0]	10	U09:%x[0,1]/%x[1,1]
2	U01:%x[-2,0]/%x[0,0]	11	U10:%x[1,1]/%x[2,1]
3	U02:%x[-1,0]/%x[0,0]	12	U11:%x[2,1]/%x[3,1]
4	U03:%x[0,0]/%x[1,0]	13	U12:%x[-3,1]/%x[-2,1]/%x[-1,1]/%x[0,1]
5	U04:%x[0,0]/%x[2,0]	14	U13:%x[-2,1]/%x[-1,1]/%x[0,1]/%x[1,1]/%x[2,1]
6	U05:%x[0,0]/%x[3,0]	15	U14:%x[0,1]/%x[1,1]/%x[2,1]/%x[3,1]
7	U06:%x[-3,1]/%x[-2,1]	16	U15:%x[-3,0]/%x[-2,0]/%x[-1,0]/%x[0,0]
8	U07:%x[-2,1]/%x[-1,1]	17	U16:%x[-2,0]/%x[-1,0]/%x[0,0]/%x[1,0]/%x[2,0]
9	U08:%x[-1,1]/%x[0,1]	18	U17:%x[0,0]/%x[1,0]/%x[2,0]/%x[3,0]

本文中单纯使用原子特征的准确率为 77.1%，经过不断改进原子特征和组合特征，最终主体识别的准确率提升到 87.7%。

3.3 谓词识别

3.3.1 谓词词典的构建

谓词词典按照领域谓词建立词典表，每个谓词下有很多别名。本文需要对人物类、电影类等 6 类问句建立 6 张谓词词典。为了构建 6 个领域的谓词词典，需要获得百度百科、互动百科、豆瓣等 6 个领域的 infobox 结构化信息，然后按照领域对所有的 infobox 谓词进行抽取、排序以及谓词合并等。将高频的谓词作为谓词词典中的谓词，同时将低频的谓词进行人工选择后设置为高频谓词的别名或二级谓词，比如“出生日期”下面有别名：“生日”、“出生年月”、“出生时间”等。最终，人物类、电影类、音乐类、图书类、游戏类、应用类的谓词词典中属性的个数分别为：35，35，35，16，22，27。每个属性下又有 1 个或者多个别名。

3.3.2 谓词识别算法

本文使用谓词词典和句法分析相结合的属性谓词识别方法。首先使用谓词词典匹配的方法查找句子中的属性谓词，如果用该方法找不到属性谓词，则用句法分析的方法找出属性谓词。句法分析的优点是当句式不发生变化但其中的某个词发生变化时，可以很准确地识别出谓词；句法分析的缺点是当句式发生变化时，由于句法分析规则有限，因此无法识别出更多的句式结构中的谓语。故谓词词典和句法分析的结合非常有必要。谓词词典可以解决句式变化但属性词不变化的情况，句法分析可以解决属性词变化但句式不变化的情况，因此，将两者结合可以同时解决句式变化和属性词变化的情况；另外，在使用谓词词典方法查找属性谓词时，需要根据问句中属性谓词出现的先后顺序确定属性谓词，否则，在间接属性问句中若属性谓词先后顺序倒置，则会导致无法理解问句。

例如，对于问句“刘德华的生日”，本文提出的属性谓词识别的具体流程如图 2 所示。

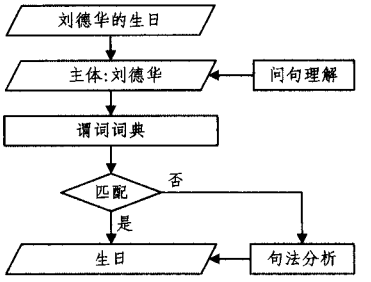

图 2 属性谓词识别方法示例

从图 2 可以看出，问句首先通过问句理解过程识别出主体“刘德华”，之后使用人物类谓词词典匹配该问句，虽然人物类谓词词典中含有属性谓词“出生日期”，但是没有“生日”这个属性谓词，因此需要通过句法分析获得该问句的属性谓词“生日”，之后再对属性谓词“生日”做属性对齐得到“出生日期”。

