文章编号：1003-0077(2019)00-0134-09

基于代表性答案选择与注意力机制的短答案自动评分

谭红叶[1]，午泽鹏[1]，卢 宇[2,3]，段庆龙[2]，李 茹[1]，张 虎[1]

(1. 山西大学 计算机与信息技术学院，山西 太原 030006；
2. 北京师范大学 未来教育高精尖创新中心，北京 100875；
3. 北京师范大学 教育技术学院，北京 100875)

摘 要：短答案自动评分是智慧教学中的一个关键问题。目前自动评分不准确的主要原因是：(1)预先给定的参考答案不能覆盖多样化的学生答题情况；(2)不能准确刻画学生答案与参考答案匹配情况。针对上述问题，该文采用基于聚类与最大相似度方法选择代表性学生答案构建更完备的参考答案，尽可能覆盖学生不同的答题情况；在此基础上，利用基于注意力机制的深度神经网络模型来提升系统对学生答案与参考答案匹配情况的刻画。相关数据集上的实验结果表明：该文模型有效提升了自动评分的准确率。

关键词：短答案自动评分；代表性答案；参考答案；注意力机制；神经网络

中图分类号：TP391 **文献标识码**：A

Using Representative Answers and Attentions for Short Answer Grading

TAN Hongye[1], WU Zepeng[1], LU Yu[2,3], DUAN Qinglong[2], LI Ru[1], ZHANG Hu[1]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
2. Advanced Innovation Center for Future Education, Beijing Normal University, Beijing 100875, China;
3. School of Education Technology, Beijing Normal University, