单位代码：__10166__

沈阳师范大学

硕士学位论文

基于NLP的小学数学学习者语料库的构建及应用

论文作者：王蕊拂

学科专业：计算机应用技术

指导教师：宋波 教授

培养单位：科信软件学院

培养类别：全日制

完成时间：2020年04月12日

沈阳师范大学学位评定委员会

编　号：

类别	全日制研究生	√
	教育硕士	
	同等学力	

沈阳师范大学

硕士学位论文

题　目：　基于 NLP 的小学数学学习者语料库的构建及应用

所　在　院　系：科信软件学院

专　业　名　称：计算机应用技术

指　导　教　师：宋波　教授

研　　究　　生：王蕊拂

完　成　时　间：2020 年 4 月

沈阳师范大学研究生处制

学位论文独创性声明

本人所呈交的学位论文是在导师的指导下取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示了谢意。

作者签名：___王总拂___ 日期：___2020.05.31___

学位论文使用授权声明

本人授权沈阳师范大学研究生处，将本人硕士学位论文的全部或部分内容编入有关数据库进行检索；有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，允许论文被查阅和借阅；有权可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。保密的学位论文在解密后适用本规定。

作者签名：___王总拂___ 日期：___2020.05.31___

基于 NLP 的小学数学学习者语料库的构建及应用

摘要

随着全球化的影响日益增大，教育已经从重视一般语言逐渐的转向为重视学科语言。数学语言是一种简洁、准确且其自身概括能力强的一种语言，它不仅在人文学科、自然学科等领域具有广泛地应用，还是所有学科语言的核心。作为国民教育起点的小学数学课程，不仅承载着打基础的重要作用，更是受到国内外教育工作者的关注。小学数学不仅是传授数学的基础知识，还可以培养小学生的心理素质及头脑的灵活性，对学生以后的成长道路具有非常大的帮助。

对小学数学语言的研究不仅有助于提高教学质量，更能加深学生对数学的理解及应用。针对利用智能技术解决小学数学语言方面问题缺乏语料库支持的现状，本文以 2013 年-2019 年考试真题和相关知识点作为语料库的研究对象，使用 MATTER 循环法对小学数学语料库进行构建研究。首先对具体的语言现象创建模型和规格说明，按照规格说明对知识点与复习题进行标注，接下来使用刚刚创建的标注语料库进行机器学习，对结果进行评价并修改模型和算法。随着建模标注循环和训练测试循环，一旦在原有模型上进行了增加或修改，MATTER 循环将重新进行一遍。这个过程虽然繁琐且费时，但可以极大地提升算法的性能和数据的准确性，为创建黄金标准语料库提供了方法论。

本文的主要研究内容包含以下三部分：

(1)小学数学考题的现象建模及标注。本文共收集了 1480 道小学数学题型，并将这些题型分为三大类：数与代数类题型、空间与图形类题型和统计与概率类题型，根据对知识点的分析建立知识结构体系，对知识点进行难易程度和考察比例的综合分析，建立标注模型并根据该模型，使用 GATE 标注工具对生语料进行标注。

(2)自动标注的实现及描述。基于刚刚创建的语料库进行机器的半监督学习，此过程将语料库分为训练集、开发—测试集和测试集三部分。训练集用于训练任务中使用的算法，开发—测试集用于错误分析，最后在预留的语料测试集上运行。根据测试结果更改模型，以便改善之后的数据更接近黄金标准，进而改善自动标注算法的性能。

(3)面向小学数学学习者语料库的有关应用。构建了基于 Web 技术的小学数学语料库系统，其主要提供知识点查询与搜索相关试题的功能。前台界面主要提供了知识点与相关题型的模糊输入，系统根据输入进行快速处理，并向用户展示相匹配的知识点与试题清单；后台界面主要面向管理员，其提供了管理员查看、录入、删除、修改等管理的功能。

通过构建的语料库，并根据知识点的定义、题型的提问方法结合相关公式归纳解题规则，这些规则将通过 Python 语言进行标注和存储，使用这些规则进行“存题”—“识题”—“解题”的机器求解过程，并举例说明 Python 的解题效果，从而说明构建的语料库拥有实用性和有效性。
关键词：语料库；小学数学语言；半监督学习；MATTER 循环；自动标注

Construction and Application of Corpus for Primary School Mathematics Learners Based on NLP

Abstract

With the increasing influence of globalization, education has gradually changed from emphasizing general language to emphasizing subject language. Mathematical language is a concise, accurate language with strong generalization ability. It is not only widely used in humanities, natural sciences and other fields, but also the core of all subject languages. As the starting point of national education, primary school mathematics curriculum not only plays an important role in laying the foundation, but also attracts the attention of educators at home and abroad. Primary school mathematics is not only to impart the basic knowledge of mathematics, but also to cultivate the psychological quality of primary school students and the flexibility of their minds, which is of great help to the growth of students in the future.

The study of primary school mathematics language not only helps to improve the teaching quality, but also deepens students' understanding and application of mathematics. In view of the lack of corpus support for solving primary school mathematical language problems by using intelligent technology, this paper takes the real exam questions and related knowledge points from 2013 to 2019 as the research object of corpus, and USES the MATTER cycle method to construct primary school mathematical corpus. Firstly, a model and specification are created for the specific language phenomenon, and the knowledge points and review exercises are marked according to the specification. Then, the annotated corpus just created is used for machine learning, and the results are evaluated and the model and algorithm are modified. With the modeling annotation loop and the training test loop, once the original model is added or modified, the MATTER loop will be repeated. Although this process is tedious and time-consuming, it can greatly improve the performance of the algorithm and the accuracy of the data, which provides a methodology for creating the gold standard corpus.

The main research contents of this paper include the following three parts:

(1) Phenomenon modeling and labeling of elementary school mathematics examination questions. This paper collected 1480 elementary school mathematics

questions, and these questions can be divided into three categories: the number and algebra class topic, space and graphics class topic and statistics and probability class topic, according to the analysis of the knowledge system of knowledge structure, degree of difficulty of knowledge points and review the comprehensive analysis of proportion, annotation model is set up according to the model, using the GATE opposite corpus annotation tool for labeling.

(2) Realization and description of automatic annotation. The semi-supervised learning is carried out based on the newly created corpus, which is divided into three parts: training set, development-test set and test set. The training set is used to train the algorithm used in the task, the development-test set is used for error analysis, and finally runs on the reserved corpus test set. Change the model based on the test results to improve the subsequent data closer to the gold standard, thereby improving the performance of the auto-tagging algorithm.

(3) application of corpus for primary school mathematics learners. Based on Web technology, the corpus system of primary school mathematics is constructed, which mainly provides the function of querying knowledge points and searching related questions. The foreground interface mainly provides the fuzzy input of knowledge points and related question types. The system can quickly process the input and show the matching knowledge points and question lists to the users. Background interface is mainly for the administrator, it provides the administrator view, input, delete, modify and other management functions.

By building a corpus, and according to the definition of knowledge points, inductive problem solving questions questions of method and combining with related formula rules, these rules will be marked by the Python language and storage, and use these rules to "save", "general questions", "problem solving" the solving process of machine, and illustrate the Python effect in solving problems, so as to build the corpus has practicability and validity.

Key Words: corpus; primary school mathematical language; semi-supervised learning; MATTER cycle; automatic tagging

目　录

第 1 章 绪　论

随着现代社会的快速发展，计算机技术与人工智能的领域也快速崛起，基于知识体系的智能系统研究已经成为计算机领域的研究热点之一。此时人们对智能系统提出了更高的要求，呆板的检索功能与问答系统已经满足不了人们探索问题的需求，人们更希望可以通过自然语言和计算机进行交流。Iphone 中的 Siri 就是一个例子，人们可以通过平常的自然语言来代替繁琐的键盘输入进行提问，但它只能理解某些关键短语的子集，并不能完全地理解所输入的自然语言，因此如何使计算机有效且快速地理解自然语言便成为当务之急。

1.1　研究背景及意义

自然语言处理(Nature Language Processing, NLP)是计算机科学与工程领域中的一个研究方向，其源自人工智能领域中的自然语言和计算语言学研究[1]。自然语言处理的主要目标是设计和创建各种有关人工智能的应用系统，这些系统可以实现人与计算机直接通过自然语言进行各种交互。自然语言处理的应用领域主要包含自动问答系统、机器翻译、语音识别、文档摘要、文档分类等[2]。但仅仅给计算机输入大量的数据就希望它能够学会说话是不可行的，应该先准备好易于计算机发现模式和推理的数据，再通过给数据集增加相关元数据来实现这一目标，由此可见语料库的研究是开发智能类人求解技术的关键环节。

随着自动标注的实现以及机器学习开始受到广泛的关注，计算机技术与语料库的结合已经成为了现代化语言研究的重要手段，其特点是定性分析+定量调查，体现出理性主义与经验主义在语料库语言学上的辩证统一[3]，是当代语言学研究的特色。现如今多数的研究方法是根据语料库对现有知识进行验证和完善，还有些研究方法是对语料库进行研究以便发现其他相关的知识[4]。纵观国内外语料库的发展，大多数是各国对自身语言和语法上的研究，在其他领域上，尤其是小学数学的领域对语料库进行的研究接近为零，针对这一现象，本文将构建一个小学数学领域的学习者语料库，使人与计算机能够交互并掌握该领域的相关知识，以增强小学生的数学学习能力，并为有关小学数学语言的研究提供素材。

1.2　NLP 国内外研究现状

在 20 世纪 50 年代末我国已经开始对自然语言进行研究，在计算机应用领域有许多研究人员从事着有关自然语言检索的研究，早期的主要研究是对在开发语言检索系统时遇到的一些问题进行理论分析，最终的研究结果也比较少。直至 20 世纪 80 年代初期自动分词的检索技术被提出，众多领域专家和研究人员在中文语言信息处理的领域中取得了非常大的进展，但还是没有办法突破处

理单个句的限制，其原理过多得依赖着统计学的算法，这也是目前 NLP 技术中遇到的最主要的问题之一[5]。

在 20 世纪 60 年代国外就已经在自然语言处理的领域取得了明显的进步，开始建立出一批基于 NLP 的应用系统，例如法国的自然语言处理系统是先将整句文字拆分成单个的成名词词组，然后再与机器内存储的词表进行匹配，这个系统在自然语言接口技术与情报语言检索结合的方向取得了巨大的成果[6]。在这个方面，美国将自然语言处理应用到了情报检索领域，并开发出 WIN 系统，FREE- STYLE 系统与 MNIS 系统，这些系统以文本信息与问题的相关性进行排序的结果为根据进行检索，实现了真正的非布尔逻辑检索[7]。

