**文章编号**：1003-0077(2016)06-0173-09

# 基于文本语义离散度的自动作文评分关键技术研究

王耀华[1]，李舟军[1]，何跃鹰[2]，巢文涵[1]，周建设[3]

（1. 北京航空航天大学 计算机学院，北京 100191；
2. 国家计算机网络应急技术处理协调中心，北京 100029；
3. 首都师范大学 成像技术高精尖创新中心 北京 100048）

**摘　要**：该文尝试从文本语义离散度的角度去提升自动作文评分的效果，提出了两种文本语义离散度的表示方法，并给出了数学化的计算公式。基于现有的 LDA 模型、段落向量、词向量等具体方法，提取出四种表征文本语义离散度的实例，应用于自动作文评分。该文从统计学角度将文本语义离散度向量化，从去中心化的角度将文本语义离散度矩阵化，并使用多元线性回归、卷积神经网络和循环神经网络三种方法进行对比实验。实验结果表明，在 50 篇作文的验证集上，在加入文本语义离散度特征后，预测分数与真实分数之间均方根误差最大降低 10.99%，皮尔逊相关系数最高提升 2.7 倍。该表示方法通用性强，没有语种限制，可以扩展到任何语言。

**关键词**：作文评分；语义离散度；神经网络

**中图分类号**：TP391　　**文献标识码**：A

# Research on Key Technology of Automatic Essay Scoring Based on Text Semantic Dispersion

WANG Yaohua[1], LI Zhoujun[1], HE Yueying[2], CHAO Wenhan[1], ZHOU Jianshe[3]

(1. School of Computer Science and Engineering, Beihang University, Beijing 100191, China;
2. National Computer Network Emergency Response Technical Team, Beijing 100029, China;
3. Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048, China)

**Abstract**: Based on the existing methods, including LDA model, paragraph vector, word vector text, we extract four kinds of text semantic dispersion representations, and apply them on the automatic essay scoring. This paper gives a vector form of the text semantic dispersion from the statistical point of view and gives a matrix form from the perspective of decentralized text semantic dispersion, experimented on the multiple linear regression, convolution neural network and recurrent neural network. The results showed that, on the test data of 50 essays, after the addition of text semantic dispersion feature, the Root Mean Square Error is reduced by 10.99% and the Pearson correlation coefficient increases 2.7 times.

**Key words**: Automatic Essay Scoring; semantic dispersion; neural network

## 1　引言

自动作文评分(Automated Essay Scoring, AES)的研究已达 50 年之久，旨在使用特定的计算机程序给作文进行打分。近些年来，随着计算机的发展以及自然语言处理技术的进步，自动作文评分系统逐渐成为一种趋势。自动作文评分系统具有如下优势：客观性，评分结果不受人为因素的影响；效率高，机器评分具有即时性；成本低，机器评分可以节省大量的人力。

### 1.1　自动作文评分的研究现状

回顾近半个世纪以来在自动作文评分上的研

**收稿日期**：2016-07-21　**定稿日期**：2016-08-05

**基金项目**：国家自然科学基金(61170189,61370126, 61202239,U1636211)；国家 863 计划(2015AA016004,2014AA015105)；北京成像技术高精尖创新中心项目(BAICIT-2016001)

究,主要分为三类:基于文本表层特征的作文评分,基于潜在语义的作文评分,以及将前面两者进行综合的作文评分。

基于文本表层特征的作文评分,是指从语言学角度来对作文进行评分。以1968年发布的PEG(Project Essay Grade)[1]系统为代表,它是由自动作文评分之父Ellis Batten Page开发,这个系统从语言学角度抽取特征来对作文进行评分,假设作文的内容质量能通过可测量的代理指标反映。但是这个系统评价的角度过于单一,评测完全依赖于统计学方法,没有直接评测作文的内在质量,没有使用自然语言处理的技术,所以容易受到技巧上的欺骗。尽管如此,PEG系统是人类在自动作文评分上的第一次尝试,影响了后续的自动作文评分系统的研究。

基于潜在语义的作文评分,是指将作文映射到潜在语义空间再对其进行评分。Peter Foltz和Thomas Landauer开发了第一款基于潜在语义的自动作文评分系统IEA(Intelligent Essay Assessor)[2]。IEA系统借助潜在语义分析模型(Latent Semantic Analysis,LSA),构建词语的共现矩阵,借助矩阵的奇异值分解(Singular Value Decomposition,SVD)将待评分作文与人工评分后的标准作文一起映射到潜在语义空间,再使用待评分作为与人工评分的作文之间的相似度作为权重,加权得到作文的评分。这种方法充分利用了文本语义信息,具有较好的反欺骗能力,但是对于文本表层信息没有直接评价,而且LSA模型过程的可解释性不强,缺乏严谨的数理统计基础,同时会丢失语序的信息。

