一种规则与统计相结合的应用题句子语义角色识别方法*

吴林静，劳传媛，范桂林，黄景修，刘清堂†
（华中师范大学 教育信息技术学院，武汉 430079）

摘 要：以应用题自动求解为目标，以高考入学考试数学试卷中的分层抽样应用题为研究对象，重点研究了分层抽样应用题的句子语义角色识别方法。根据分层抽样的原理，首先定义了分层抽样题意表征中的五种核心语义角色，分别为总体、样本、总体中的层、样本中的层和实体之间的关系。基于这五种语义角色，应用题题意理解中的核心问题被转换为对应用题文本中的句子进行语义角色判定。提出了一种基于特征词与 *n*-gram 模型相结合的句子语义角色判定方法，对分层抽样应用题文本中的句子进行语义角色判定。根据测试集中的实验结果，应用题的整题识别准确率由基于特征词的判定方法的 17.95% 提高到 64.1%。实验结果说明基于特征词与 *n*-gram 模型相结合的句子语义角色判定方法能够提高题意理解的准确率。

关键词：应用题自动求解；题意理解；语义角色；特征词；*n*-gram
中图分类号：TP391.1 **文献标志码**：A **文章编号**：1001-3695(2018)08-2299-05
doi:10.3969/j.issn.1001-3695.2018.08.015

Hybrid method based on rules and statistics for semantic role annotation in math word problems

Wu Linjing, Lao Chuanyuan, Fan Guilin, Huang Jingxiu, Liu Qingtang†
(*School of Educational Information Technology*, *Central China Normal University*, *Wuhan* 430079, *China*)

Abstract: This paper proposed a semantic role annotation method of sentences for the stratification sampling word problem in China's college entrance examination. According to the basic principles of stratification sampling, this paper defined five core semantic roles to represent the meaning of stratification sampling word problem. These five roles were population, levels in the population, sample, levels in the sample and the relation between entities. With the help of these five semantic roles, it could solve the word problem if the sentences in the word problem could be mapped with these semantic roles. To achieve this goal, this paper presented a hybrid method based on characteristic words and *n*-gram model to annotate the semantic roles of sentences in stratification sampling word problem. The experimental results show that the accuracy of the right-annotated word problems improves from 17.95% to 64.1%. It proves that the hybrid method can significantly improve the accuracy of stratification sampling word problem understanding.

Key words: word problem automatically solving; understanding of word problem; semantic roles; characteristic words; *n*-gram

0 引言

数学问题自动求解一直是人工智能领域一项富有挑战性和吸引力的工作[1,2]。许多研究者在该领域奋斗多年并取得了一系列的研究成果。例如由吴文俊[3]所提出的方法可以对初等数学中的几何定理进行机器证明，被公认为是机器证明领域的里程碑。张景中等人[4]提出了平面几何定理的可读机器证明的方法，并基于该方法研发了智能教学软件超级画板，是机器自动求解领域的一座丰碑，同时为基础教育事业作出了巨大贡献。除了几何定理的证明之外，近年来应用题的自动求解也开始受到广泛关注[5]。其中初等数学代数问题尤其是加减法的应用题求解成为研究者们最先选择的应用领域[6,7]。选择这一应用领域的主要原因在于初等数学中的加减法类应用题题意相对清晰、涉及到的参数较少、求解规则相对简单。因此，研究者们大多选择这一领域作为研究起点。

本文选择了另外一类应用题作为研究内容：国家高考入学考试中的分层抽样类应用题。该研究是国家“863”计划项目“初等数学问题求解关键技术及系统”的研究任务之一。项目的研究目标是构建一个机器求解程序，对中国高考入学考试的数学试卷进行自动求解。分层抽样类应用题是目前高考数学试卷中的必考内容之一。图 1 即为一个典型的分层抽样类应用题。

分层抽样应用题
某中学高一年级有学生 1200 人，高二年级有学生 900 人，高三年级有学生 1500 人，现用分层抽样的方法从中抽取一个容量为 720 的样本进行某项调查，则高二年级应抽取的学生数为(　)？
关键参数： 总体：高一年级 1200 人、高二年级 900 人、高三年级 1500 人 样本：720 求解：样本中高二年级的人数
方程：$900/x=(1200+900+1500)/720$ 答案：$x=180$

图 1 分层抽样应用题的实例

当学习者对上述问题进行求解时，解题过程主要包括三个步骤：a)理解题意并抽取出题目中的参数；b)确定需要使用的数学规则，并将参数代入规则；c)通过计算或推理，得出计算结果。对于学习者来说，步骤 a)相对容易；步骤 b)c)则具有一定的难度。但是对于机器自动求解来说，情况则刚好相反。机器进行题意理解的难度主要体现在以下几个方面：

a)分层抽样涉及到的参数较多，比小学加减法的参数要复杂。

b)应用题本身具有一定的情境性，随着情境的变化，实体及参数也会发生变化。例如在图 1 中，“学生”可以使用“教师”“职工”“运动员”进行代替。同时，随着情境的不同，同一