在现有的 NLP 技术基础上，本文搜集了各省的考试真题和相关知识点作为语料进行分析，研究并创建了面向小学数学学习者的语料库，以实现对用户输入信息的快速处理，并匹配展示相应的知识点与题型的功能。

1.3 小学数学语言国内外研究现状

在这个飞速发展的时代，全球化不仅是一种概念，还是一种现象过程，深刻地影响着文化、经济、政治等社会生活的各个方面。教育作为社会发展的重要基础，需要在这个进程中进行不断地创新，顺应并跨越全球化所带来的机遇与挑战。世界上许多国家都逐渐意识到数学这门知识的重要性，例如荷兰自 60 年代末就已经开始了对数学教育体制的改革[8]，直至 90 年代初期，基本上所有的荷兰小学都已经开始使用新的数学教材，其是依据现实教育思想所编撰的，锻炼学生将已拥有的知识与生活经验相结合的能力，从而达到学习和理解数学的目的。英国在这方面还形成了具体化的教育系统，数学实践应用不仅被确立为独立的教学目标，还进入了英国国家数学课程标准的规定[10]。日本传统的数学课程主要受东方文明的影响，学习的内容相对较多且教学方法多以教师灌输知识内容为主，在教学的过程中不太注重学生的感受与体验，自 90 年代末日本颁布了全新的小学数学学习的指导纲要，其要求了教师在教授过程中明确地体现出数学教育的活动化、个性化和实践性[11]。

在 20 世纪末，我国进行了全面的教育改革，以提高国民素质为根本宗旨，语言教育方面得到了前所未有的重视。整个课程的内容、目标、理念、评价、实施中最为重视的就是有关英语语言的学习与应用，不仅设立了从小学到大学各个阶段系统化的英语教学体系，还在高价环节中加重对英语语言应用能力的评价。因此我国不仅成为了世界上学习英语人数最多的国家，学生的英语学习与应用能力也得到了大幅度的提升[12]。进入 21 世纪后，随着中国国力的稳步提升和全球化程度的加深，计算机技术领域与教育领域的交流日益增多。在这个背景下，英语仅仅作为一般性意义的应用语言已经满足不了各国各领域之间的

交流，学科语言渐渐地出现在全球化的舞台上[13]。而在众多的学科语言学习中，数学语言毫无疑问的是各个学科语言的基础，曾有学者提出：只要完全掌握了数学语言的精髓，就相当于掌握了表达科学信息和实践活动中遇到的实际问题的技术[14]。这种技术主要应用于在两个方向：一是有关科学领域的研究，越发的注意到数学语言具有的独特的解释性作用；二是有关日常生活的方面，数学语言突出了其自身的简洁性与丰富性的特点。

数学作为一种学科语言，它不仅能简便而准确地表达和交流想法。一个人其自身的语言能力不仅决定了这个人是否能够顺应社会的发展，还涉及了其在社会生活中的生存的品质，而这个人其自身的数学语言能力则决定了这个人在相关领域发展的水准和研究程度[15]。本文旨在建立一个面向小学数学学习者的语料库，此语料库有助于教学质量的提高，加深学生对数学的理解及应用，不仅为机器求解提供语料库技术的支持，还可以为小学数学知识的有关研究提供相关素材。

1.4 语料库国内外研究现状

随着电子设备的普及和万维网的成长，语料库在规模上不断地扩大，起初的语料库内容较少，词的容量很小，基本没有达到百万字以上，并且全是基于各国语言学对英语进行的研究。60 年代时期 Brown 语料库横空出世，它是第一个涉及类别比较广泛的语料库，该语料库将语料分为三个层次，并且每个层次的语料所占比例都是固定的[16]，是最具有典型代表性的平行语料库之一。在 70 年代初期，伦敦大学收整了 110 多小时的新闻发言和收音机电台等英式英语的口语语料，通过对英语语法的调查将信息整理成纸质文档，由于 LLC 语料库是通过英语用法调查而建立起来的，研究者从语音语料库中发现并构建了随机模型，这使得语音识别功能的实现成为了可能[17]。

语料库技术从 80 年代末进入了中国，上海交通大学的杨惠中教授研究开发的 JDEST 英语语料库最早成型[18]。后来在中国进行分析设计并构建的还有中国专业英语学习者语料库、中国英语学习者口语语料库、大学学习者英语口语语料库等。这些语料库为千千万万想要学习英语的人们提供了参考材料，不仅有效地帮助了人们去准确迅速地学习英式英语和英语口语，还掀起了一阵年轻人热爱英语的浪潮。同时期国家语言文字应用委员会构建了第一个大规模的关于汉语书面语的平衡语料库：国家现代汉语语料库，它的第一批语料是来自于 20 世纪初期的人工录入的印刷版本语料，那时规模已经达到了 8000 万字，并且仍然不断增长着，现以网络电子文本语料为主，达到了 1 亿多字，包含了八大类和 30 个小类，它不仅是 20 世纪最大的现代汉语语料库，还是语料库中的均衡性的典型代表之一[19]。

纵观国内外语料库的发展，大多数是各国对自身语言和语法上的研究，在其他领域上，尤其是小学数学的领域对语料库进行的研究接近为零[20]，针对这一现象，本文将构建一个小学数学领域的学习者语料库，用以支持小学数学方向的类人求解问题的前提需求，并为小学数学知识的相关研究提供素材。

1.5 本文组织结构

本文的研究内容主要分为六个章节，如下安排：

第一章，绪论部分：介绍了 NLP 技术、小学数学语言、语料库的国内外研究现状和研究内容，创建小学数学语料库的意义。

第二章，技术分析：介绍了开发语料库所使用的方法论、算法以及工具。

第三章，构建语料库：根据 MATTER 循环法中的建模标注循环对生语料库进行标注并审核。

第四章，机器学习：根据 MATTER 循环法中的训练测试循环使用刚刚创建的语料库对算法进行机器学习训练，依据测试结果改进模型和算法，以达到语料的黄金标准。

第五章，语料库的应用：展示了语料库系统的功能界面，和 Python 解题方法的实现，说明了所创建的语料库拥有实用性和有效性。

第六章，总结与展望。对论文的工作内容和研究成果进行分析与总结，并对接下来要做的工作进行展望。

第 2 章 技术分析

随着自动标注的实现以及机器学习开始受到广泛的关注，将语料库知识与计算机技术相结合的研究方法已经成为当代对语言进行研究的重要手段之一，其特点是定性分析+定量调查，体现出理性主义与经验主义在语料库语言学上的辩证统一，是当代语言学研究的特色。语料库有多种类型，根据它的研究目的和用途划分，可以分成同质的、异质的、专用的和系统的语料库；根据所收集语料的语种划分，还可以分为多语的、双语的和单语的语料库；根据所收集语料的搜索单位划分，也可以分成短语的、语句的、语篇的语料库。其中的多语和双语语料库根据其语料的构成方式，又可以将语料库分成比较语料库和平行语料库，前者将描述相同内容的不同语言文本归纳到一起，用于语言对比研究的领域，后者所收集的语料构成译文关系，用于双语词典编撰、机器翻译等应用领域。

2.1 研究内容及方法

本文以 2013 年—2019 年各省份的小升初数学试卷和相关知识点作为语料库的研究对象，对其进行问题分类和数据分析以构建语料库，这不仅使学习者能够快速学习知识、提高学习效率，并为有关小学数学语言的研究提供素材。本文所研究的内容主要分为以下三点：

(1)小学数学题型分类及其表征研究。在对语料库进行研究之前，本文共收集了 21 套小升初数学试卷，针对这些没有经过处理的语料，首先需要根据小学数学的知识结构，构建可靠的知识体系并为其建立知识点索引表，对每道试题所涉及的知识点进行标注，然后统计和概括分析试题中相关知识点出现的频率，最后将所有知识点根据其难易程度和相关题型所占的比例进行综合分析，将整套试卷所考察的知识点主要分成三部分并概括其特征，使计算机能够在自然语言中捕获有关知识点的信息和已知条件从而实现“理解题意”的目标。

第一类是数与代数知识点，此类题型主要关于数的运算与简易方程，其占据了小升初试卷的 78%[21]。其中有关数的运算题型所占比例最大，并且这种题型的题意表征相比于其他题型较为简单，如果对题目意义的表征描述得正确，那么计算机对此类题型的题目理解可顺利完成。

第二类是空间与图形知识点，此类题型主要关于图形的测量及变换，其中关于图形的测量计算最常见，并且这类知识点与数与代数中的运算类题型息息相关，可以实现大部分真题的标注。

第三类是统计与概率知识点，此类题型主要关于统计与可能性的运算，其中有关可能性的计算题型所占比例最多。

(2)面向小学数学知识点语料库的构建及应用。根据 MATTER 循环方法首先进行的是“建模-标注”循环，先对具体的语言现象创建模型和规格说明，按照规格说明对知识点与复习题进行标注，在进行标注时不断地对模型和规格说明进行改进与简化，从而达到黄金标准。然后进行的是“训练-测试”循环，首先将语料库分成两个部分：开发语料库和测试语料库。其中开发语料库进一步分成两个部分：训练集和开发-测试集。其中的训练集经常用于训练算法，而开发-测试集则经常用于分析测试时遇到的错误。当算法在训练集学习结束后就可以到开发-测试集上运行算法，在这个过程中如果算法没有准确地对语料进行标注，就调整并重新训练算法然后再重新进行测试[22]，不间断地重复以上过程直至算法获得满意的训练结果。训练完成后算法在最先预留好的测试语料库上进行测试，这些数据在训练中和开发测试中从来没有使用过，因此能够获得算法在全新的数据上的表现结果，再根据这个结果不断地改进语料库的模型与自动标注的算法。

(3)小学数学学习者语料库的应用。本文使用基于 Web2.0 的技术对语料库系统进行可视化的展示，其主要提供知识点查询与搜索相关试题的功能。前台的界面主要为用户提供了知识点与相关题型的模糊输入，系统根据输入进行快速处理，并向用户展示相匹配的知识点与试题清单；后台界面主要面向管理员，其提供了管理员查看、录入、删除、修改等管理的功能。按照此前分类的知识体系，选取了考察所占比例较多的知识点进行研究分析，比如四则混合运算题型、简易方程题型、比的运用题型、图形的面积或体积题型、逻辑推理题型、统计与概率题型等。根据知识点的定义，将题型问法与相关公式相结合以用来归纳该类题型的解题规则。这些规则将通过 Python 语言进行标注和存储，使用这些规则进行“存题”—“识题”—“解题”的机器求解过程，并举例说明 Python 的解题效果，从而说明构建的语料库拥有实用性和有效性。