本文中使用的句法工具为斯坦福句法分析器，通过句法分析可以得到句子的依存关系。例如，对于句子“刘德华的生日是什么时候”，得到的依存关系如表 4 所列。

表 4 句子依存关系分析结果

编号	依存关系	关系描述
1	assmod(生日-3,刘德华-1)	关联修饰
2	case(刘德华-1,的-2)	主语和“的”关系
3	nsubj(时候-6,生日-3)	名词性主语
4	cop(时候-6,是-4)	系动词关系
5	det(时候-6,什么-5)	限定关系
6	root(ROOT-0,时候-6)	依存关系树的根

通过表 4 可以总结出该类问句的依存关系规律：当依存关系“assmod”，“nsubj”，“cop”，“det”同时出现时，“assmod”中的第一个词“生日”即为属性词。句法分析需要收集大量的句子总结句式结构规律，因此需要人工参与的过程。需要注意的是，分类之后再使用句法分析时，针对每个类别可以建立一个句法规则表，以方便管理和修改。同时，各个类别之间的句法分析规则不会产生冲突。

例如，人物类部分问句句法分析规则示例如表 5 所列。

表 5 人物类问句句法分析规则表

句法分析规则集	可解析问句示例	属性谓词
assmod=3	刘德华的老婆的老公的生日是什么时候	(配偶,配偶,出生日期)
assmod=1,nn=1	刘德华老婆的生日是什么时候	(配偶,出生日期)
assmod=1,nsubj=1,dep=0,nn=0,dobj=0,cop=0	刘德华的祖籍在哪里	(祖籍)
nsubj=1,assmod=0,dep=1,dobj=0	刘德华是干什么的	(职业)
nsubj=1,assmod=0,dep=0,dobj=1	刘德华是做什么的	(职业)
nsubj=1,assmod=1,dep=0,dobj=1	刘德华的英语名字叫什么	(英文名)
…	…	…

表 5 中第一列是属性规则匹配条件，当句法分析得到的依存关系满足其中的依存关系条件时，则可以使用依存关系之间的组合关系得到属性谓词。例如，当依存关系满足“assmod =3”时，得到 3 个依存关系“assmod(Y,X)”，“assmod(Z,Y)”，“assmod(V,Z)”，如果 X 为问句的主体，则“Y”，“Z”，“V”分别是自然语言问句的 3 个属性谓词。

3.3.3 谓词消歧

谓词消歧就是将句子中识别出的谓词和知识库中已经存储的属性谓词对齐。谓词消歧有助于将问句转化为结构化查询语言，方便查询知识库。本文使用的谓词消歧的步骤如下：

1)查询谓词词典，判断是否存在该属性谓词，若存在则跳转到步骤 5)，否则跳转到步骤 2)；

2)使用该属性谓词和谓词词典中的每个词计算语义相似度[6]；

3)如果语义相似度大于 0.5，跳转到步骤 5)，否则跳转到步骤 4)；

4)返回 UNKNOWN；

5)返回对齐后的属性谓词。

需要注意的是，本文中的谓词词典是分领域建立的，如“人物类”、“电影类”、“音乐类”等 6 类领域词典。分领域建立词典的优点在于谓词消歧时能在计算相似度的过程中减少运算量。

通过 3.3.2 小节中的基于句法分析和谓词词典相结合的属性谓词识别方法，得到自然语言问句“刘德华的生日是什么时候”的属性谓词为“生日”，但是该谓词在人物类谓词词典中不存在，人物类谓词词典中的标准谓词为“出生日期”，因此，需要将“生日”消歧为“出生日期”。

例如，对“刘德华的生日”进行谓词消歧的过程如图 3 所示。

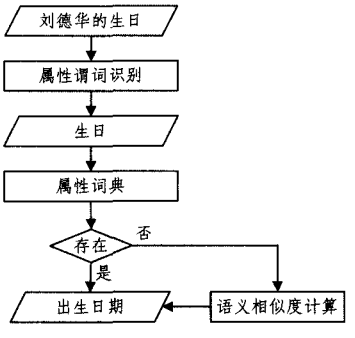

图 3　谓词消歧示例

如果属性谓词在谓词词典中，则返回标准谓词；若不存在，则计算属性谓词与谓词词典中每个词的语义相似度，如果与谓词词典中某个属性谓词的语义相似度最大且相似度值大于 0.5，则返回谓词词典中的该谓词。