收稿日期：2017-04-11；**修回日期**：2017-05-19 **基金项目**：国家“863”计划资助项目(2015AA015408)；国家教育部新世纪优秀人才计划项目(NCET-13-0818)；国家“十二五”科技支撑计划资助项目(2015BAK27B02)；中央高校基本科研业务费项目(CCNU15A02020，CCNU16A05023)

作者简介：吴林静(1987-)，女，湖北松滋人，讲师，博士，主要研究方向为语义网络、人工智能及应用；刘清堂(1969-)，男(通信作者)，教授，博士，主要研究方向为自然语言理解、人工智能与教育应用(liuqtang@mail.ccnu.edu.cn)。

实体在分层抽样中所充当的角色也会发生变化。如在图1的实例中,“高二年级”学生是总体中的一个层。而在下例中,“高二年级”学生则成为分层抽样的总体:“高二年级有学生900人,其中男生400名,女生500名。”

c)中文中存在较多的省略与上下文情境依赖,增加了题意理解的难度。如前例中的最后一句“女生500名”,若仅观察该子句,几乎无法判断该参数在分层抽样中到底充当了什么样的角色。

本文的研究目标是从分层抽样的题目文本中获取相关的核心参数,并标志其在分层抽样中的语义角色。为了达到这一目标,本文首先提出了一个分层抽样的题意表征框架,识别出分层抽样类问题中的核心语义角色和求解规则。该框架定义了分层抽样问题中的五种核心语义角色,分别为总体、总体中的层、样本、样本中的层、实体关系。分层抽样中的所有参数被映射为上述五种角色中的一种。依据这一框架,本文定义了一些特征词和模式,用于判定分层抽样题目中句子的角色。但是,由于中文表达方式的多样性和省略、指代等语法现象,并非所有的句子都可以通过特征词和模式来进行识别。不能通过特征词识别的句子,其语义往往与上下文紧密联系。根据这一特征,本文提出了一种基于*n*-gram模型的句子语义角色预测方法,对无法通过特征词判定的句子进行类型预测,从而提高题意理解的准确率。

为了验证方法的有效性,本文构建了一个包含189道分层抽样应用题的语料库,全部来源于近十年的高考真题和教材的课后习题。其中150道作为训练集,39道作为测试集。测试结果显示,基于特征词的判定方法对于单句语义角色判定的*F*值为75.33%,对于整题的判定准确率为17.95%。结合*n*-gram模型进行校正之后,单句角色判定的*F*值提高到了92.1%,整题的判定准确率提高到了64.1%,这说明本文所提出的方法在限定语料中是有效的。

1 相关研究

1.1 应用题的题意理解

题意理解的研究由来已久,最早研究数学问题求解的系统可以追溯到1964年由Bobrow[8]开发的STUDENT系统,该系统可以求解使用英语描述的代数问题。系统被设计为可以处理六种数学关系,分别是和、不同、乘积、商、幂和相等。其题意理解的过程是围绕上述六种关系建立关系模型,并使用谓词逻辑进行描述。1969年出现的CARPS系统可以求解距离与体积的微积分问题,其题意理解的方法是将题意转换为树型结构以表示对象之间的顺序关系[9]。这一时期的主要研究方法是使用能够表征数学关系的关键词。20世纪80年代,随着认知心理学的发展,人们开始从语义层面对题意理解进行研究。例如Riley[10]将应用题分为3大类14小类。Kintsch等人[11]提出了小学应用题的问题框架表征模型。Dellarosa[12]开发的ARITHPRO系统则可以通过模拟人类认知过程进行一步加减算术应用题的自动求解。该系统将题意分为三类,即关于词的知识、关于命题的知识和关于文本结构的知识。到了20世纪90年代末,统计语言模型开始被用于应用题的题意理解中。例如Wong等人[13]构建了Info-map本体知识库帮助理解一步加减数学应用题,其方法是将题目中的词汇与本体库中的词汇进行匹配,以确定问题属于不同类别的概率。国内较有代表性的研究包括程志[14]开发的小学数学应用题求解系统,马玉慧[15]开发的小学数学一学段应用题自动求解系统,以及Yu等人[16]关于小学数学应用题自动求解的研究。程志开发的小学数学应用题求解系统主要通过关键词串来实现题意理解。马玉慧开发的小学数学一学段应用题自动求解系统则以Kintsch提出的问题框架表征模型为基础,通过句模的方式来实现题意的理解。Yu等人则通过关键词和语义关系映射的方式对题意进行表征。关键词和句模成为自然语言题意理解中的首选方法。

1.2 数学问题的自动求解

数学问题的自动求解最早起源于几何领域。1977年,吴文俊[3]提出了一个证明等式型初等几何定理的新的代数方法,被公认为是机器证明领域的里程碑式突破。1992年,张景中等人[4]提出了消点算法,在几何定理可读机器证明方面取得了突破性进展。在模拟人类的解题过程方面,目前研究较多的是认知推理模型[17~19]。认知推理模型被认为是人工智能的基础之一,其主要作用是对人类认知过程进行计算机推理[20]。如文献[21]利用认知模型对代数方程式求解进行模型化。美国Carnegie Mellon大学的Anderson从认知心理学的角度提出了ACT-R认知模型,被公认为是模拟人类高级认知过程的理想模型之一。基于该模型开发的数学学习软件Carnegie Learning可以进行简单的数学问题求解[22]。国内符红光等人提出了三角函数化简过程中的认知推理模型,提出了三角函数分层化简算法,实现了三角函数的计算机自动化简。在代数应用题自动求解方面,尤其是在小学数学应用题中,较为常用的自动求解方法主要是通过建立方程或方程组来实现问题的自动求解。例如文献[6,7]均使用该方法来实现问题的自动求解。

本文的研究对象为分层抽样类应用题,这使得目前已有的方法中有许多方法可以为本文的研究提供借鉴,但并不能完全适合于本研究。因此,针对分层抽样应用题的基本特征,本文提出了一种基于核心语义角色的题意表征模型,并以该模型为基础,通过句子的语义角色分类来实现应用题的题意理解。

2 句子语义角色标注方法

2.1 分层抽样应用题的题意表征模型

为了对分层抽样应用题的题意进行表征,本文根据分层抽样问题求解过程中涉及到的各类参数定义了五种核心语义角色,分别为总体、总体中的层、样本、样本中的层、实体关系。这五种核心语义角色与问题求解密切相关,是应用题题意理解的关键。表1为五类核心角色的定义与实例。

表1 分层抽样的核心语义角色及其实例

语义角色	定义	实例
总体	对分层抽样的总体进行描述	某校共有教师156人
总体中的层	对总体的分层情况进行描述	其中男性90人 女性66人
实体关系	对不同层之间的数量关系进行描述	男女比例为2∶3 男性比女性少10人
样本	对抽取的样本进行描述	现通过分层抽样的方法抽取一个人数为52人的样本
样本中的层	对样本的分层情况进行描述	需要抽取女性多少人

从实例中可以看出,如果能够识别出应用题文本中一个句子在分层抽样问题求解中所代表的语义角色,那么就可以通过实体抽取的方法将该句子中的数值信息和离该数值最近的实体及其数量单位抽取出来,作为分层抽样问题求解的关键参数。例如表1中的总体句“某校共有教师156人”,可提供如下信息:分层抽样的总体为“教师”,数量为“156”,单位为“人”。其他句类似。

完成所有的关键参数提取之后,即可通过分层抽样的相关计算规则进行计算。分层抽样问题求解的计算规则包括:

规则1 总体的数量=总体中的层的数量之和。

规则2 抽样比=样本的数量/总体的数量。

规则3 抽样比=样本中某一层的数量/总体中对应层的数量。

将抽取的参数代入上述规则,即可得到一系列的方程组。对于方程组的求解已由相关数学引擎如maple、mathematic等完成,从而使得应用题得到解决。

因此,从上述分析可以看出,分层抽样应用题题意理解的核心在于对句子的语义角色进行识别。如果句子的语义角色能够被正确地识别,就几乎获得了求解所需要的全部参数。因此,本文的研究目的即为对分层抽样应用题文本中句子的语义角色进行识别。

为了实现目标,本文算法主要包括两个步骤:a)通过特征词和部分模式对句子的语义角色进行初略判定(后文中简称

为基于特征词和模式的方法);b)在初略判定结果的基础之上,利用 *n*-gram 模型对无法通过基于特征词的方法判定的句子进行预测,从而提高语义角色识别的准确率。

2.2　基于特征词和模式的语义角色识别

在中文中很多词汇具有明确的意义指向性。因此可以利用这些词汇来理解句子的意义,并识别句子在分层抽样中所承担的语义角色。例如"总共"常用来指代总体;"容量"则用来指代样本的容量等。表 2 给出了从实例中所筛选出的各类不同语义角色的代表性词汇和模式。

