基于局部密度的无监督作文跑题检测方法

李霞[1,2]，温启帆[2]

（1.广东外语外贸大学 语言工程与计算实验室，广东 广州 510006；

2.广东外语外贸大学 信息科学与技术学院，广东 广州 510006）

摘　要：针对现有的无监督作文跑题检测方法中使用作文内容向量表示作文存在非主题词噪音所导致的相似度不准确问题，本文提出一种基于作文主题词抽取和局部密度阈值选择的无监督作文跑题检测方法。首先使用 LDA 主题生成模型挖掘待测作文的主题词，并使用分布式表示向量寻找与题目词项语义相似的词作为对作文题目的主题词扩展，在此基础上使用本文提出的切题度计算方法计算待测作文的切题度，并使用所提出的基于作文集切题度局部密度的阈值抽取方法动态选取切题阈值，进而实现一种无需训练集和主题无关的无监督作文跑题检测方法。在以英语为母语的学习者和以汉语为母语的学习者所写的 8 个作文集共 9381 篇作文上的实验结果表明，本文提出的作文跑题检测方法能有效识别跑题作文。

关键词：作文跑题检测；主题词抽取；切题度；阈值选取

中图分类号：TP391　　**文献标识码：**A

Unsupervised Off-topic Essay Detection Based on Local Density

LI Xia[1,2], WEN Qifan[2]

(1. Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou, 510006, China;

2. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, 510006, China;)

Abstract: Existing off-topic essay detection methods mainly use content vector to express the composition which sometimes results in low accuracy due to noise words. In this paper, we propose an unsupervised off-topic essay detection method based on the topic words and the local density thresholds. Firstly, Latent Dirichlet Allocation is used to predict essay's topic distribution and the topic words are extracted according to the different weights of the topics. Secondly, we use distributed word vector representation to find the similar words as the expansion of the topic of the composition, and then compute on-topic score of all the test essays using our new similarity calculation method. Finally, we propose a local density threshold extraction method to extract the off-topic threshold automatically and ultimately determine off-topic essay. The experimental results on eight essay sets with 9381 essays in total show that our algorithm can greatly improve the F-measure value compared to our baseline method.

Key words: off-topic essay detection, topic word extraction, on-topic score, threshold extraction

1 引言

对于机器评分系统，当应试者通过拷贝、背诵、堆砌词汇等方式输入一篇与作文题目无关的作文时，系统如果不做跑题检测，可能会给该作文评出较高的分数，从而影响机器评分系统的公平性和准确性。因此，作文跑题检测对于作文自动评分系统的公平性、鲁棒性和准确性具有重要的意义。

作文跑题是指作文偏离题目所要求的主题并写成其他无关的主题作文。例如题目要求学生就“全球淡水资源短缺问题”写一篇议论文，而所写的作文是有关“对假冒伪劣商品的看法”或“对社会实践的重要性的讨论”等与淡水资源短缺无关的主题时，则该作文将被认定为跑题作文。目前作文跑题检测方法主要包括有监督作文跑题检测方法和无监督作文跑题检测方法，前者需要事先对已经标注好的大规模跑题作文进行训练，使用机器学习方法中的分类等方法实现跑题作文的检测。然而，在很多实际教学场景中，当教

基金项目：国家自然科学基金“面向中国英语学习者的英文作文全自动评分及诊断反馈技术研究”(61402119)；广东省普通高校科技创新项目“面向网络英文文本的涉华舆情分析关键技术研究”(2013KJCX0071)。

作者简介：李霞，shelly_lx@126.com，教授，主要研究方向为自然语言处理和语言测试；温启帆，本科生，主要研究方向为自然语言处理。

师给出一个新的作文题目时，往往事先并没有标注好的跑题作文数据。因此，针对事先没有作文训练集，通过作文题目的描述信息来自动检测作文是否跑题的无监督作文跑题检测研究成为近年来作文跑题检测的主要研究内容。

作文跑题检测的核心问题是判断作文的主题是否偏离作文题目给定的主题。通常，一篇作文为了论证作者的思想或观点，往往会通过几个子观点来论证其核心观点，因此一篇作文可能会包含多个子主题，而这些子主题中有些在语义层面与题目的相关度较低，这使得单纯通过抽取特征词来向量化作文和题目并基于此来计算相似度，有可能因为不相关的子主题词导致作文和题目相似度的不准确，从而影响到跑题检测的最终结果。

基于以上问题，本文提出一种不同于现有的作文内容向量表示的方法，通过使用LDA主题生成模型对待测作文生成作文所包含的主题集合，并依据主题概率所占的权重按一定比例抽取作文主题中的关键词组合作为作文最终的主题信息，并基于这些核心主题词信息和作文题目信息之间的语义相似程度来判断作文是否跑题，进而避免了传统方法中使用作文特征词来判断作文是否跑题时所引入的噪音特征词问题。在此基础上，本文还提出了有效的作文和题目的相似度计算方法和基于局部密度的阈值抽取方法，最终实现了一种无需作文训练集和主题无关的无监督作文跑题检测系统。

2 相关工作

现有研究中，Higgins 等[1-3]将作文和题目表示为包含作文内容的空间向量，使用余弦相似度来计算作文和题目之间的相关程度。与传统方法所不同的是，Higgins 等在工作中引入了参考题目，即和作文的目标题目所不同的题目集合，通过计算待测作文与目标题目以及参考题目之间的相似度，并对这些相似度进行排序，判断待测作文与目标题目之间的相似度占整个排序集合中的排名比例来判断待测作文是否跑题，例如认为与目标题目相似度排名在前 10%时认为是切题作文，否则认为是跑题作文。Persing 和 Ng[4]基于作文的丰富特征和人工事先标注好的与主题相关的句子分值，通过建立线性回归方程来构建作文与主题的一致性分值计算方法。该方法由于需要针对不同的作文主题训练得到不同的评分模型，属于有监督的主题一致性计算方法。Cummins 等[5]分别使用分布式语义和信息检索中的伪相关反馈方法对作文题目进行扩充，提高了作文主题相关性的计算结果，同时系统还将该主题相关性模型纳入到一个有监督的评分系统中，结果表明该方法可以有效提升系统对作文的综合评分性能。Rei 和 Cummins[6]在句子级别的主题相关性判别领域做了一定的研究，其使用 Word2Vec 词向量按词语权重叠加的方式表示句子向量，结合 tf*idf 特征权值，以余弦相似度计算句子和主题的相关程度，实验结果表明该方法具有较强的鲁棒性。

陈志鹏等[7]将文本中的单词采用词向量表示，并基于分布式表示扩展与其语义上相近的词，基于此提升作文和题目的相似度计算。在其后续研究中[8]提出了一种基于文档发散度的概念，通过大规模作文回归模型训练得到发散度与跑题阈值的关系模型，从而实现对不同题目动态选取不同跑题阈值的方法。该方法需要事先具有大规模来自不同主题下已经标注好的作文训练数据通过训练才能得到回归参数，所以本质上还是属于有监督的作文跑题检测方法。李晓亚[9]针对不同的应用场景分别提出了几个跑题检测模型，其中基于题目排序的跑题检测方法属于无监督跑题检测方法，该方法延续使用了空间向量模型方法并基于 WordNet 进行了词扩展来提升作文和题目的相似度比较。范弘屹等[10-11]也分别研究了基于 HowNet 或 WordNet 来计算和提升词语的语义相似度问题。梁茂成等[12-13]的工作中也涉及了作文内容分析和相似度的计算，但均需要事先标注好的作文训练语料，属于有监督的方法。

已有的无监督作文跑题检测方法从不同层面改进了作文与题目的语义相似度计算，进而提升作文跑题检测的结果，但是这些方法是将作文表示为内容向量，并采用特征词抽取方法来表示作文，依然存在非主题词被选入所导致的噪音问题。阈值抽取方面，已有的无监督方法中 Higgins 明确给出了作文跑题的阈值判断方法，该方法通过判断作文与参考题目和目标题目之间相似度差异来检测作文是否为跑题作文，但是该方法在不同的参考题目下检测结果可能存在较大的差异。

针对以上两个层面的不足，本文提出一种基于作文主题词抽取和局部密度阈值选择的无监督作文跑题检测方法，主要贡献包括：①基于 LDA 主题生成模型预测待测作文的主题分布，抽取更为准确的作文主题词信息；②提出面向作文跑题检测的有效相似度计算方法；③根据待测作文的切题度分布密度实现对跑题阈值的自动抽取。实验结果表明本文提出的作文跑题检测方法能有效识别跑题作文。

3 基于局部密度的无监督作文跑题检测方法

3.1 作文关键词抽取

通常，一篇作文为了论证作者的思想或观点，往往会通过几个子观点来论证其核心观点，因此一篇作文可能会包含多个子主题，例如一篇描写关于“大学生活”的作文，作者可能从学习和娱乐两个方面阐述其大学生活，并在作文中着重于学习方面的描写。因

此，针对跑题检测，既要抽取出作文中有关学习方法的子主题，也要抽取其有关娱乐方面的子主题，并能够依据作文的侧重点对不同主题分别对待。基于这样一个思想，本文提出使用 LDA 主题生成模型[14]对作文进行主题抽取，利用 LDA 主题模型去“理解”作文的主题分布，并根据各个主题分布概率作为权重在各个主题下提取不同数量的关键词。

首先使用现有的大规模语料利用 LDA 主题生成模型训练得到一个主题-词分布矩阵 M，当输入一篇待测作文时，依据矩阵 M 得到待测作文的文档-主题分布概率向量 $V=\{v_1,v_2,...,v_i,...,v_n\}$，其中 n 表示事先设定的主题总数，v_i 表示待测作文属于第 i 个主题的概率。对 v_1 到 v_n 按照概率的大小从高到低排序，选取前 k 个主题 $\theta_1 \sim \theta_k$ 作为待测作文的主题。设总的作文主题词抽取个数为 h 个，则每个主题 θ_i 中抽取的主题词个数为 $N_{\theta_i}=\frac{v_{\theta_i}}{\sum_{i=1}^{k} v_{\theta_i}}\times h$，依次对 $\theta_1 \sim \theta_k$ 个主题中的主题词排序，并分别抽取前 N_{θ_i} ,$1 \le i \le k$ 个主题词并组合而成主题词列表，去除重复词后得到最后表示待测作文主题的主题词列表。

表 1 分别给出了三个题目作文中某一篇待测作文提取的关键词及其概率值，表格中的主题 1~主题 5 是按照主题概率排序后的前 5 个主题，参数中设置的抽取主题词个数为 h = 30。从表 1 可以看出，作文的不同子主题抽取出了不同数量的作文主题词，从而更为真实的反映作文的内容主题。

表 1 本文方法在 3 个主题中各选取一篇待测作文抽取的特征词列表

Prompt#1 Global Shortage of Fresh Water									
主题 1		主题 2		主题 3		主题 4		主题 5	
waste	0.00521	men	0.00433	shortage	0.00568	industries	0.0192	drink	0.00209
scientists	0.003168	ways	0.0021	countries	0.00264	development	0.00049	purse	0.00101
water	0.001219	water	0.00102	fresh	0.00256	Large	0.00039	mountains	0.00083
important	0.001138	supply	0.00101	actually	0.00244	limited	0.00030		
growing	0.001004	rivers	0.00089	polluting	0.000367	required	0.00026		
population	0.000928	result	0.00055	steps	0.000015				
stop	0.000928	poluted	0.00054						
quality	0.00041	left	0.000464						
Prompt#2 The Effects Computers have on People									
主题 1		主题 2		主题 3		主题 4		主题 5	
favors	0.00209	persuade	0.000506	computer	0.00433	communicat	0.00113	local	0.000359
matter	0.00106	science	0.00049	computers	0.00244	pinish	0.00096	newspaper	0.000014
essays	0.00101	people	0.000399	thing	0.00101	homework	0.00035	bothering	0.000014
readers	0.00092	ideas	0.000307	amazing	0.00036	kids	0.00074	informants	0.000013
writing	0.00024	social	0.000261	Fault	0.000014	device	0.00021		
effect	0.00017	studies	0.000169						
love	0.00014	give	0.000016						
addicting	0.00014								
Prompt#3 The Features of the Setting Affect the Cyclist									
主题 1		主题 2		主题 3		主题 4		主题 5	
make	0.0097	drinking	0.00061	dried	0.01979	yosemite	0.000359	making	0.000196
system	0.0021	shows	0.00054	circled	0.00209	lose	0.000157	past	0.000491
heatstroke	0.00102	extreme	0.00023	growing	0.00132	features	0.000143	left	0.000261
water	0.00085	lack	0.00021	heat	0.00101	affected	0.000143	danger	0.0002
cyclist	0.00055	eased	0.00014	ends	0.00024	heat	0.000143		
stroke	0.00055	he	0.00014	story	0.00017	lot	0.000132		
rings	0.000464	dehydrated	0.00009						
shirt	0.000453								

3.2 作文题目主题词扩展

词的分布式表示是指将词表中的词映射为一个稠密的、低维的实值向量，每一维表示词的一个潜在特征，可以反映词与词之间的语义关系，通过单词的词向量形式可以找出与其语义上相近的词。Word2vec[15-16]是 Google 在 2013 年开源的词向量工具包，可以实现将词语表示成具有语义的词向量，通过词向量的余弦相似度可以测量词语间的语义距离，从而用于获取语义相近的词语。

本文首先对题目进行分词和去停用词等预处理，假定题目预处理后的特征词列表为 $T=(t_1,t_2,...,t_n)$，对该列表中的每个特征词 t_i，首先基于训练好的分布式词向量模型表示该特征词为一个向量，然后计算出和特征词 t_i 在分布式向量上余弦相似度较高的单词作为其扩展词。本文中，我们对题目的每一个特征词选取了前 10 个相似度最大的词作为扩展词，分布式表示词向量训练和采用的是 50 维词向量模型。

3.3 作文切题度的定义和计算

针对本文的研究，我们将作文切题度定义为作文与题目之间在主题内容上的相似程度，通过判断作文与题目之间的切题度来划分该作文属于切题作文还是跑题作文。首先通过 3.1 节对待测作文进行主题词抽取，然后使用 3.2 节方法对作文题目基

于语义上下文信息进行扩展，然后计算两者之间的相似度。

考虑到作文的不同子主题具有不同的重要程度，本文的相似度计算方法是针对题目扩展后的每个主题词，分别计算其与作文每个主题词的相似度，并使用最大相似度值作为当前主题词与作文主题词之间的相似度，以此类推计算下一个题目主题词与作文主题词的最大相似度，最后以题目主题词的最大相似度的平均值作为作文的最终切题度分值，详细分值计算公式为：

$$Score(essay, prompt) = \frac{\sum_{i=1}^{n} \max_{j=1}^{m} \{sim(t_i, w_j)\}}{N}$$

其中 *essay* 表示待测作文文本，$(w_1,w_2,...,w_m)$ 为作文 *essay* 中抽取得到的主题词，*prompt* 为作文的题目，$(t_1,t_2,...,t_n)$为作文题目经过扩展后的全部主题词。$sim(w_i, t_j)$为作文主题词 w_i 和题目主题词 t_j 转换为词向量后的余弦相似度，N 为题目扩展后的特征词的总数。

3.4 基于局部密度的阈值选择

跑题检测的最终目标是将待测作文划分成跑题和切题两个类别，理论上如果能够找到某一个维度或多个维度指标使得切题作文和跑题作文各自聚集成独立的两个簇，则可以比较好的找到两个簇的边界阈值，从而划分作文为跑题簇或跑题簇。我们发现，虽然跑题作文本身的内容主题差异较大，但在实际作文数据中，由于跑题作文与题目主题无关，内容差异大，因此跑题作文之间切题度的差值会大于切题作文之间切题度的差值，因此我们认为作文的局部密度可以有效划分开跑题作文和切题作文。基于此，本文提出了一种基于局部密度的阈值选择策略，具体算法描述如下：

基于局部密度的阈值选择算法

输入：从大到小排序后的待测作文集合切题度分值列表 $S=\{s_1,s_2,...,s_n\}$, n 为待测作文的总数；

输出：划分阈值 r；

1. 定义滑动窗口 w 为切题度列表中移动的区间，窗口长度设为 n/10；

2. 初始时窗口位于 S 的最高值处 w=[1, n/10]，依据式 1 计算当前窗口的密度 *density*(w)，其中|w|表示窗口 w 中作文的个数；

$$density(w) = \frac{1+|w|^2}{1+\sum_{s_i,s_j \in w} |s_i - s_j|} \quad (式1)$$

3. 将窗口 w 向下滑动 1 位使得 w=[2, n/10+1]，计算当前窗口新的密度；

4. 重复第 3 步直到窗口在样本区间全部滑动完毕，设全部窗口中密度值最大的窗口为 w_{max}=[k, k+n/10]，则切题作文簇的质心为$(s_k+s_{k+n/10})/2$，记为 $s_{on\text{-}topic}$；

5. 另起一个窗口 w'位于 S 的最小值处 w'=[9n/10, n]，并在区间[k+2n/10, n]之间从下往上滑动；

6. 窗口 w'每次向上滑动 1 位，依据式 1 计算当前滑动窗口的密度，向上滑动时，比较向上滑动后的密度与当前窗口的密度的大小，当向上滑动后的密度小于当前窗口的密度时，则终止滑动。设最终停止滑动的窗口为 w'_{max}=[l, l+n/10], 则跑题作文簇的质心为$(s_l+s_{l+n/10})/2$，记为 $s_{off\text{-}topic}$；

7. 划分阈值 $r=(s_{on\text{-}topic}+s_{off\text{-}topic})/2$；

4 实验设置与结果分析

4.1 实验数据

为了验证本文方法的有效性，分别选取了以英语为母语的学习者和以英语为二语的中国英语学习者所写的两个不同类型的作文语料库，选取了其中的 8 个作文主题下共 9381 篇作文进行测试，文中将这 8 个作文主题分别标号 Prompt #1~Prompt #8，其中 Prompt #1~Prompt #4 来自 kaggle 的作文评分比赛数据集①，Prompt #5~ Prompt #8 来自中国英语学习者语料库 CLEC 作文数据集[17]。

数据集中的跑题作文主要包括两个来源，一是从原始作文集中抽取并经过人工判定为跑题的低分作文，另一部分是从其他题目下随机抽取的不同主题的作文。其中，Prompt #1 从原始作文集中提取最低分为 0 分并经人工判断为跑题的作文 28 篇，从 kaggle 数据集中其他 3 个不同题目下分别随机抽取 100 篇共 300 篇，合计为 328 篇跑题作文。Prompt #2 从最低评分为 5 分的作文中人工判断筛选出跑题作文 31 篇，从 kaggle 数据集其他 3 个不同题目下分别随机抽取 90 篇共 270 篇，合计为 301 篇跑题作文。Prompt #3 从原始作文集中提取最低分为 0 分并经过人工判断为跑题的作文 3 篇，从 kaggle 数据集其他 3 个不同题目下分别随机抽取 91 篇共 273 篇，合计为 276 篇跑题作文。Prompt #4 从最低评分为 1 分的作文中人工判断筛选出跑题作文 43 篇，从 kaggle 数据集其他 3 个不同题目下分别随机抽取 50 篇共 150，合计 193 篇跑题作文。Prompt #5、

① https://www.kaggle.com/c/asap-aes/data

Prompt #6、Prompt #7 和 Prompt #8 分别从 CLEC 语料库其他 4 个不同题目中分别随机抽取共 44 篇、50 篇、20 篇和 30 篇跑题作文。整个实验数据集的详细描述如表 2 所示。

表 2 本文作文数据集描述

作文题目标号	作文题目	作文篇数(篇)	切题作文(篇)	跑题作文(篇)
Prompt #1	The Features of the Setting Affect the Cyclist	2054	1726	328
Prompt #2	The Effects Computers have on People	2054	1753	301
Prompt #3	The Obstacles the Builders of the Empire State Building faced in attempting to allow dirigibles to Dock	2032	1756	276
Prompt #4	Censorship in the Library	1969	1776	193
Prompt #5	My View on Job-Hopping	356	312	44
Prompt #6	Global shortage of Fresh Water	390	340	50
Prompt #7	Getting to Know the World Outside the Campus	203	183	20
Prompt #8	Haste Makes Waste	323	293	30

4.2 评价指标

采用信息检索中常用的检索正确率（precision）、召回率（recall）和 F1 度量值作为本文算法的评测指标。同时也参考了 Higgins[1-3]中使用的 FP(False Positive)和 FN(False Negative)两个指标作为辅助评价指标，相应的 5 个指标公式描述如下：

$$precision=\frac{\#正确检测出的跑题作文个数}{\#检测出的跑题作文总个数}$$

$$recall=\frac{\#正确检测出的跑题作文个数}{\#跑题作文总个数}$$

$$F1=\frac{2P\times R}{P+R}$$

$$FP=\frac{\#检测为跑题但实际为切题的作文个数}{\#切题作文总个数}$$

$$FN=\frac{\#检测为切题但实际为跑题的作文个数}{\#跑题作文总个数}$$

4.3 实验参数和比较的基准方法

将传统基于作文内容向量表示的方法作为本文的基准比较方法(文中以 tf*idf 方法来命名)，同时参考了 Higgins[2]工作中提到的使用拼写纠错和词形还原等预处理步骤，分别使用这两个预处理步骤加入到本文方法进行比较。实验还对 Higgins 中提到的基于参考作文题目的阈值划分方法与本文的局部密度阈值划分方法进行了比较和分析。

实验中 LDA 主题-词概率分布矩阵所使用的训练语料采用了路透社语料库 10788 个新闻文档共计 130 万字和 90 个主题，分布式词向量采用 11G 大小的维基百科数据源进行训练，LDA 关键词提取方法中参数 k 选取为 5，提取的关键词数量为 30，超参数 α 为 0.1，超参数 β 为 0.1，最大迭代次数为 5000。

4.4 实验结果与分析

首先我们对传统基于作文内容向量表示的方法和本文方法进行了比较，实验结果如表 3 所示。实验结果表明，本文方法相比传统的向量表示方法在 8 个作文数据集上 F1 值均有不同程度的提高。其中 tf*idf 方法最好的结果在数据集 Prompt #3 上，在该数据集上 tf*idf 方法的精度、召回率和 F1 度量值分别为 87.98%，83.15%和 85.49%，而本文方法在该数据集上的精度、召回率和 F1 度量值则分别为 94.54%，95.23%和 94.89%，分别提升了 6.56、12.08 和 9.4 个百分点。在数据集 Prompt #5 上，tf*idf 方法的 F1 度量值分别为 81.31%，而本文方法在该数据集上的 F1 度量值则分别为 95.05%，提升了 13.74 个百分点。在所有 8 个作文数据集上，整体的平均 F1 度量值分别为 71.77%和 78.03%，提升了 6.26 个百分点，整体效果提升显著。

表 3 tf*idf 方法和本文方法在 8 个数据集上的实验结果对比

数据集	tf*idf 方法				本文方法			
	precision	recall	F1	阈值	precision	recall	F1	阈值
Prompt #1	55.74%	57.16%	56.44%	0.1413	48.56%	97.87%	64.91%	0.5280
Prompt #2	84.95%	69.43%	76.41%	0.0794	88.37%	75.74%	81.57%	0.5487
Prompt #3	**87.98%**	**83.15%**	**85.49%**	0.1664	**94.54%**	**95.23%**	**94.89%**	0.4823
Prompt #4	60.49%	76.16%	67.43%	0.1586	67.51%	82.90%	74.41%	0.4832
Prompt #5	54.32%	100.00%	70.40%	0.0538	70.45%	70.45%	70.45%	0.4998
Prompt #6	**78.72%**	**84.09%**	**81.31%**	0.2574	**94.12%**	**96.00%**	**95.05%**	0.5807
Prompt #7	61.53%	80.00%	69.56%	0.1337	66.67%	80.00%	72.72%	0.4727
Prompt #8	67.50%	66.72%	67.11%	0.1415	59.09%	86.67%	70.27%	0.5082
平均值	**68.90%**	**77.09%**	**71.77%**	0.1415	**73.66%**	**85.60%**	**78.03%**	0.5129

为了进一步测试本文的方法，我们将加入拼写纠错和词形还原两个预处理方法结合到本方法中，并在 8 个数据集上进行试验，结果如表 4 所示。从表 4 可以看出，经过拼写检查预处理后，本文方法整体结果有所提升，在 8 个数据集上的平均准确率、平均召回率和平均 F1 值相比预处理前的结果分别

提升了 1.96、1.17 和 1.61 个百分点，同时本文方法相比传统 tf*idf 方法，经过拼写检查预处理后，平均 F1 值提升 7.87 个百分点。表 4 还表明，词形还原对改善跑题检测的最终实验结果没有太大提升。

表 4 各方法在 8 个数据集上实验结果对比

跑题检测方法	平均准确率	平均召回率	平均 F1 值
tf*idf 方法	68.90%	77.09%	71.77%
本文方法	73.66%	85.60%	78.03%
拼写纠错+本文方法	**75.62%**	**86.77%**	**79.64%**
词形还原+本文方法	71.81%	80.13%	75.74%
拼写纠错+词形还原+本文方法	74.32%	83.63%	77.72%

我们还对另外两个指标 FN(False Negative)和 FP(False Positive)值进行了实验对比，由于本文的数据集来源于两个类别，一个是以英语为母语学习者的作文数据集，另一个是以英语为二语的中国英语学习者语料库。我们分别计算不同方法在母语学习者和二语学习者数据集的效果，结果如表 5 所示。从表 5 可以看出，本文方法相比基准 tf*idf 方法在平均 F1 值和平均 FN 值以及平均 FP 值上，均有所提升。同时表 5 还可以看出，增加拼写检错预处理后，在二语学习者的数据集上本文方法平均 F1 值提升了 2.04 个百分点，而在母语学习者数据集上提升了 1.29 个百分点，这说明，增加拼写检查后，虽然都有所提升，但是针对二语学习者的中国英语学习者作文数据，由于作文的语法错误相对比母语学习者多，因此经过纠错后，整体提升的效果更大。

表 5 各方法在不同母语语言学习者作文上的平均 FN 和平均 FP 值的实验结果对比

跑题检测方法	以英语为母语的学习者作文			以汉语为母语的学习者作文		
	平均 F1 值	平均 FN 值	平均 FP 值	平均 F1 值	平均 FN 值	平均 FP 值
tf*idf	71.44%	25.32%	3.21%	72.10%	14.73%	6.49%
本文方法	78.95%	16.80%	2.25%	77.12%	16.72%	3.89%
拼写纠错+本文方法	**80.24%**	**15.66%**	**2.11%**	**79.16%**	**14.90%**	**3.41%**
词形还原+本文方法	75.74%	23.57%	2.72%	76.20%	18.00%	4.00%
拼写纠错+词形还原+本文方法	77.72%	21.23%	2.83%	76.49%	18.00%	4.00%

为了能够更为准确的比较本文的局部密度划分和 Higgins 提到的基于参考作文题目的阈值划分方法的差异，本文对所有作文数据均采用本文的主题词抽取和题目扩展以及基于本文的相似度计算方法，但在划分跑题和切题作文上，分别采用本文的局部密度方法和 Higgins 的基于参考作文题目的方法，其中基于参考作文题目的方法以排序后是否排在前 25%作为判断是否离题的标准。实验中参考题目选取了来自中国高考、雅思和中学作文三个领域各取 3 个参考题目共 9 个参考题目，这 9 个参考作文题目的详细描述如表 6 所示。

表 6 参考题目描述

题目标号	参考作文题目
Reference #1	Christmas in America
Reference #2	Health is more important
Reference #3	Protect our eyes
Reference #4	Some people think children different abilities together benefits everyone, others think intelligent children should be taught separately and given special treatment.
Reference #5	In many countries women are able to join the armed forces now on the equal basis of men. However, some people think only men should be members of the army, navy and Air Force.
Reference #6	Traditional cultures will be lost as technology develops. Technology and traditional cultures are incompatible.
Reference #7	Old man got recruited
Reference #8	Should students do the housework
Reference #9	The traffic jams in the highway

实验比较结果如表 7 所示。实验结果表明，本文提出的基于局部密度的阈值划分方法优于基于参考题目的方法，在 8 个作文数据集上的平均 F1 值要高出 6.52 个百分点。同时我们发现，Higgins 的基于参考题目的方法总体召回率较低，我们认为这是因为跑题作文的主题具有不确定性，对不同的参考题目，其相关度排名有可能高也有可能低，这是导致召回率较低的原因。

5. 结论

本文提出了一种基于作文主题词抽取和基于局部密度阈值选择的无监督作文跑题检测方法，该方法的创新之处在于根据待测作文的主题分布提取作文的主题关键词，并研究出一种基于作文切题度局部密度的阈值抽取方法动态选取跑题阈

值，该方法无需事先标注好的作文训练集，并可以适应不同的作文主题，具有很好的通用性。在多个真实数据集上的实验结果表明，该方法跑题检测性能优于传统的 tf*idf 向量表示方法。

实验中我们发现在部分数据集上效果不佳，如在数据集 Prompt #1 上的 F1 度量值只有 64.91%，我们分别计算 8 个作文数据集的作文切题度方差，发现 Prompt #1 的作文切题度方差在所有数据集中值最大，为 0.015，其他 7 个作文数据集的切题度分别为：0.008、0.010、0.001、0.005、0.006、0.008 和 0.006，这说明 Prompt #1 的作文主题发散性较大。同时本文在原始作文数据集中抽取离题作文时是针对分数最低的作文进行人工判断和抽取的，有可能那些分数不为 0 分或者分数不是最低分的作文也是离题作文，这也可能导致本文算法的结果降低。针对发散性很高的作文集合如何更有效的采用无监督阈值抽取方法并检测跑题作文是未来本文进一步需要研究和改进的方法。

表 7 基于参考题目和本文局部密度划分阈值实验结果对比

数据集	基于参考题目的划分阈值方法			基于局部密度的阈值方法		
	precision	recall	F1	precision	recall	F1
Prompt #1	94.82%	41.76%	57.98%	49.38%	97.56%	65.57%
Prompt #2	75.00%	87.70%	80.85%	89.28%	74.75%	81.37%
Prompt #3	99.50%	73.26%	84.38%	95.60%	95.60%	95.60%
Prompt #4	99.06%	54.92%	70.67%	71.24%	86.01%	77.93%
Prompt #5	100.00%	46.31%	63.31%	76.19%	72.73%	74.42%
Prompt #6	96.15%	90.00%	92.97%	96.00%	96.00%	96.10%
Prompt #7	100.00%	50.00%	66.67%	65.38%	85.00%	73.91%
Prompt #8	89.01%	55.21%	68.14%	61.90%	86.67%	72.22%
平均值	**94.19%**	**62.39%**	**73.12%**	**75.62%**	**86.77%**	**79.64%**

参考文献

[1] Burstein, J., & Higgins, D. Advanced Capabilities for Evaluating Student Writing: Detecting Off-Topic Essays Without Topic-Specific Training[J]. AIED, 2005, 112-119.

[2] Higgins, D., Burstein, J., & Attali, Y. Identifying off-topic student essays without topic-specific training data[J]. Natural Language Engineering, 2006, 12(02): 145-159.

[3] Louis, A., & Higgins, D. Off-topic essay detection using short prompt texts[C]. NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, 2010, 92-95.

[4] Persing, I., & Ng, V. Modeling Prompt Adherence in Student Essays[C]. 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), MD: Baltimore, June 2014, 1534-1543.

[5] Cummins, R., Yannakoudakis, H., & Briscoe, T. Unsupervised Modeling of Topical Relevance in L2 Learner Text[C]. 11th Workshop on Innovative Use of NLP for Building Educational Applications, California: San Diego, June 2016, 95-104.

[6] Rei, M., & Cummins, R. Sentence Similarity Measures for Fine-Grained Estimation of Topical Relevance in Learner Essays[C]. arXiv, June 2016: 1606.03144.

[7] 陈志鹏，陈文亮，朱幕华. 利用词的分布式表示改进作文跑题检测[J].中文信息学报. 2015,29(5)，PP: 178-184.

[8] 陈志鹏，陈文亮. 基于文档发散度的作文跑题检测[J]. 中文信息学报. 2017, 31(1), PP:23-30.

[9] 李晓亚,中国大学生英语作文跑题检测系统的研究与设计[D].中国科学技术大学, 2016.

[10] 范弘屹,张仰森. 一种基于 HowNet 的词语语义相似度计算方法[J].北京信息科技大学学报,2014, 26(4):42-45.

[11] 颜伟. 基于 WordNet 的英语词语相似度计算[D].北京：北京语言大学.

[12] 梁茂成. 中国学生英语作文自动评分模型构建[D].2005:1-241.

[13] 李霞,刘建达. 适用于中国外语学习者的英文作文全自动集成评分算法[J]. 中文信息学报，2013, 27(5):100-106.

[14] David M. Blei, AndrewY. Ng, Michael I. Jordan, Latent Dirichlet Allocation[J], Journal of Machine Learning Research, 2003, p993-1022.

[15] Tomas Mikolov,Kai Chen,Greg Corrado,Jeffrey Dean,Efficient Estimation of Word Representations in Vector Space[C]. arXiv, 2013:1301.3781.

[16] Tomas Mikolov,Ilya Sutskever,Kai Chen, Greg Corrado,Jeffrey Dean,Distributed Representations of Words and Phrases and their Compositionality[C]. arXiv, 2013: 1310.4546.

[17] 桂诗春，杨惠中. 中国英语学习者英语语料库[M]. 上海外语教育出版社，2003.