4　实验

4.1　数据集

为了测试文中语句的理解程度，需要大量的句子进行测试。因此，在本文的实验中数据主要来自：Web 句子、句子采集系统中的句子、实网手机助手所得的交互语句。Web 句子主要来自百度百科、豆瓣、360 搜索等互联网数据；句子采集系统有 300 余人参与，其中多人参与训练数据的主体标注工作；手机助手收集得到的 145371 条数据是未经过分类的原始语句。因此需要使用统计的方法进行初步分类，然后经过十余人的人工挑选和确认，最终得到数据质量较好的训练数据。最终得到的用于训练和测试的句子如表 6 所列。

表 6　句子收集情况统计表

类别	人物	电影	音乐	图书	游戏	应用
手机助手	1905	629	1638	101	365	739
Web	1427	1235	1128	1036	1317	1189
句子采集系统	2505	3318	2537	2234	2024	1883
总计	5837	5182	5300	3371	3706	3811

4.2　问句分类与主体识别实验

在 CRF 分类和主体识别模型中需要设定两个参数，即阈值(cut-off threshold)和拟合度(hyper-parameter)。阈值和拟合度并不是越高越好。阈值越高，训练数据的特征就越少；阈值越小，无效特性就越多。拟合度越高，训练数据更加耦合；拟合度越低，训练数据更加离散。因此，本文采用交叉验证的方式确定最优参数。为了获得最优组合，需要进行 5 次主体识别实验，每次实验需要阈值和拟合度交叉组合，交叉验证的主体识别准确率如图 4 所示。

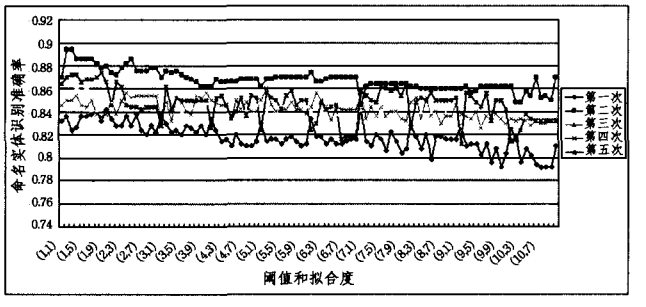

图 4　阈值与拟合度交叉验证图

图 4 中，每条线表示每一次实验中阈值和拟合度组合时的主体识别准确率，共有 5 次实验，故有 5 条折线。横坐标〈X,Y〉中，$X\in[1,5]$，$Y\in[1,10]$，其中 X 表示阈值大小，Y 表示拟合度。从图 4 所示的主体识别实验数据中得出，当阈值为 1 且拟合度为 4 时，5 次主体识别的过程的平均准确率为 86%，综合效果最好。因此，在之后的实验中，CRF 的阈值设置为 1，拟合度设置为 4。

文中分类和主体识别同时进行，有些主体虽然识别错误，但分类却是准确的。因此，问句分类和主体识别需要单独计算 F-Measure 值。另外，将表 6 中的数据分为两部分，一部分作为训练数据，另一部分作为测试数据；其数据集大小比例为：训练数据占 2/3，测试数据占 1/3。计算所得的问句分类的 F-Measure 值如表 7 所列。

表 7　问句分类 F-Mesure 值/%

类别	CRF	Maxent	SVM
人物	92.27	88.09	94.75
电影	93.29	94.14	92.60
音乐	94.37	95.78	94.21
图书	93.30	92.62	95.18
游戏	94.86	97.10	94.47
应用	92.16	93.95	93.17
平均 F-Measure	93.38	93.61	94.06

如表 7 所列，本文使用 3 种分类方法做对比实验：CRF

法，Maxent 和 SVM，这 3 种方法在分类研究中应用更多，取得了不错的效果，但是本文采用 CRF 的主要原因是它在主体识别的同时可以获取问句的类别，避免了 SVM 和 Maxent 算法先分类后主体识别的过程。

本文使用 CRF 模型对句子进行预测时，可以获得句子中的主体，然后通过表示主体的标注集的类型信息获取句子类别，因此，分类和主体识别的效果的差别不应较大，应在较小的范围之内。但是，在进行主体识别时，主体的首部和尾部有时会多一个词或者少一个词，从而导致命名识别错误，但这个错误并不影响分类效果。因此，从理论上看，分类的 F-Measure要高于主体识别的 F-Measure。具体的主体识别效果如表 8 所列。