需要特别指出的是,本文提出的基于文本语义离散度的自动作文评分方法,虽然也基于潜在语义,但是与IEA中的方法有明显的不同:IEA使用LSA的方法构建文本的向量表示,而本文使用更高级的LDA、词向量、段落向量等多种方法构建文本的向量表示,同时还为文本中的每个句子都构建了向量表示;IEA通过比较待评分作文与人工评分作文的相似度作为权重来进行评分,而我们对文本的语义离散度构建了特征向量,然后结合词汇等级、句子等级、优美句型、篇章结构等基本特征,进行回归分析。

近年来,有人提出将前面两种方法进行综合的自动作文评分。上个世纪90年代,由美国教育考试服务中心(Educational Testing Service, ETS)开发的E-rater[3],已成为一个受到广泛关注的商业性评分系统,并已成功应用到GMAT、TOEFL、GRE等考试系统当中。它与阅卷教师一起,对每篇作文进行自动评分。由于E-rater是商业化的产品,我们无法找到太多关于E-rater的实现细节。

除了E-rater以外,还有采用分类模型的贝叶斯作文评分系统[4](Bayesian Essay Test Scoring system,BETSY),但是当作文评分所使用的特征不满足相互独立的前提假设时,评分的效果将会受到影响。我国梁茂成[5]是较早涉足自动作文评分领域的人,他对比了前人的工作,并提出:一个合理的作文自动评分系统应该充分利用统计技术、自然语言处理技术、信息检索技术及其他可能利用的技术。

### 1.2 创新点与结构安排

基于前人的研究工作,结合近年来自然语言处理领域尤其是深度学习领域中对文本建模的最新成果,本文提出了一种基于文本语义离散度的自动作文评分方法(Automated Essay Scoring based on Text Semantic Dispersion,AES-TSD)。本文的主要创新点如下:

1) 提出了基于距离的文本语义离散度的表示方法,并使用统计学的方法将其向量化表示;

2) 提出了基于中心的文本语义离散度的表示方法,并使用深度学习模型将其向量化表示;

3) 基于现有的多种文本表示进行了对比实验,训练出基于文本语义离散度的自动作文评分模型;

4) 本文提出的文本语义离散度表示方法通用性强,没有语种限制,可以扩展到任何语言。

本文的结构安排如下:第一部分介绍了自动作文评分的研究现状以及本文的创新点;第二部分介绍了文本建模的三种方法以及文本离散度表示的两种方法;第三部分介绍了用于作文评分的两种回归网络模型,并重点阐述了文本语义离散度的使用方法;第四部分介绍了实验设计与实验分析;第五部分是内容总结与下一步工作。

## 2 文本语义与文本语义离散度的表示

文本语义离散度是指文本中每个句子语义的差异程度。本节介绍了四种文本的语义表示方法和两种文本语义离散度的表示方法。本文将文本和文本中的每个句子映射到同一个潜在语义空间中,依据每个句子与文本本身在语义空间中的相对位置,表征文本语义的离散度。

### 2.1 LDA 模型

LDA(Latent Dirichlet Allocation)是一种生成式概率模型，使用不可观测的话题分布来解释可观测的文本相似度。LDA 模型认为文本中每一个词是在话题和词分布的共同作用下产生的。LDA 的概率图模型如图 1 所示。

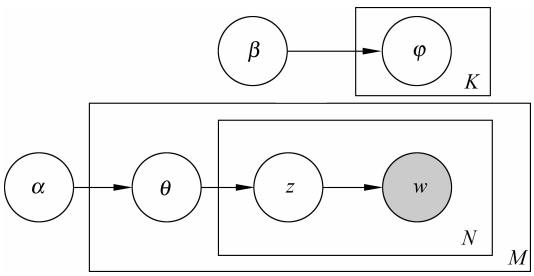


图 1　LDA 的概率图模型

其中，含阴影圆形代表可观测的变量，空白的圆形表示不可观测的变量，方框表示过程的重复，箭头表示条件依赖，$\alpha$ 和 $\beta$ 是模型中 *Dirichlet* 分布的超参数，$\theta$ 代表文本的话题分布，$\varphi$ 代表话题下的词分布，$z$ 表示每一个词的话题，$w$ 表示文本中的词，$K$，$M$，$N$ 分别代表话题个数、文本篇数和文本内的词数。

LDA 模型有两个参数需要求解：文本下的话题分布 $\theta$ 和话题下的词分布 $\varphi$。由于参数 $\theta$ 和 $\varphi$ 的耦合关系，导致难以准确估计，Griffiths[6] 提出了 Gibbs 采样的方法来近似地获取观测序列：当确定某一个维度的边缘分布时，给定其他的维度的变量值，使用该维度采样的结果来更新该维度的值，依次迭代直至收敛。在 LDA 模型中，每一步采样的概率分布可用公式(1)计算。

$$P(z_i = k \mid \vec{z}_{-i}, \vec{w}) \propto \theta_{m,k} \cdot \hat{\varphi}_{k,t} = \frac{n_{m,-i}^{k} + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k} \times \frac{n_{k,-i}^{t} + \beta_t}{\sum_{t=1}^{V} n_k^t + \beta_t} \tag{1}$$