2.2 MATTER 循环标注法

由于标注的过程不是线性的，所以在标注过程中的定义任务、标注和评价需要通过多次迭代，以获取最佳的结果。如今大部分标注工作都是采取众包标注方式的，将标注任务拆分成小任务，之后发给许多人，每个任务有少量的报酬，从而代替一小部分的高工资标注者标注整个语料库。最著名的亚马逊土耳其机器人就是使用这一方法，研究人员创建智慧任务（Human Intelligence Tasks，HIT）并将其发布在一个工作布告栏上[23]，Turkers 可以有选择地接受任务，在完成接受的标注任务后可以获得少量的报酬。

尽管亚马逊土耳其机器人这种省时且低成本的方法听起来很理想，但是人类智慧（HIT）系统并不是绝对完美的。首先它要求将每个标注任务拆分为“微

任务”，但并不是所有的标注任务都适合拆分成微任务，这会削弱标注工作人员对标注目标的直觉，可能会影响标注的最终结果。其次是数据质量问题，研究人员很难对众包的标注人员提出要求，由于很大一部分的 Turker 把人类智慧任务作为收入的主要来源，导致所收集的数据质量不用，很难分辨好坏。

本文通过 MATTER 开发循环实现小学数学语料库的建设。此过程具体步骤为：建模（Model）、标注（Annotate）、训练（Train）、测试（Test）、评价（Evaluate）、修改（Revise）[24]。图 2.1 详细地描述了这一循环的具体步骤。首先对具体的语言现象创建模型和规格说明，按照规格说明对知识点与复习题进行标注，接下来使用刚刚创建的标注语料库进行机器学习，对结果进行评价并修改模型和算法。随着建模标注循环和训练测试循环，一旦在原有模型上进行了增加或修改，MATTER 循环将重新进行一遍。这个过程虽然很繁琐且费时费钱，但可以大大地提升算法的性能和数据的准确性，为创建黄金标准语料库提供了方法论。

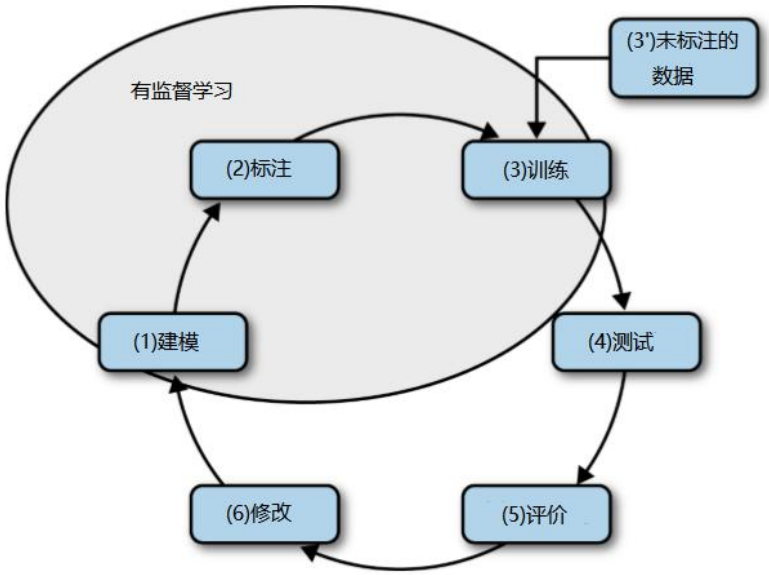

图 2.1 半监督学习下的 MATTER 循环

2.3 机器学习算法

分类是为给定的输入选择正确的类标签，在基本的分类任务中，每个输入的数据被认为是与其他所有输入的数据相隔离的[25]，并且标签集是预先定义好的。本文根据小学数学语料具有的准确性、简洁性和抽象性的特点选择了朴素贝叶斯分类算法进行机器学习。贝叶斯分类方法的理论依据是贝叶斯定理为理论，其原理采用了概率推理方法[26]，即通过计算预分类样本在各个已知类别上的后验概率，然后根据这个概率将这个样本分类为其对应的类别中。然而在计算后验概率的过程当中，需要先知道属性的条件概率，与数据集中每个类别的

先验概率。每个类别的先验概率能够使用统计的方法获得，且属性的条件概率也能够通过统计的手段或者假设的分布模型进行估计[27]。

朴素贝叶斯分类器(Naive Bayes Classifier, NBC)是一种生成式分类器，是贝叶斯分类器中使用的最广泛的模型之一[28]，其通过考虑特征概率来预测分类。在总样本 S 中，假设有 A、B 两个随机事件，当事件 A 发生时，事件 B 也发生的概率被称为事件 B 在给定事件 A 的情况下发生的条件概率(也称之为后验概率)，记作 P(B|A)。相应地将 P(A)称作无条件概率(也称之为先验概率)。其中的条件概率可以由以下公式进行计算：

$$P(B \mid A) = \frac{P(A \bullet B)}{P(A)} \quad \cdots\cdots(2.1)$$

由概率的乘法定理可将条件概率公式转换成：

$$P(A \bullet B) = P(B \mid A)P(A) \quad \cdots\cdots(2.2)$$

再假设有 n 个事件 B_1，B_2，…，B_n，为样本空间 S 的一个划分，且 $P(B_i) > 0\{i = 1，2，\cdots，n\}$，则有全概率的计算公式：

$$\begin{aligned} P(A) &= P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_n)P(B_n) \\ &= \sum_{i=1}^{n} P(A \mid B_i)P(B_i) \end{aligned} \quad \cdots\cdots(2.3)$$

根据条件概率的定义以及全概率的计算公式可得到贝叶斯定理：

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)} \quad \cdots\cdots(2.4)$$

再假设有一个变量集 U = {A,C}，其中 A = {A_1，A_2，…，A_n}代表着变量 A 包括 n 个条件属性，C = {C_1，C_2，…，C_m}代表着变量 C 包括 m 个类标签。根据朴素贝叶斯分类模型的特点假设所有的条件属性 $A_i(i = 1，2，\cdots，n)$都是类变量 C 的孩子节点。将一个待分类的样本 X = {a_1，a_2，…，a_n}分配给类 $C_i(1 \le i \le m)$，当且仅当：$P(C_i|X) > P(C_j|X)$ $(1 \le i,j \le m,i \ne j)$时，根据贝叶斯定理有：

$$P(C_i \mid X) = \frac{P(C_i)P(X \mid C_i)}{P(X)} \quad \cdots\cdots(2.5)$$

假如事先不了解类变量 C 在数据集中的发生概率时，可以先假定每个类别的发生概率全都相等。既有公式：

$$P(C_i) = P(C_j),(C_i, C_j \in C，\ i \ne j) \quad \cdots\cdots(2.6)$$

并且根据这个公式对 $P(C_i|X)$进行最大化。要不然就会最大化 $P(X|C_i)P(C_i)$。由于 P(X)在面对所有的类别是均代表为常数，因此有公式：

$$P(C_i \mid X) = \frac{P(C_i)P(X \mid C_i)}{P(X)} \propto P(C_i)P(X \mid C_i) \quad \cdots\cdots\cdots\cdots\cdots\cdots(2.7)$$

因为朴素贝叶斯分类算法中假设了条件属性是相互独立的，所以有公式：

$$P(C_i \mid X) \propto P(C_i)\prod_{k=1}^{n} P(a_k \mid C_i) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(2.8)$$

在此公式中 $P(C_i)=S_i/S$，其中 S_i 代表着类 C_i 在训练样本中的具体实例个数，S 表示着训练样本的总数。因此 NBC 模型的算法公式可以表达为：

$$NB(X) = \arg\max_{C_i \in C} P(C_i)\prod_{k=1}^{n} P(a_k \mid C_i) \quad \cdots\cdots\cdots\cdots\cdots\cdots(2.9)$$

其中概率 $P(a_1|C_i)$，$P(a_2|C_i)$，…，$P(a_n|C_i)$可由训练样本进行估值。在处理待分类样本 X 时，应需分别计算每个类别 $C_i \in C$ 的条件概率 $P(C_i)P(X|C_i)$[30]。当且仅当 $P(C_i)P(C_i|X)>P(C_j)P(C_j|X)(1\leq i,j \leq m,i\neq j)$时，样本 X 属于类别 C_i。

朴素贝叶斯分类模型算法的具体步骤如下：

①对待分类的数据集进行预处理，其中包括缺失值补充和属性值离散化[29]；

②统计训练样本的数量 S、类为 C_i 的样本数 S_i、类为 C_i 的样本中的属性为 A_k 取值为 a_k 的样本个数 S_{ik}；

③计算 $P(C_i)=\frac{S_i}{S}$ 和 $P(a_k|C_i)=\frac{S_{ik}}{S_i}$；

④利用分类模型：$NB(X) = \arg\max_{C_i \in C} P(C_i)\prod_{k=1}^{n} P(a_k \mid C_i)$ 得出待分类样本 X 的判定结果。

通过上述内容对朴素贝叶斯分类算法的分析和理解，可以总结出算法的优点主要有两点：第一，由于 NBC 算法的基本思想是假设属性条件间是相互独立的[30]，这使其在对具有不同属性特征的数据进行分类时也能保持分类方法的稳定性，并且不用考虑这些属性之间存在的关系。第二，NBC 模型的算法结构也同样十分简单，需要用户所估计的参数也相对较少，因此 NBC 算法在训练以及分类数据时使用的计算开销也相对较小，简单且高效地进行分类是 NBC 算法的主要优点。

当然，NBC 算法同样存在着不足。在算法应用到实际分类时，基本上满足不了该算法对属性条件间相互独立的假设，面对一些具有高度相关性属性的数据时[31]，假如直接运用 NBC 算法对语料进行分类，其结果很难达到预期的结果。此外，在面对处理不完整的数据时，或是出现了属性条件极度不平衡的数据时，就可能直接导致某个甚至某些属性的后验概念出现非常大的偏差，因此

将会影响最终的分类结果。不过，目前的统计学家们已经研究出了相关方法用以解决不完整的数据和过度拟合的数据，其中一个就是平滑技术，例如加法平滑，其使用已知的最大似然估计概率[32]，并根据语料库和词汇表对概率进行折减，这个技术可以在一定的程度上提高 NBC 算法的性能。