表 2　不同角色的特征词与模式

语义角色	代表性特征词与模式	实例
总体	共有,一共,总共,共,全体,全部,全,总体,总数,总人数,总	某单位教职工共有 160 人
总体中的层	依次 + … + ! 抽取, 其中 + … + ! 抽取, 分别 + … + ! 抽取	其中女性人员有 24 人
实体关系	占,比,之比,	男女职工之比为 5∶2
样本	分层抽样,样本,容量,大小,抽取,抽查,抽,选出,选取	从该单位职工中抽一个容量为 5 的样本
样本中的层	{分别} + …. + {抽[取]},{依次} + …. + {抽[取]},在/从 + … + 抽[取],其中 + … + 抽[取],样本中	请问应从女性职工抽取多少人

在表 2 中,"!"表示其后的词汇不出现,如模式"依次 + … + ! 抽取"表示"依次"后面没有出现"抽取";"{}"表示词汇在句子中出现的顺序不受限制,如"{分别} + … + {抽取}"表示"分别……抽取"和"抽取……分别"均可与该模式进行匹配;"[]"表示其中的词汇可以出现,也可以不出现。

在实际的应用题中,一个句子可能包含上述不同角色中的多个词汇,因此,根据句子中所包含的不同角色中的特征词数量,本文制定了如下的优化判定规则:

规则 4　如果一个句子仅包含一种语义角色的特征词,则将该语义角色作为当前句子的语义角色。

规则 5　如果一个句子包含两种及两种以上角色的特征词,选择包含特征词数量最多的角色作为句子的角色标注结果。

规则 6　若多个句子包含"总体"角色的特征词和模式,则选择数值信息较大的句子标注为"总体",同时将其他句子的"总体"角色特征词和模式数量置为 0(一个题目仅包含一个总体,且总体的数量应最大)。

规则 7　若一个子句同时包含"样本"角色特征词和"样本中的层"角色特征词,且两种角色的特征词数量相等,则该句优先被判定为"样本中的层",同时,将该句的样本特征词数量置为 0。

规则 8　若一个子句同时包含"实体关系"角色特征词和其他角色特征词,且数量相等,则该句被优先判定为"实体关系"。

表 3 展示了部分例句的判定过程及结果。

表 3　基于特征词和模式的句子类型判定实例

子句	特征词及模式数量	判定结果	应用规则
某校高一、高二、高三共有学生 1 200 人,	总体:1	总体	规则 1
其中高二有学生 360 人	总体中的层:1	总体中的层	规则 1
现采用分层抽样的方法从三个年级中抽取一个容量为 50 人的样本进行学习兴趣相关情况的调查	样本:4(分层抽样、容量、样本、抽取) 样本中的层:1(从…抽取)	样本	规则 2
则应在高二年级抽取____人	样本:1(抽取)样本中的层:1(在…抽取)	样本中的层	规则 4
其中男女之比为 3∶2	总体中的层:1(其中)实体关系:1(之比)	实体关系	规则 5

在表 3 的实例中,每一个句子的语义角色均被正确地识别了出来。但是在某些情况下,由于句子本身并不包含指向明确的特征词,所以无法根据特征词对其角色进行判定,如"高二年级有 200 人"。因此本文提出了一种基于上下文依赖的 *n*-gram 方法对无法通过特征词判定的句子进行角色预测。

2.3　基于 *n*-gram 模型的句子角色预测方法

n-gram 是一种常用的语言模型,在语音识别中有着广泛的应用,其基本假设是第 n 个词的出现只与前面 $n-1$ 个词相关。在本文中分层抽样应用题由一系列的短句组成。若将组成题目的所有短句看做一个序列,每一个短句作为一个基本元素,则后一个句子的语义角色与其前面句子的角色必然相关。在已经通过特征词和模式确定部分句子角色的情况下,可以利用 *n*-gram 模型对其他句子的语义角色进行预测。其基本算法分为训练与预测两个阶段。

在训练阶段,首先标注出训练集每一个题目中所有包含数值信息的句子的语义角色,形成一个角色序列,作为训练集中的一个训练样本。依次计算每种角色组合在当前训练集中的概率。在本应用中,本文取 n 为 3,进行了三元模型的训练。具体计算方法如下:

一元模型的训练为

$$P(r)=\frac{\text{number}(r)}{N}$$

其中:number(r)表示角色 r 在训练集中出现的次数;N 表示训练集中所有角色的总次数。

二元模型的训练为

$$P(r|s)=\frac{\text{number}(rs)}{\text{number}(s)}$$

其中:$P(r|s)$表示 s 确定时 r 的条件概率;number(rs)表示训练集中 rs 两种角色连续出现的次数;number(s)表示角色 s 在训练集中出现的次数。

三元模型的训练为

$$P(q|rs)=\frac{\text{number}(qrs)}{\text{number}(rs)}$$

其中:$P(q|rs)$表示 rs 确定时 q 的条件概率;number(rs)表示训练集中 rs 两种角色连续出现的次数;number(qrs)表示角色 qrs 在训练集中连续出现的次数。