参数可按照公式(2)获得。

$$\theta_{m,k} = \frac{n_{m,-i}^{k} + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k}, \quad \hat{\varphi}_{k,t} = \frac{n_{k,-i}^{t} + \beta_t}{\sum_{t=1}^{V} n_k^t + \beta_t} \tag{2}$$

以上公式中，$n_m^k$ 代表属于话题 $k$ 的词在文本 $m$ 中出现的次数，$n_k^t$ 代表词 $t$ 在话题 $k$ 下出现的次数，$z_i$ 代表第 $i$ 个词所对应的话题。

### 2.2 词向量

最早的词向量可以追溯到 one-hot 表示[7]，这个方法简单易行，但是向量过于稀疏，会造成形成维度灾难，不能反映两个向量之间的语义关系。词的分布式表示(Distributed Representations)是对 one-hot 表示的改进，1986 年由 Hinton[8] 首次提出。2003 年，Yoshua Bengio[9] 提出了神经概率语言模型，应用了词语的分布式表示，但是模型训练过程冗长。2013 年，Mikolov[10-11] 对此进行了改进，将词向量与神经概率语言模型的训练过程解耦合，提出高效的词向量训练方法：Word2Vec，这是目前广泛应用的方法。

Word2Vec 一共有两种训练方法，分别是 CBOW(Continuous Bag-of-Words)模型和 Skip-gram 模型，其网络结构分别如图 2 所示。

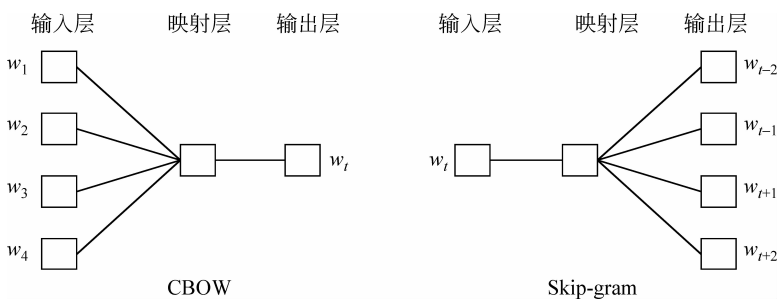


图 2　词向量的训练模型

CBOW 模型一共有三层，输入层是一个宽度固定的滑动窗口，用于捕捉局部特征，分别是 $\{w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}\}$。在映射层，将这些词向量进行拼接或平均，然后进行线性映射。在输出层，对数据进行 softmax 映射，得到 $w_t$ 的概率值 $P(w_t)$ 作为模型的输出。CBOW 模型通过极大化 $P(w_t)$ 达到模型训练的目的。

Skip-gram 模型也是一个三层的神经网络，不同之处在于输入层是一个词 $w_t$，而将上下文窗口 $\{w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}\}$ 放到输出层，在映射层依次计算 $P(w_{t+j} \mid w_t)$ 的概率之后再求和。Skip-gram 通过极大化 $w_t$ 前后 $2k$ 个词的概率和，以

达到模型训练的目的。

### 2.3 段落向量

受词向量的启发，Le 和 Mikolov[12] 提出了两种段落向量(Paragraph Vector)的训练方法，分别是 PV-DM 和 PV-DBOW 的方法。其网络结构分别如图 3 所示。