表 8 主体识别 F-Measure 值/%

迭代实验	人物	电影	音乐	图书	游戏	应用	平均 F-Measure 值
实验一	92.24	89.74	93.01	92.24	91.27	93.74	92.04
实验二	87.57	88.91	88.97	89.60	91.25	93.54	90.31
实验三	92.40	90.69	92.56	95.19	94.84	95.27	93.04
实验四	94.94	92.56	93.03	94.15	94.05	95.01	94.12
实验五	91.12	91.24	91.59	92.73	94.31	95.27	92.71
平均 F-Measure 值	91.65	90.63	91.83	92.78	93.14	94.57	92.44

如表 8 所列，为了检验主体识别的效果，本文采用 5 组训练数据和测试数据，训练数据和测试数据的数据集大小与问题分类的大小一样。不同之处在于，在分类中使用 1 组实验数据、3 种不同的分类算法；在主体识别实验中，使用 5 组数据及同一主体识别算法。如表 8 所列，5 组实验的平均 F-Measure值都在 90%以上，因此本模型能够满足主体识别的要求。

4.3 谓词识别实验

本实验中，从人物类、电影类、音乐类、图书类、游戏类、应用类各取 100 个符合表 6 中句式结构的句子来测算句法分析的准确率，即总共迭代 5 次，每次取 100 个句子做测试，计算出每次的准确率。采用 100 个问句是因为本文收集的问句采用领域词典和人工的方式来标注主体，但是没有标注问句中的谓词和谓词消歧后的结果，例如，“《平凡的世界》是谁写的”需要进行谓词消歧，得到谓词“作者”。因此，本节实验需要采用人工方式观察问句的谓词是否识别准确以及是否准确地进行了消歧等，具体结果如表 9 所列。

在表 9 中，因为游戏类和应用类没有间接属性问答，所以准确率较高。如表 9 所列，每一次实验中平均准确率为 70%左右，因此仅单纯使用句法分析来获取谓词的效果不佳，应该将它和其他方法结合才能使准确率和效率得到显著的提升。本文使用了谓词词典和句法分析相结合的方式识别句子中的谓词。

表 9 句法分析准确率结果/%

迭代实验	人物	电影	音乐	图书	游戏	应用	平均准确率
实验一	71.00	60.19	61.56	79.41	83.55	77.29	72.17
实验二	66.25	61.99	76.92	65.57	74.28	81.66	71.11
实验三	75.27	72.58	69.37	74.46	83.94	75.37	75.17
实验四	79.38	54.51	67.29	65.33	80.74	69.88	69.52
实验五	62.63	71.69	69.11	66.22	76.86	78.44	70.83
平均准确率	70.91	64.19	68.85	70.20	79.87	76.53	71.96

表 10 是将谓词词典和句法分析结合后在封闭测试环境下的谓语识别准确率结果，实验数据使用句法分析所用的数据，以便对比在同一数据集上不同方法的准确率。

如表 10 所列，将谓词词典和句法分析结合后能够显著提升谓词识别的准确率，平均提升 20%左右，因此本文所提出的谓词识别方法能够满足应用要求。需要注意的是，词典和句法分析相结合后，词典识别谓词在前，句法分析在后；另外，当使用词典识别谓词时，需要考虑谓词出现的先后顺序。

表 10 谓词词典和句法分析结合后的谓词识别效果/%

迭代实验	人物	电影	音乐	图书	游戏	应用	平均准确率
实验一	90.90	92.80	88.17	94.09	85.31	91.57	90.47
实验二	92.14	94.05	89.57	93.23	92.91	94.28	92.70
实验三	83.99	93.18	87.71	92.04	94.20	92.76	90.65
实验四	88.57	87.95	89.29	93.51	89.99	91.17	90.08
实验五	92.61	90.79	90.14	92.73	89.95	91.27	91.25
平均准确率	89.64	91.75	88.98	93.12	90.47	92.21	91.03

