# 基于篇章结构的英文作文自动评分方法

周 明[1,3] 贾艳明[2] 周彩兰[1] 徐 宁[1,3]
(武汉理工大学计算机科学与技术学院 武汉 430070)[1]
(北京博智天下信息技术有限公司人工智能与大数据研究中心 北京 100085)[2]
(武汉理工大学交通物联网技术湖北省重点实验室 武汉 430070)[3]

**摘 要** 作文自动评分(Automated Essay Scoring AES)是指使用统计学、自然语言处理及语言学等领域的技术对作文进行评价和评分的系统。篇章结构分析是自然语言处理领域的一个重要研究方向,也是作文自动评分系统的重要组成部分之一。目前国外的作文自动评分系统虽有广泛应用,但对篇章结构评分的研究还存在不足,且对中国学生英语作文的针对性不强;国内对英语作文自动评分的研究处于起步阶段,忽视了篇章结构对英语作文评分的重要性。针对这些问题,提出一种基于篇章结构的英文作文自动评分方法,在词、句、段落3个层面上提取作文的词汇、句法以及结构等特征,并使用支持向量机、随机森林以及极端梯度上升等算法对篇章成分进行分类,最后构建线性回归模型对作文的篇章结构进行评分。实验结果表明,基于随机森林的篇章成分识别模型(Discourse Element Identification based Random Forest,DEI-RF)的准确率为94.13%;基于线性回归的篇章结构自动评分模型(Discourse Structures Scoring based Linear Regression,DSS-LR)在背景介绍段(Introduction)、论证段(Argumentation)以及让步段(Concession)的均方差可达到0.02,0.11和0.08。

**关键词** 作文自动评分,篇章成分,篇章结构分析,自然语言处理,随机森林,线性回归
**中图法分类号** TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.035

## English Automated Essay Scoring Methods Based on Discourse Structure

ZHOU Ming[1,3] JIA Yan-ming[2] ZHOU Cai-lan[1] XU Ning[1,3]
(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China)[1]
(Research Center for Artificial Intelligence and Big Data, Global Wisdom Inc, Beijing 100085, China)[2]
(Hubei Key Laboratory of Transportation Internet of Things, Wuhan University of Technology, Wuhan 430070, China)[3]

**Abstract** Automated essay scoring is defined as the computer technology that evaluates and scores the composition, based on the technologies of statistics, natural language processing, linguistics and some other fields. Discourse structure analysis is not only an important research field of natural language processing, but also an important component of the AES system. Nowadays, AES system has widely application. However, there is not enough research on the structure of the essay, and the AES system does not focus on the Chinese students. The domestic researches on the AES are in infancy, ignoring the importance of discourse structure in essay scoring. In view of these problems, this paper proposed a method of automated essay scoring based on discourse structure. Firstly, the method extracts essay's features, such as vocabulary, lexical and discourse structure from levels of words, sentences and paragraphs. Then, the composition of essays is classified by support vector machines, random forests and extreme gradient boosting, and then the linear regression model with the discourse element is constructed to score the compositions. The experimental results show that the accuracy of discourse element identification based random forest (DEI-RF) can reach 94.13%, and the mean squared error of automated discourse structure scoring based on linear regression (DSS-LR) model can reach 0.02, 0.11 and 0.08 on introduction, argumentation and concession respectively.

**Keywords** Automated essay scoring, Discourse element, Discourse structure analysis, Natural language processing, Random forest, Linear regression

到稿日期:2018-01-24 返修日期:2018-05-13

**周 明**(1993-),男,硕士生,主要研究方向为自然语言处理、篇章分析,E-mail:zhoum1118@163.com;**贾艳明**(1980-),男,博士,高级工程师,主要研究方向为机器学习、算法优化;**周彩兰**(1964-),女,硕士,副教授,主要研究方向为深度学习、图像处理,E-mail:383277764@qq.com(通信作者);**徐 宁**(1968-),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为超大规模集成电路的计算机辅助设计系统、计算机体系结构、数据挖掘和算法优化。

## 1　引言

作文自动评分是指使用特定评分模型对作文内容、结构及语言等进行评分的系统,常在大型考试中应用。近几十年来,随着计算机软硬件和自然语言处理技术的发展,国内外相继出现了 AES 系统。AES 不但节省了大量人力、物力,且评分更加客观,但 AES 的研究大多停留在内容和语言方面,对篇章结构(Discourse Structures,DS)的分析研究存在不足。在高水平英语考试(如托福、雅思、GRE 等)中,作文评测更加注重学生的结构表达,尤其是雅思考试,文章的结构表达在作文评分方面所占比重较大,因此篇章结构分析逐渐成为 AES 系统中不可或缺的部分。

高水平英语考试的作文类型一般为议论文,议论文要求作者针对某个特定的主题陈述自己的观点和主张,并以严密的逻辑和合理的论证来确定其观点正确与否。篇章成分表示篇章单元对文章组织结构的贡献,如何正确区分篇章成分是分析议论文篇章结构的首要问题。Stab 等[1-2] 主要关注文章的论证过程,将篇章成分大致分为 3 类:主论点、分论点和论据。Song 等[3] 分析了中国高中学生撰写的 3 种主题共 300 篇议论文体裁作文,以句子为单位划分篇章单元,对文章中的每个句子进行篇章成分分类,共划分为 7 种类型,包括背景介绍、重述或总结提示、观点综述、主要理由、论据、总结以及无关句。Burstein 等[4] 分析了 6 种主题共 250 篇议论文体裁文章,这些文章包括高中三年级和大学一年级学生的作文;他们将篇章成分类别划分为标题、背景介绍、观点综述、主要理由、论据、总结以及无关句。对于篇章成分类别,我们分析了大量托福、雅思以及 GRE 的 Argument 官方范文和应考学生的文章,依托真实托福应考学生的 300 篇议论文,总结出中国学生撰写议论文的特点;并结合议论文的 5 个要素,即论点、理由、论据、结论和论证,将篇章成分划分为 9 个细粒度类别标签。表 1 列出了篇章成分及其定义。本文以句子为单位划分篇章单元,为每个篇章单元标注对应篇章成分。例 1 为一篇议论文作文的篇章成分标注示例。