完成模型训练后,即可利用该模型进行句子角色的预测。具体的预测方法为:

a)首先通过基于特征词和模式的方法对题目文本中的所有包含数值信息的句子的角色进行判定。若所有句子均可进行判定,则预测结束;若有部分句子不能通过该方法进行判定,则将不能判定的句子的语义角色用"*"代替,形成初始判定结果序列。

b)将每一个"*"依次用五种不同语义角色进行替换,形成题目角色判定的所有可能结果集合。集合中包含的可能结果的个数为 5^n,其中 n 为"*"的个数。

c)依次计算可能结果集合中每种可能序列的概率,任一序列 $S(S_1,S_2,\cdots,S_n)$的概率通过如下公式进行计算,其中 $1\leqslant i\leqslant n$:

$$P(S)=P(S_1)\times P(S_2|S_1)\times P(S_3|S_1S_2)\times\cdots\times P(S_i|S_{i-2}S_{i-1})\times\cdots\times P(S_n|S_{n-2}S_{n-1})$$

d)将所有序列按照概率大小进行降序排列,选择概率最大的序列作为当前题目的句子角色预测结果。

应用题句子的语义角色识别完成。

3　实验及结果

3.1　语料与实验结果评价方法

为了验证本文所提出方法的效果,本文采集了 2008—2016 年间各省及全国高考数学试卷和教科书课后习题中关于分层抽样知识点的应用题作为实验语料。由于算法的限制,在试题筛选时,仅从选择题和填空题中进行选择。解答题中的题目大多为复合知识点题目,分层抽样往往与其他知识点进行结合,因此本文不对此类题目进行测试。此外,包含图片和表格的题目也未被包含在本语料中。经过筛选后,本文一共得到了 189 道分层抽样应用题,其中随机选择 150 道作为训练集,39 道作为测试集。

为了对测试结果进行评价,本文借鉴信息分类中的相关评价方法制定了如下的结果评价指标:句子角色识别准确率 P、

句子角色识别召回率 R、句子角色识别的 F 值和整题识别准确率 PQ。句子角色识别准确率的计算方法如下：

$$P(S)=\frac{N(s_{\text{right}})}{N(ST)}$$

其中：S 表示某一特定角色；$P(S)$ 表示对于角色 S 的标注准确率；$N(s_{\text{right}})$ 表示被正确识别出的属于类别 S 的句子数量；$N(ST)$ 表示所有被测试算法标注为类别 S 的句子数量。

句子角色识别召回率的计算方法如下：

$$R(S)=\frac{N(s_{\text{right}})}{N(SM)}$$

其中：S 表示某一特定角色；$R(S)$ 表示对于角色 S 的标注召回率；$N(s_{\text{right}})$ 表示被正确识别出的属于类别 S 的句子数量；$N(SM)$ 表示测试集中所有人工判定为类别 S 的句子数量。

句子角色识别 F 值的计算方法如下：

$$F=\frac{2\times P\times R}{P+R}$$

其中：P 和 R 分别表示当前角色识别的准确率和召回率。

整题标注准确率 PQ 的计算方法如下：

$$PQ=\frac{N(q_{\text{right}})}{N(Q)}$$

其中：PQ 表示以应用题为单位的角色识别准确率，只有当一个应用题中的所有句子均被正确标注时，本文才认为该应用题被正确标注，若某一应用题中的任一句子角色被错误识别，则该应用题被判定为识别错误；$N(q_{\text{right}})$ 表示测试集中被正确识别的应用题数量；$N(Q)$ 表示测试集中所有应用题的总数。

3.2 实验结果

根据文献分析，基于特征词和模板的方法为目前应用题理解中较为常用的方法。因此，本文将基于特征词和模式的角色识别结果作为基线方法，基于特征词与 n-gram 相结合的方法作为对比算法，对实验结果进行说明。测试集的基本信息如表 4 所示。

表 4 测试集的基本信息

题目数量	句子数量	包含数值信息的句子数量	总体句数量	样本句数量	总体中的层句子数量	样本中的层句子数量	实体之间的关系
39	197	184	20	40	76	40	8

对测试集的测试结果进行统计，其中基于特征词的角色标注结果如表 5 所示。

表 5 基于特征词的角色标注结果 /%

类别	总体句	样本句	总体中的层	样本中的层	实体之间的关系	全体句子	整题识别准确率 PQ
准确率 P	94.12	95.00	100.00	100.00	100.00	97.41	
召回率 R	80.00	95.00	22.37	85.00	100.00	61.41	17.95
F 值	86.49	95.00	36.56	91.89	100.00	75.33	