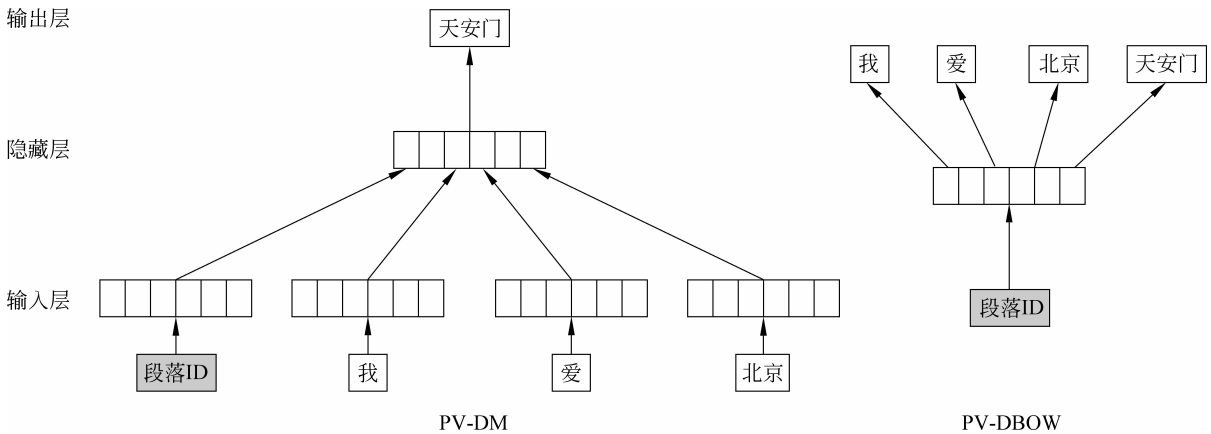


图 3 段落向量的训练模型

PV-DM 的训练过程与词向量中 CBOW 的训练方法类似，不同之处在于 PV-DM 在输入层增加了一个代表段落 ID 的向量，认为它记忆了当前上下文所缺失的语义。在训练过程中，段落向量只是段落内共享，而词向量在整个语料中共享。

PV-DBOW 的训练过程与词向量中 Skip-gram 的训练方法类似，不同之处在于网络的输入不是一个词向量，而是段落向量本身，网络的输出也不某一个词的上下文的词向量，而是该段落内随机采样的词向量。这种方法通过强制模型输出段落内词向量，而逐步调整训练段落向量的参数。

段落的范围可长可短，短为一个句子，长至一篇文本，因此段落向量的适用场景较广。

### 2.4 文本语义离散度的表示

依据上面的工作，可以将文本和文本中的每个句子映射到一个统一潜在语义空间：对于 LDA 和段落向量的方法，可以直接获取句子和文本的向量表示；对于词向量的方法，可以通过对句子、文本内词向量的算术相加或者加权(如 TF-IDF)相加的方法，获得每个句子以及文本的向量表示。基于这些工作，本文提出了两种文本语义离散度的表示方法。

对于第一种文本语义离散度的表示方法，其获取过程如下：

给定语料中的每篇文本 $m$

(1) 获取文本 $m$ 的向量表示 $v_{m0}$；

(2) 对于文本 $m$ 中的每一个句子 $s_{m,n}$，获取句子 $s_{m,n}$ 的向量表示 $v_{m,n}$；

(3) 计算每一个句子向量 $v_{m,n}$ 与文本向量 $v_{m0}$ 之间的距离向量 $d_m$。

我们称这种方法为基于距离的文本语义离散度表示方法(Text Semantic Dispersion Representation based on Distance，TSDR-D)，该方法得到的向量形式的文本语义离散度为 $d_m = [d_{m,1}, d_{m,2}, \ldots, d_{m,j}, \cdots, d_{m,n}]$，该向量中每一维度 $d_{m,j}$ 的计算公式如式(3)所示。

$$d_{m,j} = \frac{\arccos\left(\frac{v_{m,j} \cdot v_{m,0}}{\| v_{m,j} \| \cdot \| v_{m,0} \|}\right)}{2\pi} \tag{3}$$

其中，$v_{m,j}$ 表示文本 $m$ 的第 $j$ 个句子的向量，$v_{m,0}$ 表示文本 $m$ 的向量。

在第二种文本语义离散度的表示方法中，文本的向量可认为是文本语义上的中心点，以该中心点为原点，重新定义文本内每个句子的坐标，去除句子向量在不同领域、不同话题等因素偏好，获得每个句子相对于文本语义中心点的坐标。通过该过程可以得到代表文本语义离散度的矩阵 $M_m$。我们称这种方法为基于中心的文本语义离散度表示方法(Text Semantic Dispersion Representation based on Centre，TSDR-C)，由该方法得到矩阵形式的文本语义离散度为 $M_m = [M_{m,1}, M_{m,2}, \ldots, M_{m,j}, \ldots, M_{m,n}]$，该矩阵每一列 $M_{m,j}$ 的计算公式如式(4)所示。

$$M_{m,j} = v_{m,j} - v_{m,0} \tag{4}$$

其中，$v_{m,j}$ 和 $v_{m,0}$ 的含义与第一种方法中的相同。

## 3　深度学习模型与文本语义离散度的应用

上一节中提到的两种方法所获得的文本语义离散度信息仍然是原始数据。TSDR-D 的方法得到的是一个长度不确定的向量，可使用统计学方法，从原始数据中抽取信息，获得定长的向量表示，然后使用多元回归来进行自动作文评分。对于 TSDR-C 的方法得到的是一个不定长的矩阵，可使用深度学习的模型，将这些数据重新组合出一个高级特征向量，以规范化地表征文本离散度，并对其进行回归来获得作文评分。

### 3.1　卷积神经网络

卷积神经网络具有强大的特征提取能力，通过卷积、池化等操作可以将低级特征抽象为高级特征，进而有利于分类与回归，因此卷积神经网络适用于作文评分任务。

本文所使用的卷积神经网络，借鉴了 Kim[13] 的短文本分类任务中的网络结构。不同之处是：输入层，增加了基本特征，这些特征并不参与卷积等操作，而是直连到最后一层；输出层，使用了多元线性回归，而不是分类器。其网络结构如图 4 所示。