2.4 开发环境

2.4.1 GATE 标注工具

目前市面上有许多标注标注工具可以选择，比如 GATE、Knowtator、MAE、MMAX2 等。本文使用自然语言处理框架(General Architecture for Text Engineering，DATE)进行小学数学学习者语料库的构建，GATE 是开发和部署处理人类语言的软件组件的基础设施，它有将近 15 年的历史，并积极地应用于涉及人类语言的各种计算任务[33]。GATE 擅长分析各种形状和大小的文本，从大型公司到小型初创公司，甚至数百万欧元的研究联合会本科项目都使用该软件进行开发。

GATE 的核心功能有：专业数数据结构的建模和持续性；测量、评估、基准测试；可视化和编辑注释、本体、解析树等；用于快速成型和有效实施浅层分析法的有限状态转换语言；提取机器学习的训练实例；可插拔机器学习实现(Weka\SVM Light 等)。

在核心功能之上，GATE 包含用于各种语言处理任务的组件，例如解析器、形态学、标记、信息检索工具、各种语言的信息提取组件以及许多其他组件。GATE 支持多种格式的文档，包括 XML、RTF、email、HTML、SGML 和纯文本等。

2.4.2 Python 语言

Python 语言在教育、科研、工业领域的应用都十分广泛，其不仅提高了软件的质量和生产效率，还具有可维护性，因此在世界各地都备受欢迎。它的文法和语义简单易懂，并且有强大的字符串处理功能，其自带的函数非常适合处理语言数据。Python 还自带了强大的标准库，包括数值处理、网络连接和图形编程等组件。作为解释性语言，Python 便于交互式编程；作为面向对象的编程语言，Python 蕴蓄的数据和方法被封装和重用以便用户调用；作为动态语言，在程序运行时 Python 蕴蓄的属性可以添加到对象，允许变量进行自动类型转换，以提高开发效率[34]。

本文在研究时还下载了自然语言工具包(Natural Language Toolkit， NLTK)作为使用 Python 进行 NLP 编程的基础工具。NLTK 包含了大量的数据和文档，它不仅提供了与自然语言处理有关的数据集以表示基本类，还提供了文本分类、

文法分析、词性标注等任务的接口及实现，这些内容组合起来可以解决复杂的问题。

第 3 章 构建语料库

想要计算机在训练算法时能够获得预期的分类效果，首先需要对具体的语言学现象进行描述和编码，这些语言数据的特征必须足够丰富。描述这些自然语言的方法通常来自于对语言现象的理论建模，这些描述不仅是标注具体语言的基础，其自身特征还可以用于测试文本识别或训练标注算法的开发中。本章介绍了如何根据 MATTER 循环中的 MAMA(建模—标注—建模—标注)循环创建语料库并标注语料。

3.1 语料库选材

自从 2001 年《全日制义务教育数学课程标准（实验）》实施以来，现有各种版本的小学数学教材层出不穷，孙晓天先生曾在文章中指出：“迄今已有小学 6 套，中学 9 套，高中 6 套共 21 套根据《标准》要求编写的新数学教材出版使用”[35]。这些教材分别属于 13 家出版社，根据在 Internet 上对辽宁、山东、广东、河南、河北五省的小学数学教材的使用情况进行调查，其结果显示显示大部分地区使用的是人民教育出版社和北师大出版社出版的小学数学教材。本文使用北师大版数学教材对数学语言进行分析，并在 Internet 上下载 2015 年~2018 年各省份的小升初数学试卷，首先将试题归纳为 3 种题型，再分别以年份为分界线，将试题放入 Word 文档中进行存储，如图 3.1 所示。

名称	修改日期	类型	大小
2013年小升初数与代数题型整理.docx	2019/1/4 星期五 上午...	DOCX 文档	430 KB
2014年小升初数与代数题型整理.docx	2019/1/4 星期五 上午...	DOCX 文档	397 KB
2015年小升初数与代数题型整理.docx	2019/1/4 星期五 上午...	DOCX 文档	514 KB
2016年小升初数与代数题型整理.docx	2019/1/4 星期五 上午...	DOCX 文档	500 KB
2017年小升初数与代数题型整理.docx	2019/1/4 星期五 上午...	DOCX 文档	612 KB
2018年小升初数与代数题型整理.docx	2019/1/4 星期五 上午...	DOCX 文档	486 KB

图 3.1 各年小升初数与代数类题型的文档整理

3.2 建立模型与规格说明

在 MATTER 开发循环中，第一步就是“现象建模”（即 MATTER 周期中的“ M”），根据收集的相关数据，为即将进行小学数学语料库的标注任务建立模型。首先应将生语料进行分类并归纳成详细的知识体系，以数与代数类题型为例，根据人教版教学大纲的要求，将有数与代数类题型的术语整理出来并对其进行详细的描述[36]，形成对应的 EXCEL 文档，其整理部分如图 3.2 所示：

整除		整数a除以不为0的整数b，商为整数，没有余数
倍数	因数	数a能被数b整除，则a是b的倍数，b是a的因数
公因数	最大公因数	几个数公有的因数是这几个数的公因数，其中最大的数是最大公因数
公倍数	最小公倍数	几个数共有的倍数是这几个数的公倍数，其中最小的数是最小公倍数
互质数		两个数的公因数只有1
质数		一个数只有1和它本身两个因数
合数		一个数除了1和它本身还有别的因数
加法		两个数合并成一个数的运算
减法		已知两个加数的和与其中一个加数，求另一个加数的运算
乘法		求几个相同加数的和的简便运算
除法		已知两个加数的和与其中一个加数，求另一个加数的运算
四则混合运算		在一个算式里，含有加减乘除四种运算中任意两种以上的运算
简易方程		含有一个未知数的等式

图 3.2 数与代数类题型数术语整理(部分)

该类题型中有 36 个知识点术语作为考察对象，可以将这些知识点分为两大类：数的运算和寻找规律，这些知识点间的联系如图 3.3 所示。

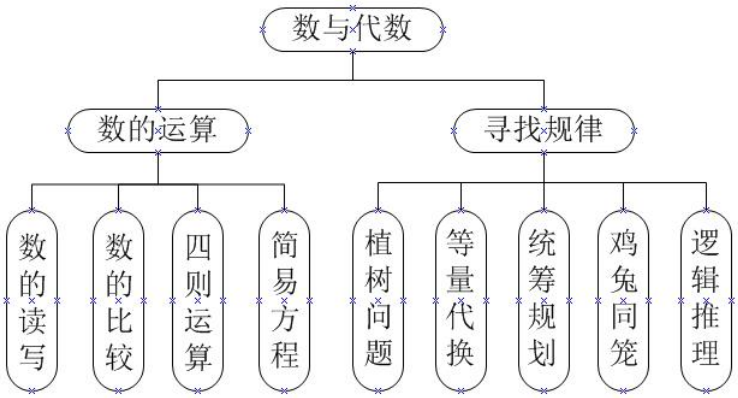

图 3.3 知识点之间的联系

在这些知识点中占比最多的是四则运算与简易方程，在小升初真题中大部分题型以考察学生的计算能力与逻辑推理能力。然后根据以上知识点的整理，使用 XML 文档类型定义(DTD)表示该模型，这类文件可以清晰地描述任务的名称、元素和属性，是目前很流行的一种通用标记语言[37]。结合数与代数类知识点的考察情况，以简易方程知识点题型为例，对知识点进行建模，具体代码如下所示：

```
<!ELEMENT equation ( title | name | time | figure | unit | relation | question)>
<!ATTLIST relation arith ( add | sub | multi | divi | mix)>
```

3.3 标注与审核

在创建一个语料库和模型之后，便开始进行实际的标注过程（即 MATTER 周期中的“A”）。根据标注模型指南对生语料进行标注，本文使用自然语言处理框架(General Architecture for Text Engineering，DATE)对小学数学语料库进行标注。GATE 是开发和部署处理人类语言的软件组件的基础设施，它用于各种语言处理任务的组件，例如解析器、形态学、标记、信息检索工具、各种语

言的信息提取组件以及许多其他组件[38]。GATE 支持多种格式的文档，包括 XML、RTF、email、HTML、SGML 和纯文本。

以小学数学简易方程知识点题型为例，根据上节所分析的模型与规格说明对试题进行标注。首先将小学数学简易方程题型的 word 文档导入 GATE 中，如图 3.4 所示：