表 1　篇章成分的定义

Table 1　Definition of discourse element

| 篇章成分 | 定义 |
|---|---|
| Background Introduction (BI) | 文章背景介绍 |
| Main Claim(MC) | 主要论点,描述了作者对某一问题的主要观点 |
| Reason(RS) | 理由,支持作者的主要论点,并与其构成因果关系 |
| Reason Interpretation (RI) | 理由诠释,理由的补充说明 |
| Evidence(ED) | 论证理由是否充分,与理由构成因果关系 |
| Concession(CS) | 让步,重申文章论点的同时承认反方观点的部分优点,并指出反方观点的局限性,以此来消除论证过程的片面性 |
| Conclusion(CC) | 总结,论证过程的自然结果,用于总结全文的核心思想,可能是理由的总结,也可能是观点的重申 |
| Transition(TST) | 过渡,用于衔接前后两个篇章单元 |
| Irrelevant(IRL) | 与文章无关的句子,对论证过程不构成有意义的贡献 |

例 1　议论文作文的篇章成分标注

〈*BI*〉Friendship is essential for everyone, but different people have different principals to make friends.〈/*BI*〉〈*MC*〉From my perspective, I will not keep on the friendship when an old friend does some things I dislike.〈/*MC*〉〈*TST*〉In what follows, I will offer reasons and examples to demonstrate my argument.〈/*TST*〉

〈*RS*〉The friendship is unhealthy that one ought to be tolerant of the other.〈/*RS*〉〈*RI*〉If you do not prepare to finish the relationship in which your old friend does something you do not like, you need give tolerance to him or her.〈/*RI*〉〈*RI*〉On the contrary, the condition should be enjoyable and be relaxing with each other in genuine friendship.〈/*RI*〉〈*RI*〉Otherwise the accumulation of tolerance will cause more severe contradiction which is harmful for the both sides.〈/*RI*〉〈*ED*〉For instance, Linda dislike a stranger to visit her home, whereas her roommate takes different men to their house constantly, but Linda chooses to be patient to keep friendship even though she feels unpleasant every time.〈/*ED*〉〈*ED*〉The grievance of tolerance affects not only their relationship but also her own work.〈*ED*〉

〈*CS*〉Admittedly, stopping the friendship decidedly may manifest you are not generous and comprehensive enough.〈/*CS*〉〈*CS*〉However, the empathy built on mutual understanding and positive status on the basis of enjoyment of friendship are the most crucial parts of the relationship.〈/*CS*〉

〈*CC*〉In a nutshell, the people who recognize with you for a long time do something you do not like are the people who do not attach great importance to you.〈/*CC*〉

对篇章结构的分析是自然语言处理中的核心问题之一。E-rater[5] 是由美国教育考试服务中心(Educational Testing Service,ETS)开发的作文自动评分系统,在作文自动评分过程中抽取篇章特征来识别 Burstein 等提出的篇章成分,并通过计算不同篇章成分的数量来完成对作文的篇章结构评分。E-rater 已被广泛应用于 ETS 的托福作文评测中,但其在篇章结构评分中仅考虑了文章的篇章结构特征,研究深度不够,准确度不高,而且由于 E-rater 是商业化产品,其篇章结构评分模型的具体内容对外保密,无法了解其篇章结构评分模型的具体实现细节。在不同的社会文化背景下,学生思考问题的方式的不同会导致写作风格有所差异[6],因此国外的 E-rater 系统对中国学生作文的评价针对性不强。随着国内在线教育的兴起,国内 AES 系统也在逐步发展。句酷批改网在国内大学英语作文评分中的应用较为广泛,但其对篇章结构的评分效果有限,且无法对文章的整体衔接性和连贯性进行评阅,还需要不断改进。

基于已有的研究工作,针对传统研究方法的不足,结合近年来自然语言处理技术对作文篇章结构分析的最新成果,本文提出了一种基于细粒度篇章结构的作文自动评分方法。本文的主要贡献如下:1)分析了从国内知名在线出国留学平台朗播网采集的真实托福应考学生的 300 篇议论文,共 6083 个句子,以句子为单位划分篇章单元,并为每个篇章单元标注篇章成分,且以段为单位对作文的篇章结构进行评分,以此自建语料库;2)国外 AES 系统一般仅针对以英语为母语的学生作

文进行评测，本文方法考虑到中国学生的写作特点，将其应用于指导实验中的特征工程，在一定程度上提高了模型的适用性；3）国外仅对篇章结构进行研究，鲜有依据篇章结构对作文进行评分的研究，国外的 AES 系统中也仅有 E-rater 系统在评分时考虑了篇章结构，而国内的篇章结构分析主要集中在篇章关系的识别上，很少对细粒度的篇章成分进行识别，也没有依据篇章结构对作文进行评分。本文通过篇章成分的识别和专家评分的拟合实现了作文篇章结构的自动评分。作文自动评分方法主要分为两个模块：1）篇章成分自动识别（Discourse Element Identification，DEI），这个模块提取篇章的结构、词汇及句法等特征，并基于支持向量机（Support Vector Machine，SVM）[7]、随机森林（Random Forest，RF）[8]和极端梯度上升（Extreme Gradient Boosting，XGBoost）[9]等算法构建 3 种篇章成分自动识别模型 DEI-SVM（Discourse Element Identification based Support Vector Machine），DEI-RF（Discourses Element Identification based Random Forest），DEI-XGB（Discourse Element Identification based Extreme Gradient Boosting），然后分别使用这 3 种模型来识别不同类型的篇章成分，如表 1 所列；2）篇章结构自动评分（Discourse Structures Scoring，DSS），该模块利用第一个模块识别出的篇章成分构建基于线性回归的篇章结构自动评分模型，以评测文章篇章结构的优劣。

本文第 2 节介绍了篇章成分识别及篇章结构自动评分的相关工作；第 3 节详细说明了本文提出的篇章成分及其特征提取方法，并阐述了在篇章成分识别的基础上对篇章结构的评分方法；第 4 节给出了实验和结果分析；最后总结全文并展望未来。

## 2 相关工作

当前对篇章成分自动识别的研究主要可以分为两个子任务：篇章单元分割和篇章结构的识别。篇章单元分割方法的研究相对成熟且分割准确率较高，而篇章结构的识别是篇章结构分析的重点和难点。Song 等[3]通过探索整篇文章中篇章单元之间关系的内聚性来分析文章的篇章结构，将篇章结构的识别作为一个分类问题来看待，使用 SVM 模型和基于线性链的条件随机场模型（Conditional Random Fields，CRF）来对句子的篇章成分进行分类。修辞结构理论（Rhetorical Structure Theory，RST）是篇章结构分析中的重要理论之一，文本可以将其转化成修辞结构树进行分析[10]，修辞结构树如图 1 所示。修辞结构树的叶子节点对应基本篇章单元，通常篇章单元具有主次之分，包含主要内容的单元称为核（nucleus），包含次要内容的单元称为卫星（satellite）。非叶子节点对应连续的文本跨度，代表不同篇章单元的修辞关系。Burstein 等[4]使用 RST 构建修辞结构树，提取文章中每个篇章单元的修辞结构特征，结合词汇、结构等与篇章成分相关的特征，使用 C5.0 方法来对篇章单元所属的篇章成分进行分类。Duverle 等[11]使用 RST 重点研究篇章单元之间修辞关系的判定，通过提取更加丰富的篇章结构特征，采用传统机器学习方法 SVM 来构建篇章结构分析模型，最终提高了篇章结构识别的性能。Feng 等[12]采用贪婪策略的自下而上的方法，结合句子级和文本级两个篇章结构的分析层次，使用两个线性链 CRF 模型将篇章结构分析的准确度提升至58.2%。

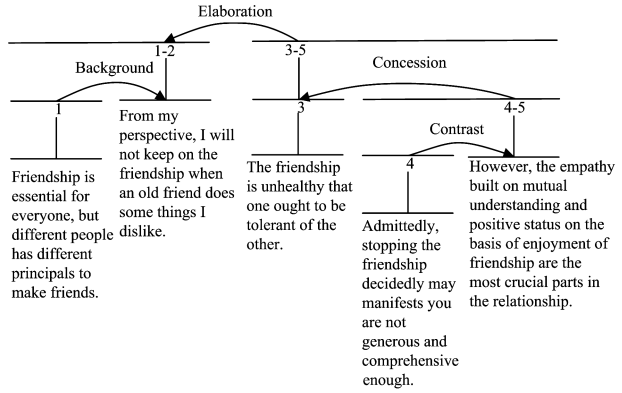


图 1 修辞结构树

Fig. 1 Rhetorical structure tree

国内篇章结构分析领域的研究主要集中在篇章关系的识别方面，有效识别篇章关系有利于实现篇章文本的结构化[13]。随着 RSTDT，PDTB 以及 HIT-CDTB 等语料库的发布，篇章关系的研究迎来了更多的机会与挑战。篇章关系分析主要包括篇章语义关系的识别、基于 RST 的篇章结构及其修辞关系的分析。篇章关系主要分为显式和隐式两种。李生等[14]在 PDTB 语料之上提出了一个基于 CRF 模型的显式篇章分析平台，其包含连接词的识别、显式篇章关系的分类及关系论元的提取，提升了显式篇章关系分类的准确度。显式篇章关系通常有连接词或特定的指示词，更容易识别，但显式篇章关系在整个篇章关系中的占比较小，因此篇章关系主要的研究难点在于识别隐式篇章关系。徐凡等[15]探索篇章中的浅层语义特征以及句子的情感特征，提出了一种基于树核的隐式篇章关系的识别方法，提高了 PDTB 中四大类别（temporal，contingency，comparison 和 expansion）的隐式篇章关系的识别准确度。蒋玉茹等[16]自建语料库，利用话题句在篇章中的位置、语法以及邻接性等细粒度特征，结合穷举策略指导候选话题句的生成过程，减少了候选话题句的数量，尽管提升了系统效率，但整体话题句识别的准确率仍然偏低。

回顾 AES 系统的发展，其有着大量基于文章内容和语法的研究，但鲜有基于篇章结构的研究。篇章结构是作文评分的重要方面，表示作者对文章组织结构以及论证逻辑的掌握。在篇章结构的研究中，国内主要分析篇章的主次关系，对篇章成分的分析研究不足，而国外的研究对中国学生作文的针对性不强。本文自建真实应考学生的托福议论文体裁语料库，主要侧重篇章成分分析，并在此基础上对作文的篇章结构进行评测。本文中假设文章可以被分割成篇章类别序列，且以句子为分割单位，一个句子即为一个篇章单元，使用特征提取程序提取与篇章成分相关的特征。与其他方法不同的是，本文分别使用 SVM，RF 和 XGBoost 算法构建篇章成分分类模型，同时使用线性回归算法来评测作文篇章结构的表现。

## 3 篇章结构自动评分方法

通过对 300 篇真实托福应考学生的议论文进行分析，将每个独立的篇章单元与表 1 中的标签关联，这些标签的设置

符合中国人撰写英文议论文的特点。本节首先针对语料库中的所有篇章单元给出正确的篇章成分类别,并且评测文章中每个段落的篇章结构分数,然后提取篇章成分的词汇、句法及结构等特征,最后分别构建 DEI-SVM,DEI-RF 和 DEI-XGB 篇章成分自动识别模型与 DSS-LR 篇章结构评分模型,从而实现篇章成分的识别及篇章结构的自动评分。

### 3.1　篇章成分分析

为了保证篇章成分识别的客观性与一致性,邀请 3 位专家参与篇章成分标注工作,其中 2 位专家按照表 1 所定义的篇章成分标签独立对 300 篇文章的 6083 个句子进行标注;若 2 位专家对同一个句子标注了不同的标签,则引入第三位专家,让其在事先不知道前两位专家的标注信息的情况下对该句子进行重新标注,如果第三位专家的标注结果与前两位专家中的某一位的标注结果一致,则该句子的最终标注为那个相同的标注结果,否则由 3 位专家仔细讨论后再给出该句子最终的标注。表 2 列出了语料库的基本统计信息以及篇章成分的分布情况。

表 2　语料库的基本统计信息

Table 2　Basic statistical information of corpus

| 统计对象 | 统计信息 |
|---|---|
| 文章 | 300 |
| 句子 | 6083 |
| BI | 670 |
| MC | 285 |
| RS | 564 |
| RI | 1232 |
| ED | 1579 |
| CS | 778 |
| CC | 706 |
| TST | 151 |
| IRT | 118 |

### 3.2　特征提取

特征提取是机器学习中至关重要的步骤,直接影响模型的性能。特征提取是将机器学习算法无法直接识别的原始数据转化为可识别的特征数据的过程。本文首先使用特征提取程序提取语料库中每个句子与篇章结构相关的特征,其次利用特征整合程序将提取的特征整合成特征向量,然后将每个句子的特征向量与其对应的标签向量作为模型输入,最后识别每个句子的篇章成分类别。本文随机地将数据分为 78% 的训练集和 22%的具有相同类别分布的测试集,并且使用十折交叉验证来确定性能最好的模型。本文提取的特征由以下内容组成。

#### 3.2.1　结构特征

结构特征主要提取的是篇章单元的词统计特征、位置特征以及标点符号特征。Biran 等[17]发现论据类篇章单元的平均长度比其他类别的篇章单元更长,因此本文将篇章单元中单词的数量作为词统计特征加入到特征集中。另外,还考虑了篇章单元的位置特征以及标点符号特征。篇章单元位置特征由两部分组成:篇章单元所在段落在文章中的位置和篇章单元在段落中的位置。标点符号分为句号、问号及感叹号,本文把篇章单元的结束标点符号转化为数值特征存入特征集中。

对于复杂的篇章结构,若只分析单个篇章单元的结构特征将很难正确识别其篇章成分标签,特别是在具有隐含关系的推理过程中,因此上下文特征的引入在篇章成分识别中发挥了重要作用。例如,在背景介绍后作者通常会发表自己的观点,陈述理由之后往往也会有论据分析。为了进一步探索篇章单元的上下文特征,依据评分经验,本文将篇章单元的前一个单元的篇章成分作为一个结构特征。为保证数据的完整性,将文章首个篇章单元的前一个单元的篇章成分补充为 0。

对于篇章结构评分,实验将作文的段落分为背景介绍段(Introduction)、论证段(Argumentation)、让步段(Concession)和总结段(Conclusion)。我们分析了每个段落中包含的篇章成分分布,将段落中所包含的 9 类篇章成分的个数作为 9 个数值特征。在分析语料库时,发现如果学生在一篇文章中的某些段落的得分较高,那么其段落得高分的概率较大,因此将段落前一段的篇章结构评分作为该段落的上下文特征,将背景介绍段的前一段的篇章结构分数设置为 −1。同时,由于段落中不同篇章成分的论证逻辑与篇章成分的顺序存在相关性,因此我们计算了段落的正确篇章成分序列与预测的篇章成分序列之间的最小编辑距离,并将其作为一个数值特征加入到篇章结构评分特征集中。

#### 3.2.2　词汇特征

本文将 n-grams、指示词、人称代词、情态动词以及专有名词等特征定义为词汇特征。n-gram 模型是一种被广泛应用的统计语言模型,它认为每个预测变量与长度为 $n-1$ 的上下文有关[18]。为了获取具有更大辨别力的 n-gram 语言模型,实验使用三元 n-gram 模型。基于实验语料库,我们使用 SRILM 工具搭建 n-gram 语言模型,对同属于一类篇章成分的篇章单元进行训练,构建这一类篇章成分的 n-gram 语言模型,并计算测试集中篇章单元的 n-gram 词组在不同篇章成分语言模型下的概率,进而计算每个篇章单元的 n-gram 困惑度。

**算法 1**　n-Gram 语言模型生成算法

输入:带篇章成分标签的训练/测试语料库 TC

输出:测试语料库中每个篇章单元的 n-gram 困惑度集合

Step1　将 TC 按句分割成篇章单元;

Step2　按篇章单元类别分成 9 个标签语料库;

Step3　从标签语料库中生成 n-gram 计数文件;

Step4　.基于上一步生成的计数文件,使用插值平滑(Interpolate)以及打折法(Kneser-Ney)训练 n-gram 语言模型;

Step5　利用 n-gram 语言模型计算测试语料库的每个篇章单元的概率,并计算篇章单元的困惑度结果;

Step6　返回至 Step3,对下一个标签语料库重复执行上述步骤,直至计算完每个篇章单元在 9 类篇章成分的 n-gram 语言模型下的困惑度结果;

Step7　输出测试语料库中每个篇章单元不同篇章成分的 n-gram 困惑度集合。

篇章单元中一些特定的词汇表达能直观地反映其篇章成分类别。比如,当一个篇章单元中出现“as a result”“in a word”或“all in all”等指示词时,它更有可能是总结句而不是背景介绍句;当出现“therefore”“thus”或“consequently”等指示词时,我们更倾向于将该篇章单元标注为理由句。我们从最新发布的宾州篇章树库(The Penn Discourse TreeBank,PDTB)2.0 注释手册[19]的指示词列表中删除了不曾出现在

语料库中的指示词，结合语料库常见的指示词，共提取了 64 个指示词，并将这些指示词是否出现在篇章单元中作为指示词特征。

3.2.3 句法特征

为了提取篇章单元的句法特征，我们分析了每个篇章单元的句法结构，使用 Stanford 分析器构建句法分析树，并从中提取树的深度作为句法特征。

例如："Happening things you dislike illustrates that you do not take over the important parts in their hearts."的句法分析树的结构如图 2 所示。

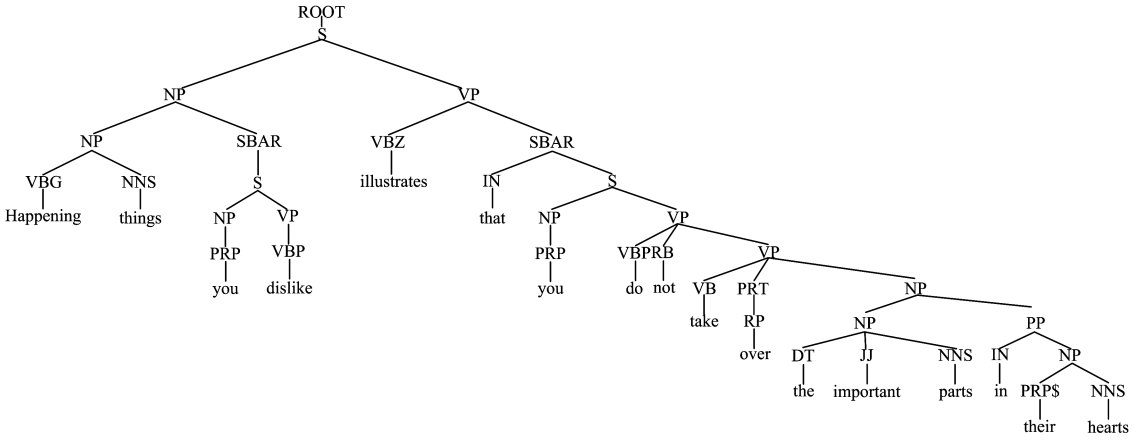


图 2 句法分析树的结构

Fig. 2 Structure of parsing tree

根据上下文无关语法，进一步分析句法分析树的产生式。一般来说，产生式的左端是语法的开始符号 S，所有符合语法规则的树都必须由这个符号作为它们的根标签。SBAR 通常用来标记子句的开始，我们将句法树中 S 与 SBAR 的数量作为篇章单元的子句数。由于论据一般使用过去时态，而论点通常使用现在时态[20]，因此将篇章单元的谓语时态作为一个句法特征。

另外，通过分析语料库发现，在论据中出现人称代词和专有名词的词频明显高于其他篇章成分，情态动词如"should"和"could"也频繁出现在论点句中。实验使用 NLTK 中的 POS 工具对篇章单元的单词进行词性标注，将其中的人称代词(PRP)、专有名词(NNP)及情态动词(MD)的词频作为词汇特征的一部分。

例如："Linda will feel uncomfortable because she cannot accept the smelling of smoke."对应的词性标注为：[('Linda','NNP'),('will','MD'),('feel','VB'),('uncomfortable','JJ'),('because','IN'),('she','PRP'),('can','MD'),('not','RB'),('accept','VB'),('the','DT'),('smelling','NN'),('of','IN'),('smoke','NN')]。

### 3.3 篇章结构评分方法

基于篇章结构的作文自动评分方法分为篇章成分识别和篇章结构评分两部分。首先通过 3.2 节的特征提取方法提取篇章成分特征，将篇章单元结构、词汇及句法等 86 个特征转化成特征向量。基于 RF 构建篇章单元识别模型 DEI-RF，以此预测测试集中每个篇章单元的篇章成分。然后基于 LR 构建篇章结构自动评分模型 DSS-LR，以此完成对文章段落的篇章结构评分。DSS-LR 算法的详细内容如算法 2 所示。

**算法 2** DSS-LR 算法

输入：带标签和评分的作文数据 D={(i,e,s)|u=(i,e,s)∈U}

输出：篇章结构评分 S

```
discourseUnits=NLTK.tokenize(D) //分割语料库
FOREEACH du∈discourseUnits DO
    w=get_wordcout(du)//获取篇章单元的词长度
    h=StanfordParser.row_parse(du).height()//构建句法分析树，并获取树的深度
    pos=NLTK.pos_tag(du)//对篇章单元进行词性标注，提取句法特征
    c=get_tag_context(du)//提取上下文特征
    p=get_punctuation(du)//提取标点符号特征
    FOREACH e∈E DO //计算篇章单元在不同篇章成分下的 ngram 困惑度
      ppl=SRILM.ngram_ppl(du,e)
    FOREACH i∈I DO//提取篇章单元在指示词集合中的特征
      IF du.contain(i) THEN //篇章单元包含指示词
        indicators.append(1)
      ELSE
        indicators.append(0)
    F(du)=addToFeature(w,h,pos,c,p,ppl,indicators)//构建特征向量
normalize(F(du))
compute_min_edit_distance(F(du))//计算最小编辑距离
DEI_Predict=DEI_Modal(F(du),E)//构建模型识别出篇章成分
S=DSS_LR(DEI_predict,s)
return S
```

## 4 实验结果及分析

本文分析了 300 篇中国学生的托福写作议论文，共 6083 个标点句。3 位专家以句子为篇章单元进行篇章成分标注，以段为单位进行篇章结构评分，以此构建带标签及评分的英语议论文语料库。测试集为其中的 66 篇文本，共 1250 个标

点句,训练集为其余的 234 篇文本,共 4 833 个标点句。为了得到性能更好的模型,本文使用十折交叉验证进行测试实验,并利用 Accuracy,Precision,Recall 和 F1-score 指标对模型的性能进行评测。实验分为两部分:篇章成分识别和篇章结构自动评分。

### 4.1　篇章成分识别

本节通过 3.2 节中的方法从篇章单元的结构、词汇及句法中提取 86 个特征,将特征向量化,然后输入到 DEI-RF,DEI-XGB 和 DEI-SVM 模型中进行对比实验。RF 算法在许多现实任务中都展现出了强大的性能,被誉为"代表集成学习技术水平的方法"[21],实验表明 DEI-RF 的性能普遍优于 DEI-XGB 和 DEI-SVM,对比实验结果如表 3 所列。

表 3　篇章成分识别准确率的对比结果

Table 3　Comparison results of accuracy rate of DEI

(单位:%)

| 模型 | 准确率 |
|---|---|
| DEI-RF | 94.13 |
| DEI-XGB | 92.90 |
| DEI-SVM | 80.69 |

网格搜索可以通过调节模型的每个参数来跟踪评分结果,实验使用十折交叉验证的网格搜索来选取不同模型上的最优调节参数。实验结果表明,在测试集上使用 DEI-RF 模型可以获得 94.13%的准确率。

查准率(Precision)和查全率(Recall)常常用来衡量分类模型的性能。我们逐个把篇章成分 $d$ 作为正例,将其他篇章成分作为反例来进行预测实验,计算每次实验的不同篇章成分的 *Precision* 和 *Recall*。令 $P(d)$ 为预测篇章成分为 $d$ 的集合,$T(d)$ 为测试集中 $d$ 的集合,则 *Precision* 和 *Recall* 的计算公式如下:

$$Precision=\frac{\sum_{d\in D}|P_{(d)}\cap T_{(d)}|}{|P_{(d)}|}$$

$$Recall=\frac{\sum_{d\in D}|P_{(d)}\cap T_{(d)}|}{|T_{(d)}|}$$

F1-score 是基于 Precision 和 Recall 的调和平均定义的[21]。实验计算了 3 种模型在不同篇章成分下的 *Precision*,*Recall* 和 *F1-score*,3 种模型的对比结果如表 4—表 6 所列。实验以 *Precision* 为纵坐标、*Recall* 为横坐标绘制不同篇章成分的 P-R(Precision-Recall)曲线,3 种模型的 P-R 图如图 3—图 5 所示。

表 4　DEI-RF 篇章成分的识别性能对比

Table 4　Performance comparison of DEI-RF

| 篇章成分 | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| BI | 0.95 | 1.00 | 0.97 |
| MC | 1.00 | 0.87 | 0.93 |
| TST | 1.00 | 0.93 | 0.96 |
| RS | 1.00 | 0.97 | 0.98 |
| RI | 0.88 | 0.99 | 0.93 |
| ED | 0.97 | 0.91 | 0.94 |
| CS | 0.96 | 0.96 | 0.96 |
| CC | 0.90 | 0.96 | 0.93 |
| IRL | 0.96 | 0.75 | 0.84 |

表 5　DEI-XGB 篇章成分的识别性能对比

Table 5　Performance comparison of DEI-XGB

| 篇章成分 | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| BI | 0.96 | 0.93 | 0.94 |
| MC | 1.00 | 0.83 | 0.90 |
| TST | 1.00 | 0.80 | 0.89 |
| RS | 0.95 | 0.97 | 0.96 |
| RI | 0.90 | 0.95 | 0.92 |
| ED | 0.93 | 0.91 | 0.92 |
| CS | 0.92 | 0.96 | 0.94 |
| CC | 0.94 | 0.92 | 0.93 |
| IRL | 0.89 | 0.97 | 0.93 |

表 6　DEI-SVM 篇章成分的识别性能对比

Table 6　Performance comparison of DEI-SVM

| 篇章成分 | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| BI | 0.95 | 0.91 | 0.93 |
| MC | 0.86 | 0.83 | 0.84 |
| TST | 0.91 | 0.67 | 0.77 |
| RS | 0.89 | 0.89 | 0.89 |
| RI | 0.71 | 0.81 | 0.76 |
| ED | 0.74 | 0.85 | 0.79 |
| CS | 0.91 | 0.83 | 0.87 |
| CC | 0.69 | 0.75 | 0.72 |
| IRL | 0.94 | 0.53 | 0.68 |

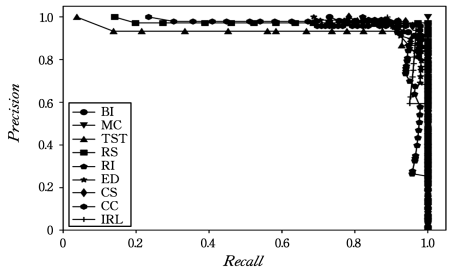


图 3　DEI-RF 模型的 P-R 图

Fig. 3　P-R diagram of DEI-RF

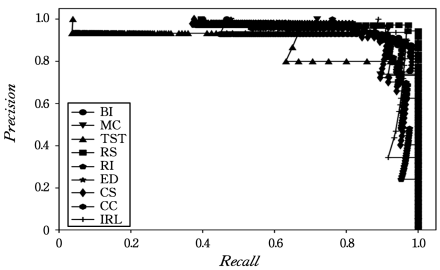


图 4　DEI-XGB 模型的 P-R 图

Fig. 4　P-R diagram of DEI-XGB

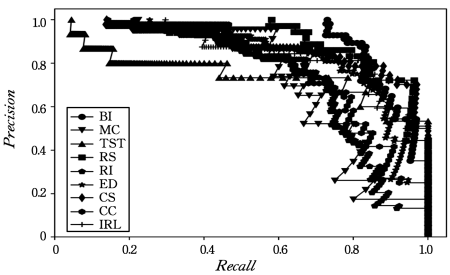


图 5　DEI-SVM 模型的 P-R 图

Fig. 5　P-R diagram of DEI-SVM

分析以上 3 种模型的 Precision,Recall 和 F1-scores 性能

度量数据以及 P-R 图可以看出，DEI-RF 和 DEI-XGB 模型对篇章成分的识别都有较好的效果；特别是在识别 MC，RS，RI，ED，CS 及 CC 等篇章成分时，相比 DEI-SVM，DEI-RF 和 DEI-XGB 的表现更佳，DEI-SVM 对 BI 的识别能力也较好。值得注意的是，3 种模型在识别 TST 与 IRL 时的表现均不如其他篇章成分，而且在 DEI-SVM 模型中 TST 与 IRL 的 *Recall* 值分别为 0.67，0.53。从表 2 的语料集的基本统计信息中可以看出，TST 与 IRL 的数量较其他篇章成分偏少，因此模型在这两个篇章成分上的特征提取能力较弱，容易将其错误识别为其他篇章成分。

将 DEI-RF 模型与现有篇章成分识别模型（DEI-RF 与 Stab 等[2] 提出的 SVM 模型、Song 等[3] 提出的 CRF 模型和 Burstein 等[4] 提出的基于概率的模型）在识别相同类别篇章成分时的 F1-scores 进行对比，结果如表 7 所列。

表 7 不同模型的 *F1-scores* 对比

Table 7 Comparison of *F1-scores* of different models

| 篇章成分 | DEI-RF | SVM | CRF | 概率模型 | 平均提升 |
|---|---|---|---|---|---|
| BI | 0.97 | — | 0.90 | 0.57 | +0.235 |
| MC | 0.93 | 0.63 | 0.75 | 0.73 | +0.226 |
| RS | 0.98 | 0.54 | 0.70 | 0.81 | +0.296 |
| ED | 0.94 | 0.83 | 0.90 | 0.91 | +0.060 |
| CC | 0.93 | — | 0.93 | 0.84 | +0.045 |
| IRL | 0.84 | — | — | 0.75 | +0.090 |

相较于 SVM 模型、CRF 模型与基于概率的模型，DEI-RF 模型在识别相同类别篇章成分时的性能均有提升，特别是在识别 BI，MC 以及 RS 时效果更加明显。

同时，实验还研究了 3.2 节中能有效识别篇章成分的具体特征。实验在整个数据集中评估了篇章单元的结构、词汇及句法特征对 DEI-RF 模型的性能影响，结果如表 8 所列。

表 8 DEI-RF 不同特征的 *F1-scores* 对比

Table 8 Comparison of different features of *F1-scores* in DEI-RF

| 篇章成分 | 结构特征 | 词汇特征 | 句法特征 |
|---|---|---|---|
| BI | 0.93 | 0.52 | 0.19 |
| MC | 0.86 | 0.47 | 0.22 |
| TST | 0.80 | 0.86 | 0.17 |
| RS | 0.97 | 0.66 | 0.40 |
| RI | 0.86 | 0.60 | 0.31 |
| ED | 0.88 | 0.60 | 0.49 |
| CS | 0.90 | 0.52 | 0.21 |
| CC | 0.88 | 0.56 | 0.18 |
| IRL | 0.80 | 0.71 | 0.58 |

从表 8 中可以看出，结构特征在十折交叉验证中表现良好，可以较好地识别出数据集中的 BI，RS，ED，CS 和 CC 等篇章成分，而这些部分也正是一篇英文议论文的核心部分。由于在 TST 和 IRL 这两类篇章成分中一般不含有指示词，且词汇较为单一，n-gram 特征相比其他篇章成分更为明显，因此词汇特征对识别这类篇章成分的帮助较大。句法特征对于识别篇章成分的作用不明显，但相对于其他篇章成分类别，句法特征在识别 ED 和 IRL 的篇章成分上有更高的准确率。

### 4.2 篇章结构自动评分

DSS-LR 以段为单位对语料库的篇章结构进行评分，段落根据语料库特点分为：背景介绍段（Introduction）、论证段（Argumentation）、让步段（Concession）以及结论段（Conclusion）。背景介绍段主要包含的篇章成分有 BI，MC 和 TST，论证段主要包含的篇章成分有 RS，RI，ED 和 TST，让步段主要包含的篇章成分有 CS，MC 和 TST，结论段则只包含 CC 篇章成分。语料库中 4 种段落的数量信息如表 9 所列。

表 9 段落数量信息

Table 9 Number information of paragraph

| 段落 | 数量 |
|---|---|
| Introduction | 294 |
| Argumentation | 575 |
| Concession | 276 |
| Conclusion | 277 |

我们在 4.1 节的基础上使用 DEI-RF 模型识别出测试集中每个篇章单元的篇章成分，利用特征提取程序提取不同段落的篇章成分分布特征、上下文特征和最小编辑距离特征，并利用基于高斯概率密度的异常点检测方法 EllipticEnvelope 剔除异常点，然后构建 DSS-LR 模型。

周志华[21] 提到，均方误差是回归任务中最常用的性能度量。本文使用均方差（Mean Squared Error，MSE）并结合 R 方（R-squared，RS）来评测 DSS-LR 模型在测试集上的性能。值得注意的是，专家在评分过程中会产生一定的随机误差，而在实际评分中一般认为分数在较小范围内的偏差不影响作文的实际质量，比如分数为 90.3 分与 90 分的作文的质量差别不大。于是，在实验中我们对篇章结构评分设置了一个容忍度，根据评分经验将容忍度值设为 0.5 分，即在这个容忍度值的范围内，预测分数是相似的，由此我们可以计算出模型的容忍分数（Tolerance Score，TS）。不同段落的 *MSE*，*RS* 和 *TS* 如表 10 所列。

表 10 不同段落评分性能的对比

Table 10 Performance comparison of different paragraph scores

| 段落 | *MSE* | *RS* | *TS* |
|---|---|---|---|
| Introduction | 0.02 | 0.91 | 0.93 |
| Argumentation | 0.11 | 0.87 | 0.89 |
| Concession | 0.08 | 0.82 | 0.96 |
| Conclusion | 0.21 | 0.58 | 0.84 |