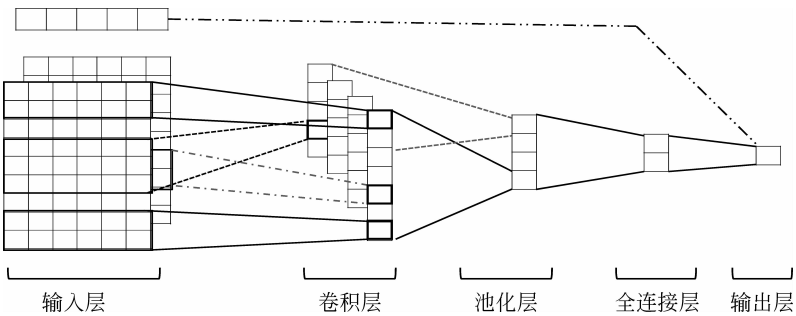


图 4　卷积神经网络模型图

本文的卷积神经网络分为五层，分别是：输入层、卷积层、池化层、全连接层和输出层。在输入层，每个句子向量为 $x_i$，$x_{i:i+h-1}$ 是 $x_i$ 到 $x_{i+h-1}$ 的向量拼接，一篇文本当中含有 $n$ 个句子，$bf$ 代表该文本的基本特征。在卷积层，一次卷积的过程会产生卷积核 $c=\{c_1,c_2,\ldots,c_{n-h+1}\}$，为捕捉不同粒度的局部信息，本文选择不同大小的过滤器，从而得到不同的特征映射(feature map)。在池化层，使用最大池化的方式来获取一个特征映射当中的最大值，然后把所有的池化结果拼接得到向量 $C$。由于在卷积层和池化层损失了部分信息，所以添加了一个全连接层做信息的修复得到 $C'$，此时的 $C'$ 就是反映文本离散度的特征向量。将 $C'$ 与文本的基本特征 $bf$ 进行拼接，一起输入到多元线性回归层中，从而得到对该文本的评分。

### 3.2　循环神经网络

循环神经网络不必要求输入数据之间独立同分布，它对信息具有一定“记忆性”，适合处理序列问题。一篇文档可以视为由句子序列组成，因此循环神经网络也适合作文评分任务。

循环神经网络当中含有环形结构，为了清楚地描述其过程，我们将循环神经网络的环按照时间打开，在本文中，使用的循环神经网络架构如图 5 所示。

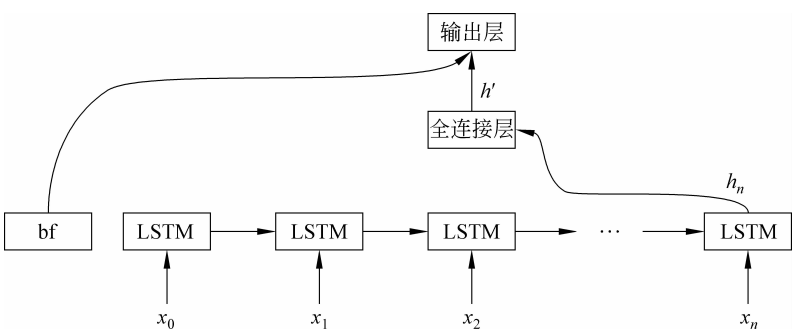


图 5　循环神经网络模型图

图5中的循环神经网络共有三层，分别是循环层、全连接层和输出层。在循环层，每个节点的输入由两部分组成：前一个节点的状态值 $s_{t-1}$ 和当前节点的输入 $x_t$。将循环层的最后一个单元的输出值 $h_n$，输入到全连接层进行信息的修复得到 $h'$，此时的 $h'$ 就是反映文本离散度的特征向量。再将 $h'$ 与文本的基本特征向量 $bf$ 进行拼接，一起输入到输出层当中进行多元线性回归，从而得到对该文本的评分。

一篇文本中句子数目较多，在时间轴上网络层次会很深，为了克服由于层次过深而导致的梯度消失和梯度爆炸现象[14]，我们引入 LSTM[15] 单元。在 LSTM 单元内部，分别有遗忘门、输入门、输出门、状态器等方式来进行信息的保持和遗忘，使得模型可以训练成功。

## 4 实验及结果分析

### 4.1 实验设计

我们收集了 495 篇具有人工评分的初中生作文，每篇作文平均含 18.8 个句子，考虑到初中生作文命题简单且这些作文均为抒情类作文，故在本实验中视为一类作文统一处理。并从词汇等级、句子等级、优美句型、篇章结构四个角度，对这些作文抽取了其基本特征，利用多元线性回归进行自动作文评分，以此作为基线方法。本文使用清华大学 THUCNews① 语料训练中文词向量，该语料包含 74 万篇新闻语料。实验系统设计如图 6 所示。