4.4 问句理解实验

以上所有实验对本文所涉及到的问句分类以及问句中的主体识别、谓词识别、属性消歧等环节做了评估。问句理解不单纯是主体的识别或是谓词的识别及消歧，它需要将问句中的主体和谓词识别出来，同时谓词需要与知识库中的谓词对齐，这样才能称为完整地理解了问句的语义。因此，问句理解的准确率为问句的主体和谓词都识别正确且谓词消歧也准确的问句的数量除以问句的总数，即在问句集合中准确的问题元的个数所占的比例。

使用谓词识别所用过的 5 组实验数据进行实验，即从人物类、电影类、音乐类、图书类、游戏类、应用类中各取 100 个句子测算问句理解的准确率。总共迭代 5 次，每次取 100 个

句子做测试。主体识别和谓词识别都准确且谓词消歧也准确时认为该问句被准确地理解，否则被理解错误。采用人工观察问题元是否准确的方式统计问句理解的准确率。本实验为封闭实验，实验结果如表 11 所列。

表 11　问句理解的准确率/%

迭代实验	人物	电影	音乐	图书	游戏	应用	平均准确率
实验一	81.60	83.19	85.63	90.25	80.57	87.64	84.81
实验二	84.19	79.27	75.39	81.47	86.94	84.68	81.99
实验三	77.67	73.92	78.34	87.15	79.15	83.55	79.96
实验四	81.91	86.17	75.28	81.63	77.93	80.61	80.59
实验五	78.57	79.83	78.64	84.77	82.81	84.69	81.55
平均准确率	80.79	80.48	78.66	85.05	81.48	84.23	81.78

如表 11 所列，在实验中选择的 100 条人物类问句的结构比较复杂，有些问句一次间接属性或者二次间接属性类问句；另外，部分问句的属性谓词使用了同义词代替。因此，人物类问句的理解准确率较低。音乐类问句的准确率为 78.66%，主要是因为音乐类问句含有一些播放某歌曲或者下载某歌曲的问句，而且歌曲更新较快，导致使用 CRF 算法识别主体时效果较差。图书类问句较为规整，因此问句理解的准确率是最高的。电影类、游戏类、应用类由于主体更新较慢，且训练使用的问句中包含了这些最新类型的问句，因此主体识别效果较好，能获得较好的问句理解程度。

结束语　本文首先介绍了问句的类别定义和问句理解的流程；然后介绍了如何使用 CRF 算法同时进行问句分类和主体识别工作；最后介绍了谓词词典和句法分析相结合的谓词识别和谓词消歧方法。将谓词词典和句法分析相结合，可以解决问句句式变化但谓词固定和问句谓词变为同义词但句式固定这两种情况。通过以上工作，基本能够理解人物类、电影类、音乐类、图书类、游戏类、应用类的问句。另外，本文提出的方法也取得了不错的实验结果，问题分类的平均 F-Measure值为 93.88%，主体识别的平均 F-Measure 值为92.44%，谓词识别的平均准确率为 91.03%，问句理解的准确率为 81.78%。

参 考 文 献

[1] ZHOU Z, ZHANG L J, HE X F, et al. Expert Finding for Question Answering via Graph Regularized Matrix Completion[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(4): 993-1004.

[2] LIU D R, CHEN Y H, SHEN M X, et al. Complementary QA Network Analysis for QA Retrieval in Social Question Answering Websites[J]. Journal of the Association for Information Science & Technology, 2015, 66(1): 99-116.

[3] WU C H, LIU C H, SU P H. Sentence Extraction with Topic Modeling for Question-Answer Pair generation[J]. Soft Computing, 2015, 19(1): 39-46.

[4] KONKOL M, BRYCHÍN T, KONOPÍK M. Latent semantics in Named Entity Recognition[J]. Expert Systems with Applications, 2015, 42(7): 3470-3479.

[5] LIU Q L, TOMASZ J, JINHO C. Real-Time Community Question Answering: Exploring Content Recommendation and User Notification Strategies[C]// Proceedings of the 20th International Conference on Intelligent User Interfaces. Atlanta: ACM, 2015: 50-61.

[6] WEI C Y, ZHAN Q, FAN X Z, et al. Event Information Enhanced Question Semantic Representation for Chinese Question Answering System[J]. Journal of Chinese Information Processing, 2015, 1(29): 147-154. (in Chinese)
魏楚元，湛强，樊孝忠，等. 融合事件信息的中文问答系统问题语义表征[J]. 中文信息学报，2015，1(29)：147-154.

[7] HUANG P J, HUANG Q, WU X P, et al. Question Understanding by Combining Grammar and Semantic for Chinese Dialogue System[J]. Journal of Chinese Information Processing, 2014, 28(6): 71-78. (in Chinese)
黄沛杰，黄强，吴秀鹏，等. 语法和语义相结合的中文对话系统问题理解研究[J]. 中文信息学报，2014，28(6)：71-78.

[8] LIU C T, LI Z S. Research on Question Analysis Based on Sentence Type Recognition of Interrogative Sentences[J]. Computer Science, 2008, 35(12): 151-153. (in Chinese)
刘朝涛，李祖枢. 基于疑问句句型识别的问题理解研究[J]. 计算机科学，2008，35(12)：151-153.

[9] CHEN Y P, YANG S C, MAO W S, et al. Question Interpretation Based on Theme and Focus in Chinese Question Answering System[J]. Computer Systems & Applications, 2011, 20(6): 56-60. (in Chinese)
陈永平，杨思春，毛万胜，等. 中文问答系统中基于主题和焦点的问题理解[J]. 计算机应用，2011，20(6)：56-60.

[10] MA L, TANG S Q, CHEN L N, et al. Improved Question Interpretation Method Based on Matching Algorithm of Sentence Template[J]. Computer Engineering, 2009, 35(20): 50-52. (in Chinese)
马莉，唐素勤，陈立娜，等. 改进的基于句模匹配算法的问句理解方法[J]. 计算机工程，2009，35(20)：50-52.

[11] HOU Y S, ZHANG Y Y, WANG X L, et al. Recognition and Retrieval of Time-sensitive Question in Chinese QA System[J]. Journal of Computer Research and Development, 2013, 50(12): 2612-2620. (in Chinese)
侯永帅，张耀允，王晓龙，等. 中文问答系统中时间敏感问句的识别和检索[J]. 计算机研究与发展，2013，50(12)：2612-2620.

[12] XU K, FENG Y S, ZHAO D Y, et al. Automatic Understanding of Natural Language Questions for Querying Chinese Knowledge Bases[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2004, 50(1): 85-92. (in Chinese)
许坤，冯岩松，赵东岩，等. 面向知识库的中文自然语言问句的语义理解[J]. 北京大学学报(自然科学版)，2004，50(1)：85-92.

[13] UNGER C, BUHMANN L, LEHMANN J, et al. Templatebased Question Answering over RDF Data [C]// Proceedings of the 21st International Conference on World Wide Web. New York: ACM, 2012: 639-648.

(下转第 254 页)

果表明，该算法能高效地产生频繁项集，解决了 Apriori 算法产生大量候选项集和多次扫描全局事务数据库所产生的时间开销的问题。理论分析和对比实验表明，本文算法是有效可行的，相比传统的关联规则挖掘算法，本方法可以推广到数据规模庞大的移动应用网页推荐中。

参 考 文 献

[1] HUANG Y B,CHEN M Y. Architecture Characteristics and Analysis of Mobile Device Applications[J]. Chinese Journal of Computers,2015,38(2):386-396. (in Chinese)
黄永兵，陈明宇. 移动设备应用程序的体系结构特征分析[J]. 计算机学报，2015，38(2)：386-396.

[2] MENG X W,HU X,WANG L C,et al. Mobile recommender systems and their applications[J]. Journal of Software,2013,24(1):91-108. (in Chinese)
孟祥武，胡勋，王立才，等. 移动推荐系统及其应用[J]. 软件学报，2013，24(1)：91-108.

[3] AGRAWAL R,IMIELIMSKI T,SWAMI A. Mining Association Rules between sets of items in large databases[C]// Proceedings of the ACM SIGMOD Conference on Management of Data. Washington DC,1993:207-216.

[4] AGRAWA A,SRIKANT R. Fast algorithms for mining association rules[C]// Proceedings of the VLDB International Conference. 1994:487-499.

[5] SCHLEGEL B,KIEFER T,KiSSINGER T. pcApriori:Scalable Apriori for Multiprocessor Systems[C]// Proceedings of International Conference on Scientific and Statistical Database Management. 2013:1-12.