从表 10 可以看出，DSS-LR 模型对 Introduction 段、Argumentation 段及 Concession 段的评分效果较佳，而对 Conclusion 段的评分效果较差，主要原因在于 Conclusion 段仅包含一种篇章成分，DSS-LR 模型在该段上能提取的篇章成分特征远少于其他 3 类段落。在实际作文数据中，Introduction 段、Argumentation 段以及 Concession 段的篇幅占 90%以上，其中 Argumentation 段的篇幅占 50%左右，因此判断一篇作文篇章结构的好坏，主要取决于 Introduction 段、Argumentation 段以及 Concession 段。

**结束语** 本文的主要工作分为两步，首先是篇章成分识别，然后在此基础上进行作文的篇章结构的自动评分。通过自建真实应考学生的作文语料库，以句子为单位划分篇章单元，并提取篇章单元的结构、词汇以及句法等 86 个细粒度特征，使用 DEI-RF，DEI-XGB 和 DEI-SVM 3 种模型来识别篇章成分，最后构建了 DSS-LR 模型对作文的篇章结构进行自动评分。实验完整地完成了特征提取、篇章成分识别以及篇章结构评分，并提高了篇章成分识别准确率和在背景介绍段、

论证段及让步段上的篇章结构评分准确率。但在篇章成分识别中,特征工程经验在一定程度上决定了模型结果的优劣,之后的工作可以使用递归神经网络等深度学习技术自主学习特征;在篇章结构评分过程中,在论证逻辑识别和语义分析方面还需要进行进一步的研究;值得注意的是,本文主要针对作文的篇章结构进行评分,考虑到自动作文评分系统的完整性,后续可以结合作文的内容和语法再进行进一步的研究。

## 参 考 文 献


[1] STAB C,GUREVYCH I. Parsing Argumentation Structures in Persuasive Essays[J]. Computational Linguistics,2017,43(3):619-659.

[2] STAB C,GUREVYCH I. Identifying Argumentative Discourse Structures in Persuasive Essays[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:46-56.

[3] SONG W,FU R,LIU L,et al. Discourse Element Identification in Student Essays based on Global and Local Cohesion[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:2255-2261.

[4] BURSTEIN J,MARCU D,KNIGHT K. Finding the WRITE Stuff:Automatic Identification of Discourse Structure in Student Essays[J]. IEEE Intelligent Systems,2003,18(1):32-39.

[5] YIGAL A,JILL B. Automated Essay Scoring with E-rater® v. 2. 0 [J]. The Journal of Technology,Learning,and Assessment,2006,4(2):1-21.

[6] PALTRIDGE B. Discourse Analysis for the Second Language Writing Classroom[M]// The TESOL Encyclopedia of English Language Teaching. John Wiley & Sons,Inc.,2017.

[7] HSIEH C J,CHANG K W,LIN C J,et al. A dual coordinate descent method for large-scale linear SVM [C]// International Conference on Machine Learning. Helsinki, Finland: IEEE press,2008:1369-1398.

[8] BREIMAN L. Random Forests[J]. Machine Learning,2001,45(1):5-32.

[9] CHEN T,GUESTRIN C. XGBoost:A Scalable Tree Boosting System[C]// Acm SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2016:785-794.

[10] MANN W. Rhetorical Structure Theory:Toward a Functional Theory of Text Organization[J]. Text & Talk,2009,8(3):243-281.

[11] DUVERLE D A,PRENDINGER H. A novel discourse parser based on support vector machine classification[C]// International Joint Conference on Natural Language Processing of the Afnlp. ACL,2010:665-673.

[12] FENG V W,HIRST G. A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing[C]// Proceeding of the 52nd Annual Meeting of the Association for Computational Linguistics. ACL,2014:511-521.

[13] YAN W R,XU Y,ZHU S S,et al. A Survey to Discourse Relation Analyzing[J]. Journal of Chinese Information Processing,2016,30(4):1-11. (in Chinese)
严为绒,徐扬,朱珊珊,等. 篇章关系分析研究综述[J]. 中文信息学报,2016,30(4):1-11.

[14] LI S,KONG F,ZHOU G D. A PDTB-Based Automatic Explicit Discourse Parser[J]. Journal of Chinese Information Processing,2016,30(2):18-25. (in Chinese)
李生,孔芳,周国栋. 基于 PDTB 的自动显式篇章分析器[J]. 中文信息学报,2016,30(2):18-25.

[15] XU F,ZHU Q M,ZHOU G D. Implicit discourse relation recognition based on tree kernel[J]. Chinese Journal of Software,2013,24(5):1022-1035. (in Chinese)
徐凡,朱巧明,周国栋. 基于树核的隐式篇章关系识别[J]. 软件学报,2013,24(5):1022-1035.

[16] JIANG Y R,SONG R. Topic clause identification method based on specific features[J]. Journal of Computer Applications,2014,34(5):1345-1349. (in Chinese)
蒋玉茹,宋柔. 基于细粒度特征的话题句识别方法[J]. 计算机应用,2014,34(5):1345-1349.

[17] BIRAN O,RAMBOW O. Identifying Justifications in Written Dialogs[J]. International Journal of Semantic Computing,2011,5(4):363-381.

[18] XING Y K,MA S P. A Survey on Statistical language Models [J]. Computer Science,2003,30(9):22-26. (in Chinese)
邢永康,马少平. 统计语言模型综述[J]. 计算机科学,2003,30(9):22-26.

[19] PRASAD R,MILTSAKAKI E,DINESH N,et al. The penn discourse treebank 2. 0 annotation manual[J]. Proceedings of Lrec,2007,24(1):2961-2968.

[20] PALAU R M,MOENS M F. Argumentation mining:the detection,classification and structure of arguments in text[C]// International Conference on Artificial Intelligence and Law. ACM,2009:98-107.

[21] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.