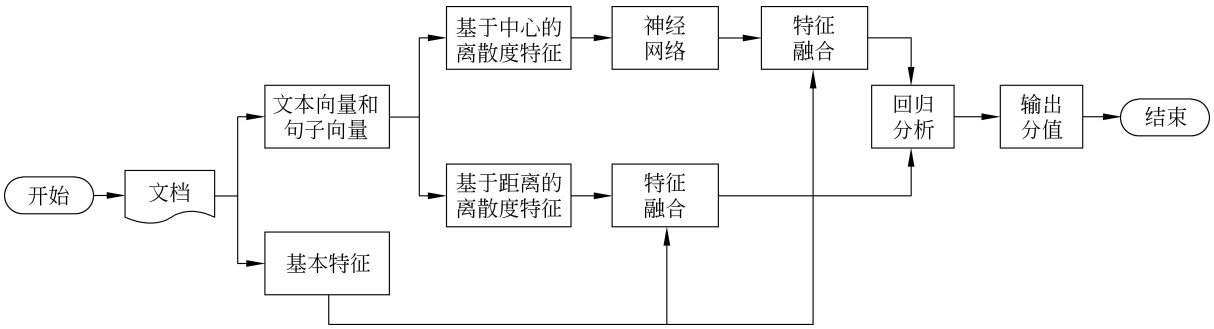


图6 系统设计图

首先对文本抽取基本特征，再对文本和每个句子向量化表示，分别获得基于中心和基于距离的文本语义离散度特征。

对于基于距离的文本语义离散度向量，本文从中抽取均值、方差、个数、最大值、最小值、变异系数、偏度、峰度等 18 个统计特征，然后对原始数据做一阶和二阶差分，再次抽取统计特征，共得到 54 个特征，由此得到规范化的文本语义离散度向量。将语义离散度特征与基本特征进行拼接，一起进行多元线性回归来预测作文的评分。

对于基于中心的文本语义离散度矩阵，本文分别使用卷积神经网络和循环神经网络来进行特征的抽取，卷积神经网络对语义离散度矩阵进行不同粒度的卷积，抽取不同范围的局部特征，然后经过池化等操作，获得向量化的文本语义离散度表示。循环神经网络按照句子出现的次序，依次输入每一个去中心化后的句子向量，捕捉句子间的时序信息，获得向量化的文本语义离散度表示。在神经网络倒数第二层，将语义离散度向量与基本特征进行拼接，一起输入到多元线性回归层，使用梯度下降的方式调整网络的参数，最终获得作文评分的模型。

需要特别指出的是，基于中心的文本语义离散度的方法，本质也是多元线性回归。深度学习的引入仅仅是为了将语义离散度矩阵转换为语义离散度向量化，以便统一地进行多元线性回归。本实验中，在神经网络网络的最后一层，将抽象后的文本语义离散度向量与基本特征拼接，一起进行了多元线性回归。由此可以说明：引入神经网络的目的是为了保留更多的离散度信息，增强了文本语义离散度特征。而实验效果的提升，从根本上来说，得益于文本语义离散度特征的引入。

本文使用均方根误差(Root Mean Square Error，RMSE)来评价回归拟合的效果，但是由于阅卷者本身具有给高分或者低分的倾向，为了去除阅卷者本身给分偏好的因素，本文还使用了皮尔逊相关系数(Pearson's correlation coefficient，R)来评价预

① http://thuctc.thunlp.org/

测分数与人工打分之间的相关度。因此，模型的损失函数如式(5)和(6)所示。

$$Loss = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}} + \lambda_1 * sign(R) * R^{-1} - \lambda_2 * (1 - sign(R)) * R + \lambda_3 * \| w \| \quad (5)$$

$$R = \frac{\sum_i (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} * \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (6)$$

在训练过程中，本文综合了 RMSE、R 以及模型复杂度三者，作为损失函数来进行优化。其中 sign($x$)为取符号函数，当 $x$ 大于 0 时为 1，否则为 0。

在神经网络的训练中，为了防止过拟合，本文使用了 dropout layer[16]的技巧。为了加快收敛速度，卷积神经网络的卷积层的非线性函数使用了 ReLU 函数。为了防止每层的输入数据方差随着该层的大小而扩散，对于使用 ReLU 函数的卷积层，其参数初始化遵循 He 等人[17]提出的规则。

### 4.2　实验分析

为了避免随机性，本文在数据集上进行十折交叉验证，以产生不同的训练集和测试集，然后将本文提出的方法与基线方法进行对比实验。本实验中向量取 100 维，训练集上使用 mini-batch 的方法调参，每个 batch 的大小为 5，在验证集中直接预测 50 篇作文的分数，神经网络的实验使用每篇文档的前 20 个句子作为输入。为了避免分数预测偏离评分范围，对预测分数进行了归一化处理如式(7)所示。

$$score' = FullScore * sigmoid\left(\frac{score - \overline{score}}{s/\sqrt{n}}\right) \quad (7)$$