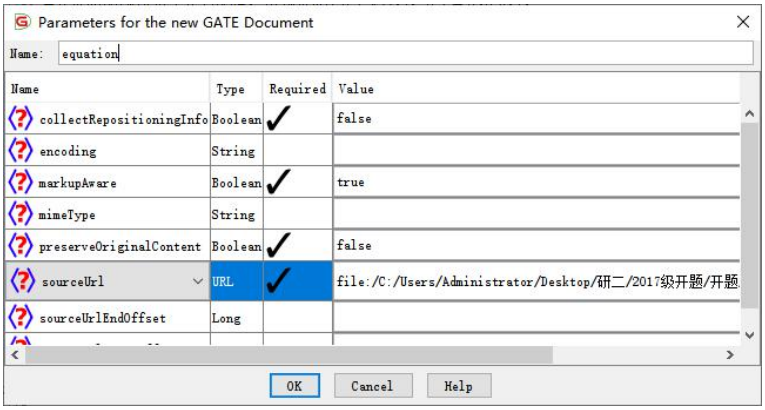

图 3.4 导入小学数学简易方程题型文档

然后按照之前所创建的规格说明进行人工标注，如图 3.5 所示：

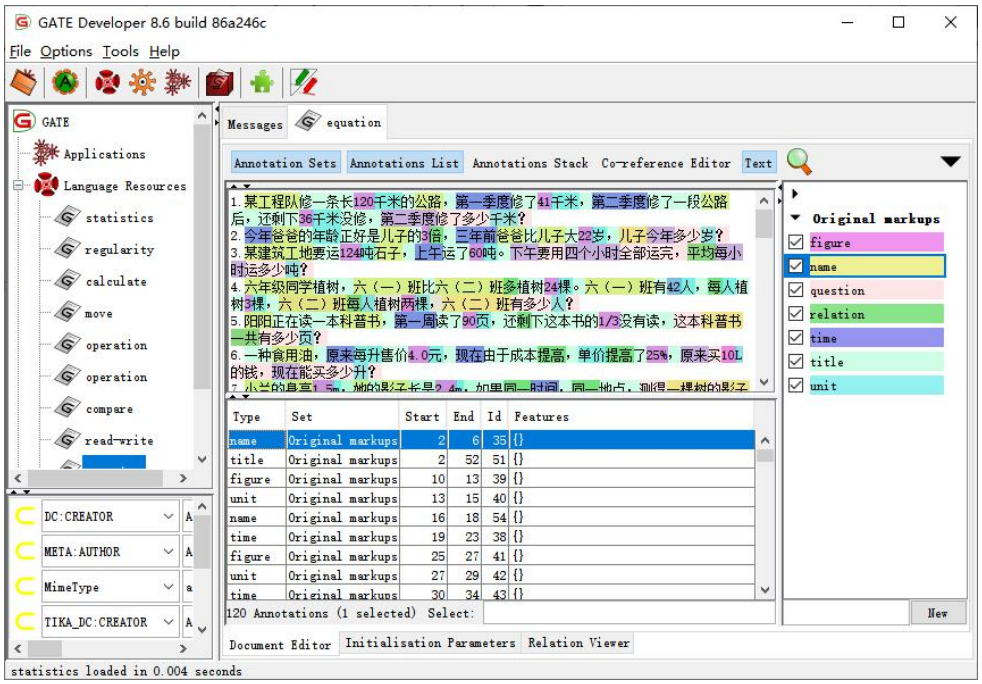

图 3.5 标注小学数学简易方程题型文档

最后输出的 XML 格式文档就意味着完成了对生语料的标注。但是在标注过程中录入信息时也可能出现问题，如果标注时过于疲劳或注意力不够集中，就可能偶然地填入错误的标签。因此在标注完成后，将标注的数据计算标注一致性（IAA）得分，如果这些得分较低，则修改模型然后重新标注[39]；如果得分比较理想，就可以在数据上进行审核以产生黄金标准语料库，然后使用它训练和测试机器学习算法。这一阶段称为 MAMA（建模-标注-建模-标注）循环，

如图 3.6 所示，其中关键部分是审核取得标注人员的标注结果，并使用它产生适用于训练机器学习的语料库。

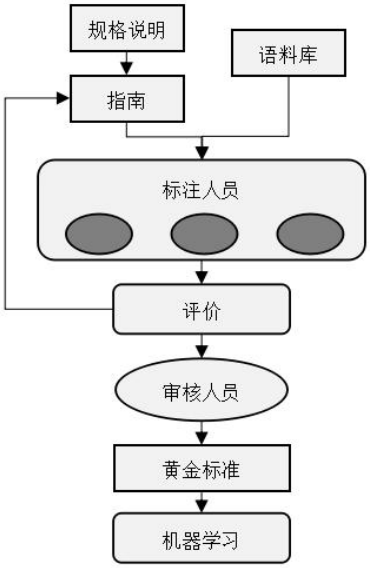

图 3.6 标注过程

第 4 章　机器学习

计算机在学习时面临的问题是如何处理在训练语料库中没有出现过的数据，计算机缺乏足够多的数据来计算，这就是所谓的数据稀疏问题。在前文中曾经谈到使用数据集的相对频率来计算给定类别的先验概率，即最大似然估计(MLE)方法。这种方法将给任何未见过的事件赋予零概率，但这个数值在预测新语料库上的行为时起不到任何作用。为了解决该问题，统计学家已经开发了许多方法来降低已知事件的概率以便可以为语料库中未出现的事件赋以非零概率，其中一个技术是平滑技术。

4.1　选择算法

创建标注语料库并将它应用于合适的机器学习算法是一件困难的任务，本文采用有监督学习算法通过已标注的数据逐渐提高系统的性能，从而实现自动地对文本进行分类和标注（即 MATTER 周期中的“T”）。在小学数学测试题中一些问题会包括两类知识点的考核，普通的决策树学习可能会导致标注的不准确性，而贝叶斯是一种生成式分类器，通过考虑特征概率来预测分类。根据创建的小学数学语料库的模型与标注，本文选择了具有一般性和广泛应用性的贝叶斯学习算法来实现自动标注，贝叶斯理论常用于解决不确定性的问题，它基于了数学领域中的统计学和概率论，因此含有牢固的数学基础[40]。朴素贝叶斯分类方法训练方法简单、且具有相当的高效性和健壮性，现已成功地应用到模型选择、分类、聚类等数据挖掘任务中[41]。

假设 Ccx 是训练集合中类别为 C_i 且属性 X 取值为 X_j 的例子总数，Cc 为类别是 C_i 的例子总数，N_x 是属性 X 所有可能取值的个数(例如，假设 X 可以取 1，2，3 三个值，则 N_x 为 3)，则 $p(X_j/C_i)$的一般计算公式为：$p(X_j/C_i)=Ccx/Cc$，该公式中当 Ccx 为零时，将会导致整个分类公式的计算结果为零，这个就是所谓的数据稀疏问题。为克服零概率和过度拟合问题，现一般研究都直接选用最早提出的且简单有效的拉普拉斯平滑方法[42]。其假定了统一的先验概率，即假定每个属性的取值都至少拥有一个训练例子，因此估计条件概率 $p(X_j/C_i)$的公式变为：$p(X_j/C_i)=(Ccx+1)/(Cc+N)$。

在查阅文献时发现了一种更为一般的概率估计方法——M 估计方法[43]，其基本思想就是用户可以根据领域的特点适当地扩大训练数据集的实例数目，即为训练数据集添加 m 个等效的实例，计算公式为：$p(X_j/C_i)=(Ccx+mp)/(Cc+m)$，其中 m 是附加实例的个数，具体取值由领域专家指定，p 是概率的先验估计。确定精确的先验概率 p 比较困难，一般取统一的先验概率 $1/N_x$。

接下来本文使用 NLTK 中自带的数据集对以上两种估计方法进行分类结果

对比，其准确率曲线对比图如图 4.1 所示。从图中可以看得出来，基于 M 估计的平滑方法在有一个、两个、三个和九个属性参与了分类的情况下，其分类效果明显好于 Laplace 方法。因此本文使用基于 M 估计的平滑方法以改进多关系朴素贝叶斯分类器，并通过标注过的和未标注过的语言数据来训练机器学习算法。

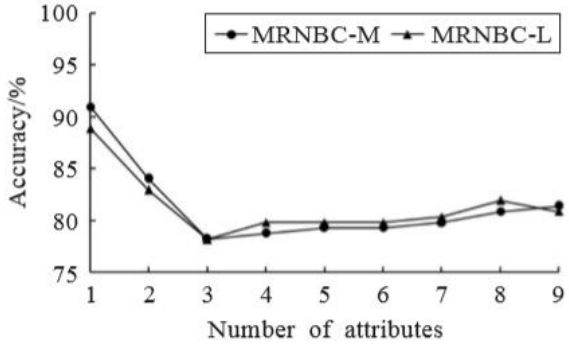

图 4.1 MRNBC-M 和 MRNBC-L 的分类准确率比较

4.2 测试算法

在选择了算法和算法所用的特征后，就可以实际地用黄金标准语料库测试算法并评价测试结果。本文将语料库分成了开发语料库和测试语料库两个部分。其中开发语料库又分成训练集和开发-测试集两个部分。训练集用于训练算法，开发-测试集用于错误分析。当算法在训练集学习结束后就可以到开发-测试集上运行算法，在这个过程中如果算法没有准确地对语料进行标注，就调整并重新训练算法然后再重新进行测试，不间断地重复以上过程直至算法获得满意的训练结果[44]。训练完成后算法在最先预留好的测试语料库上进行测试，这些数据在训练中和开发测试中从来没有使用过，因此能够获得算法在全新的数据上的表现结果，再根据这个结果不断地改进语料库的模型与自动标注的算法。这一阶段称为 TTER（训练-评价）循环，如图 4.2 所示：

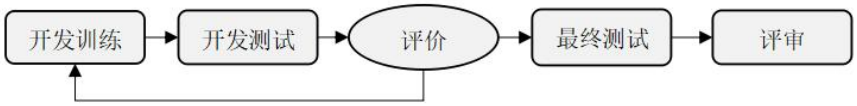

图 4.2 训练-评价循环

计算机会根据算法自动生成 XML 文件，以一道简易方程类题型为例。例：某工程队需要修一条长 120 千米的公路，第一季度修了 41 千米，第二季度修了一段公路后，还剩下 36 千米没有修，请问第二季度修了多少千米？实现的 XML 文件如下：

<title>

<name>工程队</name>

<figure>120</figure><unit>千米</unit>
<name>公路</name>
<time>第一季度</time>
<figure>41</figure><unit>千米</unit>
<time>第二季度</time>
<name>公路</name>
<relation>剩下</relation>
<figure>36</figure><unit>千米</unit>
</title>
<question>
<time>第二季度</time>
<unit>千米</unit>
</question>

判断分类器性能的优劣，可以通过计算复杂度、分类准确率、健壮性和可伸缩性进行衡量[45]。本事使用一个混淆矩阵来帮助分析算法在任务的哪些部分是成功的、哪些部分是失败的，如图 4.3 所示。其准确率曲线图如图 4.4 所示。

		测试命名	测试时间	测试数量	测试单位	测试关系	测试题目	测试问题
黄金标准	命名	3340	0	0	0	0	0	0
黄金标准	时间	0	699	75	0	0	0	0
黄金标准	数量	0	114	6810	0	0	0	0
黄金标准	单位	17	0	0	2475	0	0	0
黄金标准	关系	0	0	0	0	1384	0	0
黄金标准	题目	0	0	0	0	0	780	0
黄金标准	问题	0	0	0	0	0	0	1480

图 4.3 小学数学数据标注任务的混淆矩阵

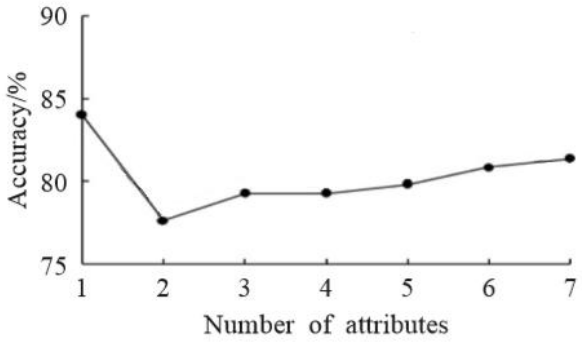

图 4.4 MRNBC-M 的分类准确率曲线

当算法准确率的评价方法比较容易计算时，最终得到的数据不一定能揭示出可能影响结果的所有错误来源。在进行机器学习训练时 TTE 循环将语料库分

为 3 个部分：开发训练、开发测试和最后测试。然而如果语料库非常大的话，用这种方式来分割语料库是非常困难的，应确保有足够多的数据来进行训练，但如果测试集太小那么即使少量的错误也会使算法看起来表现很差。这时就可以采取 k-折交叉验证分析法对算法进行评价，它允许将数据分成 k 个分区，用 k-1 折来训练算法，剩下一折用于测试[46]。然后选择另一个分区用作测试并重复这个过程，直到所有分区都曾用作测试集为止。例如，如果把语料库分成 5 个分区(k=5)，那么应该用其中的四个分区训练算法，并在第五个分区上进行测试[47]，如图 4.5 所示。