[6] GUO J,RENG Y G. Research on association rule mining in Book sales under cloud computing environment[J]. Computer Applications and Software,2014,31(11):50-53. (in Chinese)
郭健，任永功. 云计算环境下的关联规则挖掘在图书销售中的研究[J]. 计算机应用与软件，2014，31(11)：50-53.

[7] LUO D,LI T S. Research on Improved Apriori Algorithm Based on Compressed Matrix[J]. Computer Science,2013,40(12):75-80. (in Chinese)
罗丹，李陶深. 一种基于压缩矩阵的 Apriori 算法改进研究[J]. 计算机科学，2013，40(12)：75-80.

[8] WANG B L,SHEN Y G. Improvement of Apriori algorithm based on boolean matrix[J]. Adwanced Materials Research,2011,159:144-148.

[9] LIN M Y,LEE P Y,HSUEH S C. Apriori-based Frequent Itemset Mining Algorithm on Mapreduce[C]// Proceedings of the 2nd International Conference on Ubiquitous Management and Communication. 2012:1-8.

[10] LAZCORRETA E,BOTELLA F,FERNÁDEZ-CABALLERO A. Towards personalized recommendation by two-step modified Apriori data mining algorithm[J]. Expert Systems with Applications,2008,35(3):1422-1429.

[11] TANG J W,WANG X F. Design and Implementation of Apriori on GPU[J]. Computer Science,2014,41(10):238-243. (in Chinese)
唐家维，王晓峰. 基于 GPU 的并行化 Apriori 算法的设计与实现[J]. 计算机科学，2014，41(10)：238-243.

[12] LIU D Y,FENG J,LI X F. Logic-based Frequent Sequential Pattern Mining Algorithm[J]. Computer Science,2015,42(5):260-264. (in Chinese)
刘端阳，冯建，李晓粉. 一种基于逻辑的频繁序列模式挖掘算法[J]. 计算机科学，2015，42(5)：260-264.

[13] 韩家炜，等. 数据挖掘概念与技术(第 3 版)[M]. 范明，等译. 北京：机械工业出版社，2012：158-162.

[14] OLIVEIRA S R M,ZAIANE O R. A unified framework for protecting sensitive association rules in business collaboration [J]. International Journal of Business Intelligence and Data Mining,2006,1(3):247-287.

[15] JEFFREY D,SANJAY G. Mapreduce:Simplified Data Processing on Large Clusters[J]. Proceedings of the Sixth Symposium on Operating System Design and Implementation,2004,51(1):107-113.

(上接第 221 页)

[14] FERRUCCI D,BROWN E,CHU-CARROLL J,et al. Building Watson:an Overview of The Deep QA Project[J]. AI Magazine,2010,31(3):59-79.

[15] UNGER C,CIMIANO P. Pythia:Compositional Meaning Construction for Ontology-based Question Answering on the Semantic Web[C]// International Conference on Natural Language Processing and Information Systems. Salford:Springer Berlin Heidelberg,2011:153-160.

[16] BERANT J,CHOU A,ROY F,et al. Semantic Parsing on Freebase from Question-Answer Pairs[C]// The 2013 Conference on Empirical Methods on Natural Language Processing. Seattle:Association for Computational Linguistics,2013:1533-1544.

[17] BERANT J,LIANG P. Semantic Parsing via Paraphrasing[C]// The 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore:Association for Computational Linguistics,2014:479-485.

[18] BORDES A,CHOPRA S,WESTON J. Question Answering with Subgraph Embeddings[C]// The 2014 Conference on Empirical Methods on Natural Language Processing. Stroudsburg:Association for Computational Linguistics. 2014:1535-1545.

[19] YAO X,DURME B. Information Extraction over Structured Data:Question Answering with Freebase[C]// The 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore:Association for Computational Linguistics. 2014:753-770.

[20] FADER A,LUKE Z,OREN E. Open Question Answering Over Curated and Extracted Knowledge Bases[C]// Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD). 2014:1156-1165.

[21] BAO J W,DUAN N,ZHOU M,et al. Knowledge-Based Question Answering as Machine Translation[C]// The 52nd Annual Meeting of the Association for Computational Linguistics,Baltimore:Association for Computational Linguistics. 2014:1272-1294.