特征词与 n-gram 模型相结合的角色标注结果如表 6 所示。

表 6 特征词与 n-gram 模型结合的角色标注结果 /%

类别	总体句	样本句	总体中的层	样本中的层	实体之间的关系	全体句子	整题识别准确率 PQ
准确率 P	64.29	95.24	98.51	97.37	100.00	92.35	
召回率 R	90.00	100.00	86.84	92.50	100.00	91.85	64.1
F 值	75.00	97.56	92.31	94.87	100.00	92.10	

从实验结果可以看出，在测试集中，通过基于特征词和模式的方法即可对实体之间的关系进行准确识别，准确率和召回率均可达到 100%。其原因在于高考题的语法和用词较为规范，而实体之间的关系描述一般都包含意义明确的特征词，所以此类句子比较容易识别。总体句、样本句和样本中的层也可以达到较高的识别准确率，F 值均在 85% 以上。因为在大部分情况下，这三种类型的句子都包含相关的特征词或模式。而对于总体中的层的识别率则比较低，其召回率仅为 22.37%。经过分析发现，从单句结构上看，此类句子经常使用“名词 + 数词 + 量词”的结构，如“高一年级 500 人”“老年职工 120 人”“黑球 50 个”等。此类句式往往不包含明显的特征词，因此难以直接通过基于特征词和模式的方法进行判断。基于特征词和模式的方法的整题识别准确率仅为 17.95%。

在通过 n-gram 模型对基于特征词的判定结果进行校正之后，除了对总体句的识别效果有所下降之外，其他类型的句子的识别率均得到了提高。对总体句的识别 F 值由 86.49% 下降到了 75%，召回率则从 80% 上升到了 90%。其他类别的句子识别率的 F 值则均达到了 90% 以上，尤其是对于总体中的层的识别率得到了大大提高，召回率从 22.37% 提高到了 86.84%。其原因在于，在分层抽样的题目文本中，对总体中的层的描述往往会跟在对总体的描述之后。若总体句通过基于特征词的方法被识别之后，对总体中的层的句子有很大的概率会跟在总体句之后，因此，通过基于 n-gram 的概率模型可以极大地提高此类句子的识别率。总体中的层的识别率的提高也大大地提高了整题的识别准确率，使其由 17.95% 提高到了 64.1%。经过分析发现，准确率提高的主要原因在于，高考题用词规范，应用题文本中的短句排列顺序会存在一定的规律和相似性，为 n-gram 统计模型的应用提供了基础，从而大大提高了识别准确率。

3.3 错误分析

对于基于特征词的角色标注方法，出现识别错误或无法识别的主要原因在于句子的描述方法发生了一定的变化或者句子中不包含指向明确的特征词，所以规则无法覆盖。如以下例子：

例 1 一批灯泡 400 只（子句 1），其中 20 W、40 W、60 W 的数目之比为 4∶3∶1（子句 2），现用分层抽样的方法产生一个容量为 40 的样本（子句 3），三种灯泡依次抽取的个数为_____（子句 4）。

该例中，子句 2、3、4 均包含明显的特征词，可以明确判定其句子角色，而子句 1 由于不包含明显的总体特征词，所以通过基于特征词的方法未能将其识别。通过 n-gram 预测，子句 1 作为总体出现的概率远大于其他角色的概率，因此 n-gram 模型将其正确判定为总体句。

例 2 某鱼贩一次贩运草鱼、青鱼、鲢鱼、鲤鱼及鲫鱼分别为 80 条、20 条、40 条、40 条、20 条（子句 1），现从中抽取一个容量为 20 的样本进行质量检测（子句 2），若采用分层抽样的方法抽取样本，则抽取的青鱼与鲤鱼共有_____条（子句 3）。

在该例中，子句 1 和 2 均可以通过基于特征词的方法进行正确判断，而子句 3 则由于包含了样本特征词“抽取”和总体特征词“共有”两类特征词，所以无法通过特征词的方式进行判定。此句中的特征词“共有”其实并非描述抽样中的总体数量，而是用于描述一种“求和”的操作。n-gram 可以正确将此句判定为样本中的层，但进行题意理解时，仍然需要对此句进行进一步的处理，即将“共有”转换为“求和”的数学关系。

对于 n-gram 模型预测方法，其结果依赖于语料库中的统计信息，因此，语料的分布会影响预测的结果。基于 n-gram 的统计模型的错误结果主要集中于以下情况：将题目中的第一句预测为总体句，但是实际上该句为对总体中的层进行描述。如以下例子：

例 3 某企业有初级职称 90 人（子句 1），中级职称 45 人（子句 2），高级职称 15 人（子句 3），现抽取 30 人进行分层抽样调查，则各职称被抽取的人数分别为（ ）。

子句 1 ~ 3 均无法通过特征词进行判定，在使用 n-gram 进行预测时，将子句 1 判定为总体，子句 2 和 3 判定为总体中的层。出现该判定结果的原因在于，训练集中有大量的题目第一句均为对总体数量的描述，这也符合分层抽样出题的一般规律，但这导致了本题中子句 1 预测错误。

4 结束语

本文对高考入学考试数学试卷中的分层抽样类应用题的题意理解问题进行了研究。为了实现应用题的计算机自动求解，需要通过题意理解从应用题文本中获取出相关参数，并与问题的求解规则进行对应。本文根据分层抽样的原理及求解方法，定义了一个包含五个核心语义角色的分层抽样题意表征框架，分别为总体、样本、总体中的层、样本中的层和实体之间的关系。以该框架为基础，分层抽样的题意理解问题可以转换

为文本中句子的分类问题。通过对句子的语义角色进行识别,将每一个包含数值信息的分句与五种核心语义角色中的一个角色相对应,最终实现问题的求解。为了实现上述目标,本文提出了一种基于特征词与 *n*-gram 模型相结合的方法对应用题文本中的句子进行角色识别。为了验证方法的有效性,本文采集了 189 道分层抽样应用题,其中 150 道作为训练集,39 道作为测试集。实验结果表明,与仅通过特征词和模式进行判断的方法相比,基于特征词与 *n*-gram 模型相结合的方法有效地提高了句子角色判定的准确率,尤其是整题的识别率从 17.95% 提高到了 64.1%,这也证明了本文所提出方法的有效性。但是该结果依然存在很大的可提升空间。如何增强模型的适应性、拓展模型的覆盖率、进一步提升角色判定的准确率是下一步的工作重点。

参考文献:

[1] Feigenbaum E A, Feldman J. Computers and thought[M]. Cambridge, MA: MIT Press, 1963.

[2] Schoenfeld A H. Mathematical problem solving[M]. [S. l.]: Elsevier, 2014.

[3] 吴文俊. 初等几何判定问题与机械化证明[J]. 中国科学:数学, 1977, 20(6): 507-516.

[4] 张景中, 杨路, 高小山, 等. 几何定理可读证明的自动生成[J]. 计算机学报, 1995, 18(5): 380-393.

[5] Schoenfeld A H. Reflections on problem solving theory and practice [J]. The Mathematics Enthusiast, 2013, 10(1): 9-34.

[6] Hosseini M J, Hajishirzi H, Etzioni O, *et al*. Learning to solve arithmetic word problems with verb categorization[C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2014: 523-533.

[7] Kushman N, Artzi Y, Zettlemoyer L, *et al*. Learning to automatically solve algebra word problems[C]//Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014.

[8] Bobrow D G. Natural language input for a computer problem solving system[J]. Semantic Information Processing, 1964, 9(3): 281-288.

[9] Charniak E. Computer solution of calculus word problems[C]//Proc of the 1st International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1969: 303-316.

[10] Riley M S. Development of children's problem-solving ability in arithmetic[R]. [S. l.]: National Institute of Education, 1984.

[11] Kintsch W, Greeno J G. Understanding and solving word arithmetic problems[J]. Psychological Review, 1985, 92(1): 109-129.

[12] Dellarosa D. A computer simulation of children's arithmetic word-problem solving[J]. Behavior Research Methods, Instruments & Computers, 1986, 18(2): 147-154.

[13] Wong W K, Hsu S C, Wu S H, *et al*. LIM-G: learner-initiating instruction model based on cognitive knowledge for geometry word problem comprehension[J]. Computers & Education, 2007, 48(4): 582-601.

[14] 程志. 小学数学应用题自动解答系统的研究——以整数一、二步和分数基本应用题为例[D]. 北京: 北京师范大学, 2008.

[15] 马玉慧. 小学算术应用题自动解答的框架表征及演算方法研究[D]. 北京: 北京师范大学, 2010.

[16] Yu Xinguo, Wang Mingshu, Zeng Zhizhong, *et al*. Solving directly-stated arithmetic word problems in Chinese[C]//Proc of International Conference of Educational Innovation through Technology. Washington DC: IEEE Computer Society, 2015: 51-55.

[17] Wang Yingxu, Chiew V. On the cognitive process of human problem solving[J]. Cognitive Systems Research, 2010, 11(1): 81-92.

[18] Krawec J, Huang Jia, Montague M, *et al*. The effects of cognitive strategy instruction on knowledge of math problem-solving processes of middle school students with learning disabilities[J]. Learning Disability Quarterly, 2013, 36(2): 80-92.