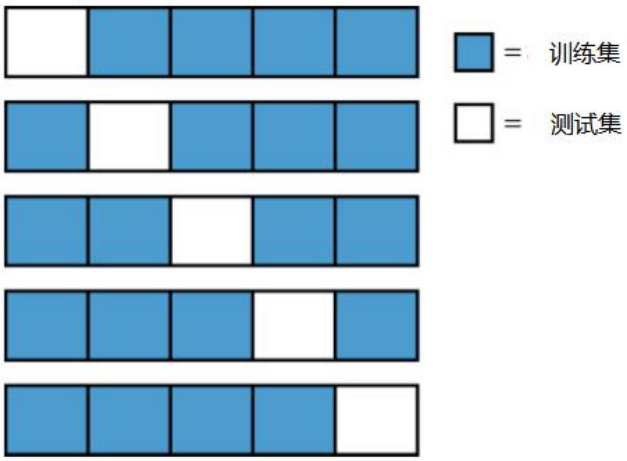

图 4.5 k 折交叉验证可视化图

4.3 修改与总结

在机器学习的训练、测试和评价阶段可能一直需要修改项目，但每一步的调整只针对当前步骤的修改，因此在此阶段本文采用倒退一步的方法考察项目中需要修改的一些“重点”条目，（即 MATTER 周期中的“R”）。在这个阶段需要考虑的范围有：

①语料库分布和内容：是否依然具有平衡性和代表性。可能在对语料进行标注时就已经多次改变过标注模型，以至于需要收集一些新的数据样本；或者在训练和测试阶段发现了一些新的特征，但是这些特征却没有在语料库中表现出来。这做这些修改时请永远记得“平衡性”问题，不要过多的表示某一特征来创建一个不平衡的语料库[48]。

②标注模型和规格说明：根据 ML 结果返回去检查标准模型和规格说明，其标注特征与标注方案是否相关联。这里并不是意味着需要完全改变标注任务来满足使算法能够有最佳的表现，而是思考是否各种不同的标签和属性都能形成各自独特的分类，是否需要将语料库中的某些数据的信息进行合并。在 MAMA 循环中，当你一直密切关注文本、标签和属性时，容易全神贯注于细节而忽略了全局，一旦到了 MATTER 循环的末端，就更容易发现标注任务中哪些

部分需要精简。

③ML 修正和训练调整：回到语料库中发现更多对训练和测试有用的语言学信息。具体来说，如果专注于算法中的标注依赖型特征，可以回去查看语料库中是否有可利用的结构依赖型特征[49]。可以考虑的方面还包括文档结构(章节标题或重要的短语)、像题材或语体等的元信息、介词短语的影响等。

目前，没有黄金标准或 ISO 列表对标注语料库应该包含哪些信息做出规定。创建黄金标准语料库并进行机器学习训练是一件困难的任务，而且因为很多变量会影响到项目的结果，所以在此阶段要保证语料库的平衡性和代表性。语料库是一种语言的选择性子集，而不是包含某种语言的所有可能使用的例子，其包含的不同类型的文本相应的比例应该与有依据的和根据直觉的判断相一致。

第 5 章　语料库的应用

本文将语料库系统进行了可视化的展示，使用了基于 Web2.0 的技术，其主要提供了对小学数学知识点查询与搜索相关试题的功能。前台界面主要提供了知识点与相关题型的模糊输入，系统根据输入进行快速处理，并向用户展示相匹配的知识点与试题清单；后台界面主要面向管理员，其提供了管理员查看、录入、删除、修改等管理的功能。

5.1　系统运行环境

系统的运行环境由服务器端和客户端两部分组成，其中服务器端包括：

①Java 开发包：JDK1.8

②Web 服务器：Tomcat8.0

③数据库：MySQL8.0

客户端：浏览器

5.2　数据库设计

本系统的数据主要分为数据库内所导入的数据和原始所收集的试题整理的 word 文档数据。数据库使用 MySQL 数据库进行数据存储方面管理，word 文档则以较好的格式排版而又图文并茂地保存题目的原始信息。

所有的知识点信息是以表结构将信息存储于 MySQL 数据库的 knowledges 表中。一个知识点对应着表中的一条记录。一条记录包含 ID、知识点描述、知识点序号、父知识点序号等元素(如表 5-1)。表的信息来源于前文所归纳的知识点结构图，这样就可以很好的将知识点信息放入数据库中进行存储，并能在 MySQL 数据库中实现对知识点信息的增加、删除、修改等管理功能。

表 5.1　knowledges 表

字段名称	数据类型	字段大小	是否为主键	是否可以为空	说明
ID	int	4	是	否	编号
course	varchar	50	否	否	科目名
kid	int	4	否	否	知识点序号
text	varchar	50	否	否	知识点描述
parentID	int	4	否	否	父知识点序号

所有的题目文本信息也是以表结构将信息存储于 MySQL 数据库的 search 表中。一道试题对应着表中的一条记录。一条试题信息记录包含 ID、知识点集合、年份、省市、题型、第几题、链接等元素(如表 5-2)。表的信息来源于对小升初数学真题的统计分析过的 EXCEL 表。这样就将题目信息很完整的录入了数据库系统，并能在 MySQL 数据库中实现题目信息的增加、删除、修改等管理功能。

表 5.2　search 表

字段名称	数据类型	字段大小	是否为主键	是否可以为空	说明
ID	int	4	是	否	编号
kids	int	50	否	否	知识点集合
year	int	4	否	是	年份
area	varchar	50	否	是	省市
type	varchar	50	否	否	题型
qnumber	int	4	否	是	第几题
qlink	varchar	200	否	否	链接

5.3　系统的功能实现

5.3.1　前台的功能实现

用户首先需要进行登录才可以进行知识点或题目的搜索，新用户则先需要进行注册(如图 5.1)。在注册时用户需要填写自己的个人信息，包括用户名、密码、性别、email 地址等，在第二遍输入密码是需要与第一遍的密码一致否则注册失败。注册成功后返回系统首页，未登录时只能查看首页的动态页面，登陆后可进入搜索系统(如图 5.2)。

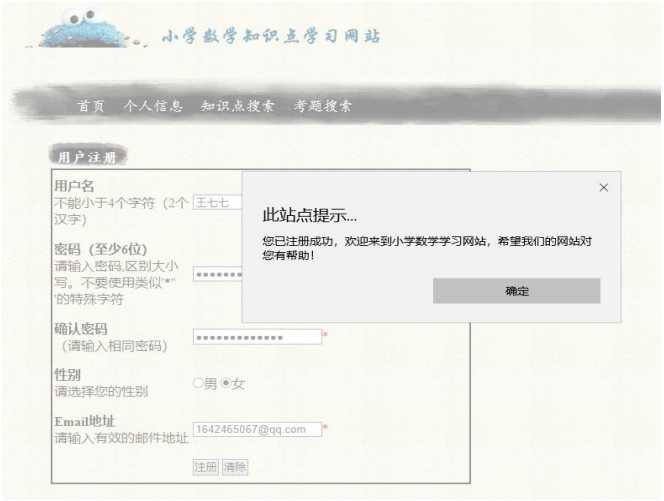

图 5.1 用户注册界面

图 5.2 用户登录界面

前台可供输入知识点信息，依据输入的信息，提供给后台进行数据匹配处理(图 5.3 给出了此步骤的可视化表示)，将部分匹配的所有知识点信息集合以列表的形式展示在当前页，并且显示符合该搜索的知识点个数(如图 5.4)。

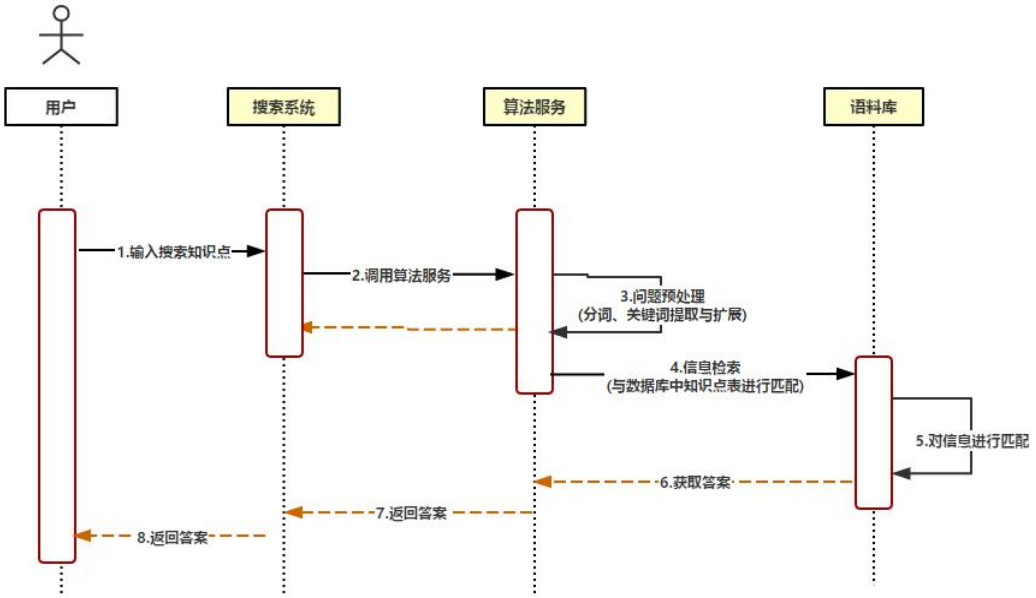

图 5.3 系统时序图

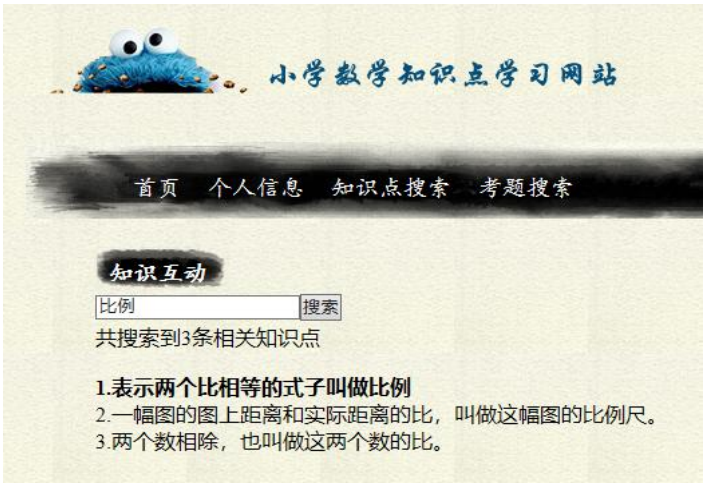

图 5.4 知识点搜索界面

前台可供输入的试题信息的文本有年份、分数、题型等等，依据输入的试题信息，传输到后台进行数据信息匹配处理，将与试题信息相匹配的所有题目信息以及知识点信息集合以列表的形式展示在当前页，并且显示符合该搜索的题目个数(如图 5.5)。考题显示部分，若点击“点击查看详情”，则可以显示出完整的题目和解题内容，例如搜索到的 2019 年辽宁省关于“鸡兔同笼”问题的题目与解题过程。