其中，$\overline{score}$ 为这批作文分数的平均值，$s$ 为这批作文分数的样本标准差，$n$ 为这批作文的个数。按照本文提出的方法进行自动作文评分，实验结果如表 1 所示。

**表 1　实验结果对比表**

| 方法 | 多元线性回归 | | 卷积神经网络(CNN) | | 循环神经网络(RNN-LSTM) | |
|---|---|---|---|---|---|---|
| | ΔRMSE/% | ΔR/% | ΔRMSE/% | ΔR/% | ΔRMSE/% | ΔR/% |
| LDA | −5.12 | −66.26 | 6.17 | 177.27 | 2.48 | 99.88 |
| 段落向量 | 4.14 | 115.58 | 4.14 | 146.81 | −5.94 | −48.68 |
| 词向量权重加和 | −5.64 | −45.05 | 2.48 | 197.72 | 10.99 | 270.30 |
| 词向量算术加和 | 2.33 | 74.05 | 3.84 | 202.46 | 3.91 | 78.62 |

表 1 中 ΔRMSE 是本实验相对于基线方法均方根误差的下降值，ΔR 是本实验相对于基线方法皮尔逊相关系数的提升值。本文使用 t 检验来测试 RMSE 和 R 分布是否明显不同于基线方法的实验值。对于 RMSE 和 R，本文的假设如式(8)和(9)所示。

$$H_0: \overline{RMSE} = RMSE_{\mu_0} \quad H_1: \overline{RMSE} \leqslant RMSE_{\mu_0} \quad (8)$$

$$H_0: \bar{R} = R_{\mu_0} \quad H_1: \bar{R} \geqslant R_{\mu_0} \quad (9)$$

其中 $RMSE_{\mu_0}$ 和 $R_{\mu_0}$ 为基线方法得到的值。检验结果如表 2 所示。

**表 2　实验结果的 t 检验表**

| 方法 | p 值(RMSE) | p 值(R) |
|---|---|---|
| 多元线性回归 | 0.70 | 4.30e-07 |
| 卷积神经网络(CNN) | 0.01 | 1.06e-08 |
| 循环神经网络(RNN-LSTM) | 0.47 | 1.42e-06 |

由表 2 可以看出：在 5%的显著性水平下，本文提出的三种方法中，卷积神经网络所对应的 RMSE 具有显著降低，而使用多元线性回归和循环神经网络的方法并没有得到明显的降低，但是三种方法在 R 指标上具有显著提升。由此可以证明，基于文本语义离散度的特征是明显的，它增强了预测分数与真实分数之间的相关度。

进一步分析表 1，可以看出：即使不引入深度学习模型，而只采用多元线性回归，段落向量和词向量的算术加和也能取得良好的效果，R 值分别提升了 115.58%和 74.05%，RMSE 分别降低了 4.14%和 2.33%。这个效果的提升直接说明，引入了文本语义离散度的特征，使得 RMSE 明显降低，R 值明显提升。

在多元回归的基础之上，深度学习模型的引入进一步增强了文本语义离散度的特征。在卷积神经网络上，四种向量均取得良好效果，LDA 的方法取得了最好的 RMSE 值，降低了 6.17%，词向量的算

术加和取得了最好的 R 值，相比基线提升了 202.46%。这个效果的提升除了得益于基于中心的文本语义离散度表示方法保存了更多的信息之外，也得益于卷积神经网络的强大的特征处理能力。在循环神经网络上，词向量的权重加和取得了最好的效果，RMSE 值降低了 10.99%，R 值提升 270.30%。这个效果还得益于循环神经网络对序列的强大处理能力。