[19] Singer F M, Voica C. A problem-solving conceptual framework and its implications in designing problem-posing tasks[J]. Educational Studies in Mathematics, 2013, 83(1): 9-26.

[20] Hooshyar D, Ahmad R B, Yousefi M, *et al*. SITS: a solution-based intelligent tutoring system for students' acquisition of problem-solving skills in computer programming[J]. Innovations in Education and Teaching International, 2016, 64(4): 787-808.

[21] Qin Yulin, Carter C S, Silk E M, *et al*. The change of the brain activation patterns as children learn algebra equation solving[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(15): 5686-5691.

[22] Kao Y S, Douglass S A, Fincham J M, *et al*. Traveling the second bridge: using fMRI to assess an ACT-R model of geometry proof[J]. Ajr American Journal of Roentgenology, 2012, 135(1): 164-166.

(上接第 2294 页)择性能进行仿真,实验随机选取六组数据进行仿真测试,六组数据的最大特征数为 20 个,将稀疏分数特征选择算法运用于这六组数据,以便对六组数据的特征重要性进行标签并排序。实验仿真结果如图 8 所示。

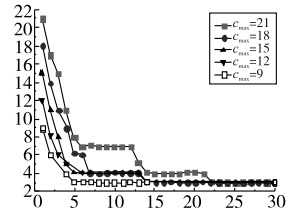

图7　不同数据集合规模的聚类结果

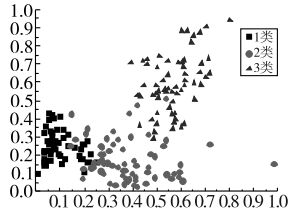

图8　稀疏分数特征选择聚类结果

从图 8 可以看出,六组数据大致被聚类成三类,聚类的依据是根据六组数据最重要的两个特征,即特征 7 和 13,将六组数据分为了三类。稀疏分数特征选择在聚类中最重要的作用,是在 20 个特征中找出了最重要的特征 7 和 13 两个特征。这是稀疏分数特征选择算法降低高维度数据独特的优势。

4　结束语

本文采用熵加权的方法对大数据集的局部结构进行划分加权运算,提高了聚类算法的稳定性,在局部结构的特征选择过程中,采用稀疏分数表示法,将高维度数据的相似局部结构进行去冗余,降低数据维度,以求得到更有效的聚类结果。

参考文献:

[1] 李晓瑜, 俞丽颖, 雷航, 等. 一种 K-means 改进算法的并行化实现与应用[J]. 电子科技大学学报, 2017, 46(1): 61-68.

[2] 邓强, 杨燕, 王浩. 一种改进的多视图聚类集成算法[J]. 计算机科学, 2017, 44(1): 65-70.

[3] Serdah A M, Ashour W M. Clustering large-scale data based on modified affinity propagation algorithm[J]. Journal of Artificial Intelligence & Soft Computing Research, 2016, 6(1): 23-33.

[4] Li Yangyang, Yang Guoli, He Haiyang, *et al*. A study of large-scale data clustering based on fuzzy clustering[J]. Soft Computing, 2016, 20(8): 3231-3242.

[5] Si Fuming, Bu Tianran. Design of a large data clustering algorithm based on Hadoop cloud computing platform[J]. Journal of Chuxiong Normal University, 2016, 31(3): 49-55.

[6] Zhang Yanfeng, Chen Shimin, Yu Ge. Efficient distributed density peaks for clustering large data sets in MapReduce[J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(12): 3218-3230.

[7] Delgado A, Romero I. Environmental conflict analysis using an integrated grey clustering and entropy-weight method[J]. Environmental Modelling & Software, 2016, 77(C): 108-121.

[8] Zhang Lijun, Zhao Fangfang. Application for technological achievements evaluations model based on entropy weight and matter-element analysis[J]. Science & Technology Management Research, 2016(6): 70-73.

[9] 邱保志, 贺艳芳, 申向东. 熵加权多视角核 K-means 算法[J]. 计算机应用, 2016, 36(6): 1619-1623.

[10] 高翠芳, 黄珊维, 沈莞蔷, 等. 基于信息熵加权的协同聚类改进算法[J]. 计算机应用研究, 2015, 32(4): 1016-1018.

[11] 蒋亦樟, 邓赵红, 王骏, 等. 熵加权多视角协同划分模糊聚类算法[J]. 软件学报, 2014, 25(10): 2293-2311.

[12] 吴杰祺, 李晓宇, 袁晓彤, 等. 利用坐标下降实现并行稀疏子空间聚类[J]. 计算机应用, 2016, 36(2): 372-376.

[13] 岳温川, 王卫卫, 李小平. 基于加权稀疏子空间聚类多特征融合图像分割[J]. 系统工程与电子技术, 2016, 38(9): 2184-2191.