图 5.5 考题搜索界面

5.3.2 后台的功能实现

后台功能主要展示了所有知识点信息和题目信息的原始数据信息，这是与数据库其中的两张表一一对应的，并且通过后台管理功能管理员可以对知识点和题目信息进行简单的添加、删除、修改等操作(如图 5.6，图 5.7)。后台管理界面的设计分为左右两部分，左边是导航菜单，主要分为三块：知识点管理、考题管理和自动标注，右侧是以表格的形式将所有的数据展示出来。

图 5.6 知识点管理界面

图 5.7 试题管理界面

管理员的添加知识点模块相对来说比较简单，只需添加知识 ID、父序号、科目和描述等参数并进行保存即可(如图 5.8)。

图 5.8 知识点添加界面

对于添加搜索题目的模块，不仅要上传对应的 Word 文档，还要对题目进行详细标注，需要添加 ID、知识点序号、年份、省市、第几题、题型等进行保存(如图 5.9)。

图 5.9 知识点添加界面

在自动标注的界面中，管理员可以通过输入三个信息后实现对试题的自动标注，分别是标题描述、运算关系描述和问题描述。小学数学问题的标准 XML 文件是由这三个模块组成，其中标题和问题的描述模块是必不可少的，运算关系模块则视具体题型而定，无运算关系的就可以不写(默认值为 null)。管理员可以从语料库或备用试题库中选择其中一个小学数学的考题，进行对以上三个模块进行描述，信息输入完成后单击“生成 XML 文件”，于是就可以生成以便计算机识别的 XML 文件，并提供了管理员的下载和保存的功能(如图 5.10)。

图 5.10 自动标注界面

5.4 解题规则的实现

通过自动标注将试题信息转化为计算机可识别的 XML 语言，但是计算机又该如何处理不同的题型，这里就涉及到解题规则的构建。无论是数与代数、空间与图形还是统计与概率类题型，想要让计算机像人类大脑一样进行解题，首先需要从题意中获取相关的试题信息，然后通过具体的规则来达到解题的目的。前文已经介绍了这些题型具体的题意描述，因此计算机可以从 XML 文件中获取必要的试题信息，根据解题规则实现机器求解的目的。

计算机根据 XML 文件中对试题信息进行分析，将得到的主要考察知识点作为依据，根据题型问法将知识点的定义与相关推导公式相结合以制定相应的解题规则，这些规则将以 Python 语言存储在计算机内。Python 语言在教育、科研、工业领域的应用都十分广泛，其不仅提高了软件的质量和生产效率，还具有可维护性，因此在世界各地都备受欢迎。它的文法和语义简单易懂，并且有强大的字符串处理功能，其自带的函数非常适合处理语言数据。Python 还自带了强大的标准库，包括数值处理、网络连接和图形编程等组件。作为解释性语言，Python 便于交互式编程；作为面向对象的编程语言，Python 蕴蓄的数据和方法被封装和重用以便用户调用；作为动态语言，在程序运行时 Python 蕴蓄的属性可以添加到对象，允许变量进行自动类型转换，以提高开发效率。

在使用 Python 对试题信息进行提取时，首先使用句子分割器，将试题的原始文本分割成句，使用分词器将每个句子进一步细分为词。接下来对每个句子进行标注，在下一步命名实体识别中将证明这是非常有益的。在这一步，寻找每个句子中提到的潜在的实体。最后使用关系识别搜索文本中不同实体间的可能关系。如图：5.11 所示

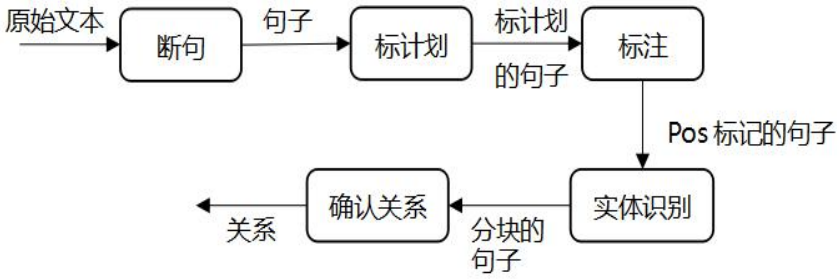

图 5.11 信息提取系统的简单流程

以一道陈述题为例，例：在亚特兰大经营的组织有乔治太平洋公司，在纽约经营的组织有宏盟集团，即在纽约又在亚特兰大经营的组织有天高集团，哪些组织在亚特兰大经营？该系统将题目文档的原始文本作为输入，将生成的(entity,relation,entity)元组链表作为输出。即生成元组([ORG:'Georgia-Pacific'] in [LOC:'Atlanta']) 、 ([ORG:'BBDO South'] in [LOC:'Atlanta']) 、 ([ORG:'BBDO South'] in [LOC:'New York'])、([ORG:'Omnicom'] in [LOC:'New York'])。那么这个问题：“哪些组织在亚特兰大经营？”可翻译如下。

```
>>> print [org for (e1, rel, e2) if rel == 'IN' and e2 == 'Atlanta']
['BBDO South', 'Georgia-Pacific]
```

在小学数学考题中最常见的题型就是解方程类的题型这种题型只含有一个未知数，解方程过程比较简单：首先弄清题意找出未知数并用 x 表示，接下来找出题中数量间的相等关系后列出等式方程，然后根据等式的特性，求出等式中未知数的值，最后将未知数的值代入原方程进行检验。例：今年爸爸的年龄正好是儿子的 3 倍，3 年前爸爸比儿子大 22 岁，儿子今年多少岁？

分析：由于二人的年龄差永远不变，可得出等量关系：今年爸爸的年龄-今年儿子的年龄=他们的年龄差。

解：设儿子今年 x 岁，根据题意得

```
3x-x=22
(3-1)x=22
2x=22
x=11
```

那么这个问题：“儿子今年多少岁？”机器可进行计算如下。

```
>>> 3x-x=22
>>> print ("x ="+x)
x=11
```

这个例说明了前文所制定的解题规则是可以通过 Python 语言实现解答问题的，充分地说明了所构建的规则具有实用性和有效性。

第 6 章 总结与展望

6.1 研究总结

本文主要研究了面向小学数学学习者语料库的理论和技术，解决了构建小学数学学习者语料库的关键问题，在此基础上，具体设计并实现了该语料库系统的功能，本文完成的具体工作如下：

(1). 小学数学题型分类及其表征研究：针对这些没有经过处理的语料，首先根据数与代数、空间与图形和统计与概率的知识点描述，构建了具体的知识体系并为其建立知识点索引表，对每道试题的知识点进行标注，然后统计和分析试题中相关知识点出现的频率，最后将所有知识点根据其难易程度和相关题型所占的比例进行综合分析，选择最具有代表性的题型进行表征研究。

(2). 面向小学数学学习者语料库的构建：根据 MATTER 循环方法首先进行的是“建模-标注”循环，首先对根据小学数学知识点创建模型和规格说明，按照规格说明对试题进行标注，在进行标注时不断地对模型和规格说明进行改进与简化，从而达到黄金标准。然后进行的是“训练-测试”循环，先将算法在训练集中进行学习，学习结束后就可以到开发-测试集上运行算法，在这个过程中如果算法没有准确地对语料进行标注，就调整并重新训练算法然后再重新进行测试，不间断地重复以上过程直至算法获得满意的训练结果。训练完成后算法在最先预留好的测试语料库上进行测试，再根据这个结果不断地改进语料库的模型与算法，以实现对小学数学语料的自动标注。

(3). 面向小学数学学习者语料库的应用。本文使用基于 Web2.0 的技术对语料库系统进行可视化的展示，其主要提供知识点查询与搜索相关试题的功能。前台界面主要提供了知识点与相关题型的模糊输入，系统根据输入进行快速处理，并向用户展示相匹配的知识点与试题清单；后台界面主要面向管理员，其提供了管理员查看、录入、删除、修改等管理的功能。按照此前分类的知识体系，选取了考察所占比例较多的知识点进行研究分析，比如四则混合运算题型、简易方程题型、比的运用题型、图形的面积或体积题型、逻辑推理题型、统计与概率题型等。根据知识点的定义，将题型问法与相关公式相结合以用来归纳该类题型的解题规则。这些规则将通过 Python 语言进行标注和存储，使用这些规则进行“存题”—“识题”—“解题”的机器求解过程，并举例说明 Python 的解题效果，从而说明构建的语料库拥有实用性和有效性。

6.2 问题与工作展望

本文对面向小学数学学习者的语料库展开了初步的研究和设计实现，但是由于时间和能力有限，尚存在以下问题：

(1). 本文所统计的数据是 2013-2019 年辽宁省、山东省和河北省的小升初考题以及复习资料中的知识点巩固习题，相对于其他的语料库，级别还远远达不到要求；

(2). 对于小学数学考试真题的 XML 表征目前只完成了数与代数、空间与图形和统计与概率类有关计算题型的研究，没有对空间与图形的图形题进行研究，且仅仅制定了相应的解题规则，没有进行深入的探讨；

(3). 语料库系统实现功能较为简单，没有记录输入过词汇的记录，且无热门搜索推荐。

针对以上问题，后续开展如下工作：

(1). 扩大语料的来源，统计 2013-2019 年的河南省、安徽省、浙江省的小升初小学数学考题内容，以此作为研究对象，并添加到生语料库进行标注与测试算法；

(2). 研究图形的特征，完成对图形题的知识点标注，并实现自动标注；

(3). 完善语料库系统的界面及后台管理界面，优化题目存储管理，进一步实现增加图片及其链接的存储。

参考文献

[1]Judith Holler,Stephen C. Levinson. Multimodal Language Processing in Human Communication[J]. Trends in Cognitive Sciences, 2019, 23(8):38-40.