为更进一步说明文本语义离散度特征的有效性，本文以分数的中位数作为边界，区分出低分与高分作文，使用 PCA 的方法将离散度向量降低到二维平面，如图 7 所示。

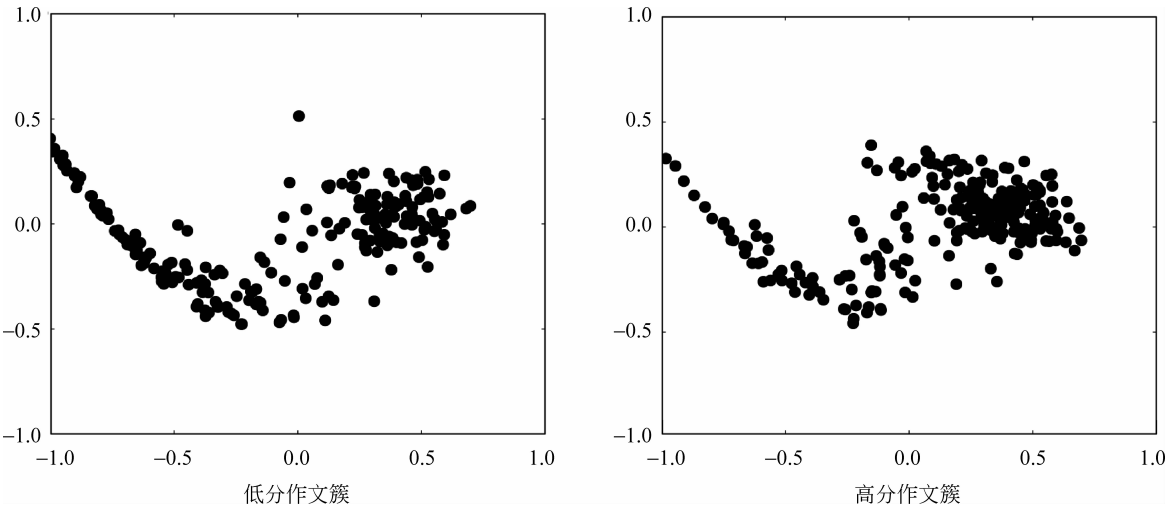


图 7 语义离散度可视化图

由此可以看出，在降维损失部分信息之后，右侧高分作文簇相对于左侧低分作文簇仍表现出显著不同：高分作文簇的数据点在右侧更加聚集。

为了定量说明这个问题，本文使用 PCA，分别将低分与高分作文的文本语义离散度向量降低到一维，然后使用 Welch's t-test 的方法进行假设检验，如式(10)所示。

$$\begin{aligned} H_0 &: \overline{Dispersion_1} = \overline{Dispersion_2} \\ H_1 &: \overline{Dispersion_1} \neq \overline{Dispersion_2} \end{aligned} \tag{10}$$

其中 $\overline{Dispersion_1}$ 和 $\overline{Dispersion_2}$ 是高低分作文的平均值。得到假设检验的 p 值为 2.35e-11。在 5%的显著性水平下，高分作文和低分作文的离散度显著不同。由此验证了基于文本语义离散度的作文评分的方法是有效的。

## 5 结束语

本文提出了两种文本语义离散度的表示方法，并给出了其求解流程、计算公式和使用方法。同时，使用 LDA、词向量、段落向量等方法进行多种对比实验，以验证文本语义离散度信息对自动作文评分效果的影响。实验结果表明：加入文本语义离散度特征后，预测分数与真实分数之间的皮尔逊相关系数有显著提升，说明该特征是有效的。

下一步的研究工作包括：研究文本语义离散度矩阵在更深层次的卷积神经网络中的特征提取效果；结合卷积神经网络的特征提取与循环神经网络的记忆性的优势，组合出新型网络来降低预测的误差；深入分析不同维度的文本向量表示，对自动作文评分效果的影响。

## 参考文献

[1] Page E B. Project essay grade: PEG[J]. Automated essay scoring: A cross-disciplinary perspective, 2003: 43-54.

[2] Hearst M A. The debate on automated essay grading [J]. Intelligent Systems and their Applications, IEEE, 2000, 15(5): 22-37.

[3] Valenti S, Neri F, Cucchiarelli A. An overview of current research on automated essay grading[J]. Journal of Information Technology Education, 2003, 2: 319-330.

[4] Rudner L M, Liang T. Automated essay scoring using Bayes' theorem [J]. The Journal of Technology, Learning and Assessment, 2002, 1(2):3-18.

[5] 梁茂成，文秋芳. 国外作文自动评分系统评述及启示[J]. 外语电化教学，2007，05：18-24.

[6] Steyvers M, Griffiths T. Probabilistic topic models

[J]. Handbook of latent semantic analysis, 2007, 427(7): 424-440.

[7] Harris, David and Harris, Sarah. Digital design and computer architecture (2nd ed.)[M]. San Francisco, Calif.: Morgan Kaufmann. Elsevier, 2012: 129.

[8] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.

[9] Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]. Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.

[10] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the Advances in neural information processing systems. 2013: 3111-3119.

[11] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of ICLR Workshop. 2013.

[12] Le Q V, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of ICML, 2014.

[13] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of EMNLP, 2014.

[14] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. Neural Networks, IEEE Transactions, 1994, 5(2): 157-166.

[15] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[16] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[17] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1026-1034.

王耀华(1991—)，硕士研究生，主要研究领域为自然语言处理。
E-mail: wangyaohua@buaa.edu.cn

何跃鹰(1975—)，通信作者，高级工程师，主要研究领域为数据挖掘与网络安全。
E-mail: hyy@cert.org.cn

李舟军(1963—)，教授，主要研究领域为数据挖掘与信息安全。
E-mail: lizj@buaa.edu.cn