[2]王晶. 基于深度学习的文本表示和分类研究[D].北京：北京邮电大学,2019.

[3]廖若飞,廖海.一种基于 NLP 的机器人查询系统[J].电脑知识与技术,2018,14(21):97-98.

[4]Edoardo Maria Ponti,Helen O'Horan,Yevgeni Berzak,Ivan Vulić,Roi Reichart,Thierry Poibeau,Ekaterina Shutova,Anna Korhonen. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing[J]. Computational Linguistics, 2019, 45(3)33-35.

[5]郭红梅,梁媛元.基于 NLP 的英语口语教学模式探究[J].教育现代化,2019,6(72):56-59.

[6]Hang Li.Deep learning for natural language processing:advantages and challenges[J].National Science Review, 2018, 5(01):24-26.

[7]赵光亮,令狐雨薇,朱德孙等.基于 Python 的通用论坛正文提取研究[J].电脑知识与技术,2018,14(24):259-260.

[8]武银平.现代信息技术在小学高年级数学教学中的运用[J].科技风,2020(08):81.

[9]余杰.关于小学数学教学中培养学生独立思考能力问题的研究[J].学周刊,2020(09):29-30.

[10]马玉山.基于核心素养视角开展小学数学教学[J].学周刊,2020(09):91-92.

[11]李红.小学生数学学习能力培养探究[J].学周刊,2020(08):39-40.

[12]汪会荣.基于信息技术的小学数学教学创新研究[J].学周刊,2020(08):53-54.

[13]吴骧.数学语言教学和小学数学教学[J].课程教育研究,2020(05):160.

[14]余根钦,何琳,索磊.美国小学阶段数学课程标准及其启示[J].教学与管理,2019(35):56-58.

[15]胡瑭. 中美小学数学教材数与运算内容的比较研究[D].江西：江西师范大学,2019.

[16]吴冰冰. 数学史融入小学数学教学的实施策略及评估研究[D].厦门：集美大学,2019.

[17]卞成德.基于语料库的数学定义研究[D].厦门：厦门大学,2009.

[18]郑泽之.数学教材语言与语料库建设[C].国家语言资源监测与研究中心教育教材语言分中心、人民教育出版社:福建省语言学会,2008:4.

[19]钱永红,陈新仁.语料库方法在语用学研究中的运用[J].外语教学理论与实践,2014(02):15-20+26+94.

[20]陈菁,符荣波.国内外语料库口译研究进展(1998-2012)——一项基于相关文献的计量分析[J].中国翻译,2014,35(01):36-42+126.

[21]李文中.语料库标记与标注:以中国英语语料库为例[J].外语教学与研究,2012,44(03):336-345+478.

[22]周立君,汪涛.亚马逊土耳其机器人:科学研究的众包网络平台研究综述[J].科技进步与对策,2014,31(08):156-160.

[23]梁志剑,谢红宇,安卫钢.基于 BiGRU 和贝叶斯分类器的文本分类[J].计算机工程与设计,2020,41(02):381-385.

[24]张朝辉,刘佳佳,冉惠.基于贝叶斯与神经网混合算法的电商信用评价方法研究[J].情报科学,2020,38(02):81-87.

[25]徐光美,刘宏哲,张敬尊,王金华.用平滑方法改进多关系朴素贝叶斯分类[J].计算机工程与应用,2017,53(05):69-72.

[26]王仁强,陈和敏.基于语料库的动词与构式关系研究——以 sneeze 及物动词用法的规约化为例[J].外语教学与研究,2014,46(01):19-31+158.

[27]Gaétan Poelman,Saeid Hedayatrasa,Joost Segers,Wim Van Paepegem,Mathias Kersemans. Multi-Scale Gapped Smoothing Algorithm for Robust Baseline-free Damage Detection in Optical Infrared Thermography[J]. NDT and E International, 2020(02):88-89.

[28]Jixiang Yang,Dingwei Li,Congcong Ye,Han Ding. An analytical C 3 continuous tool path corner smoothing algorithm for 6R robot manipulator[J]. Robotics and Computer-Integrated Manufacturing, 2020(03):64.

[29]常志鹏,徐娟.基于朴素贝叶斯算法的网络教学平台响应时间研究[J].数字技术与应用,2019,37(12):112-115.

[30]仁青吉.藏语 N-gram 语言模型中的平滑技术研究[J].西北民族大学学报(自然科学版),2019,40(04):26-30.

[31]何伟. 基于朴素贝叶斯的文本分类算法研究[D].南京：南京邮电大学,2018.

[32]韩素青,成慧雯,王宝丽.三支决策朴素贝叶斯增量学习算法研究[J/OL].计算机工程与应用:1-16[2020-03-16].http://kns.cnki.net/kcms/detail/11.2127.TP.20191122.1516.008.html.

[33]周钢,郭福亮,崔良中,李永杰,郭晖.基于 Python 的大学计算机基础课程实践教学改革[J].计算机教育,2020(02):96-100.

[34]韩文煜.基于 python 数据分析技术的数据整理与分析研究[J].科技创新与应用,2020(04):157-158.

[35]孙晓天.近年来我国中小学数学教材建设述要[J].数学教育学报,2008.8,(4).

[36]彭冬生.数学学术文献自然语言处理中的若干问题[D].吉林：吉林大学,2018.

[37]郑重.面向初等数学概率与统计语料库的构建研究[D].武汉：华中师范大学, 2016.

[38]尹梅. 中日小学 4-6 年级数学教科书内容比较研究[D].内蒙古：内蒙古师范大学,2019.

[39]王志玲. 小学六年级学生数学交流推理能力教学研究[D].上海：华东师范大学,2019.

[40]曲聪. 中日小学六年级学生数学能力与学习品质的比较研究[D].山西：山西医科大学,2016

[41]范国凤,刘璟,姚绍文,栾桂凯.基于语义依存分析的图网络文本分类模型[J/OL].计算机应用研究:1-5[2020-03-16].https://doi.org/10.19734/j.issn.1001-3695.2019.08.0522.

[42]黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[J].软件工程,2020,23(03):1-3.

[43]Inequalities and Applications; Findings from Taiyuan Normal University Has Provided New Data on Inequalities and Applications (A semi-smoothing augmented Lagrange multiplier algorithm for low-rank Toeplitz matrix completion)[J]. Journal of Mathematics, 2020(02):31-32.

[44]薛金成,姜迪,吴建德.基于 word2vec 的专利文本自动分类研究[J].信息技术,2020,44(02):73-77.

[45]陈清.基于 Python 的网站爬虫应用研究[J].通讯世界,2020,27(01):202-203.

[46]韦专.基于 Python 语言的高中人工智能课程教学分析[J].贵州教育,2020(02):28-31.

[47]杜兰,刘智,陈琳琳.基于 Python 的文献检索系统设计与实现[J].软件,2020,41(01):55-59.

[48]唐琳,何天宇.基于 Python 的自然语言数据处理系统的设计与实现[J].电子技术与软件工程,2018(16):160-162.

[49]吴爽.基于 python 语言的 web 数据挖掘与分析研究[J].电脑知识与技术,2018(27):1-2.

致　谢

四年的本科生学习生活和三年的研究生学习生活，转眼间七年的时光飞逝。沈阳师范大学，这所如印记刻在我身上的学校，值得我一生所回忆的美丽校园陪伴我度过了最美好的青春岁月。在沈阳师范大学所结识的良师益友和所学到的专业知识都是我一生中最宝贵的财富。现在，即将毕业之际，回顾过去，感慨万千。

感谢我的导师宋波教授，感谢您对我学习生活和生活上无私的帮助和无微不至的关怀。在毕业论文的撰写中，从论文的选题、框架构造、问题探讨到论文完成的各个阶段，宋老师都给予了我精心的指导。学术上宋老师严谨的科学态度和治学精神都深深的感染并激励着我。很多时候，我晚上向宋老师请教学术问题时，宋老师都会即时地为我解惑，真心希望老师不要太辛苦，要多注意自己的身体。生活上，宋老师既严谨又和蔼，教会了我很多为人处世的道理。在我犯错误时您会严厉的批评，但当我有困难时又能设身处地为我着想、尽可能地帮助我。在此，向宋老师表示诚挚的谢意和最中忠心的祝福。祝您身体健康，一切顺利。

感谢软件学院的各位领导和老师，感谢您们对我的教导。在您们的帮助下，我学习到了很多的知识，也明白了很多做人的道理。您们是我人生中的榜样，我会谨记您们对我的尊尊教诲。

感谢我身边的同学们，我们一起在沈阳师范大学度过了快乐的大学生活。前世的五百次回魔才换来今生的擦肩而过，我会谨记和珍惜这份难得的缘分，祝福你们以后工作、生活等各方面都顺利。希望我们无论天涯海角，都不忘彼此。

感谢我的父母、男朋友和所有关心我的亲人。在我学习生涯中，给我无微不至的关爱和支持，包容我的任性和莽撞，使我有更加坚定的信念面对和克服困难，更勇敢更积极地面对人生中所遇到的困难与挫折。

最后感谢百忙中抽出宝贵时间参加论文评审及答辩的各位专家、教授，向您们致以最诚挚的敬意。

个人简历

王蕊拂，女，1995 年 9 月 3 日出生于辽宁省沈阳市。

2017 年 9 月至今，在沈阳师范大学软件学院攻读计算机应用技术专业研究生。

2013 年 9 月至 2017 年 6 月，于？？学习，获？？？学士学位。

在学期间发表的学术论文及研究成果

1. 辽宁省高等学校基本科研项目“基于虚拟学习社区的辽宁省中小学教师研训体系的构建”（项目编号：2017L317）。

2. 辽宁省教育科学十三五规划课题——面向高校招生的智能问答系统的研究与实现(项目编号：JG18DB451)。

3. Rui-fu Wang,Bo Song,Xiao-mei Li.The Research on the Construction of Primary Mathematics Corpus Based on MATTER Cycle Method.International Journal of Information and Education Technology. vol. 10, no.4, 2020. EI.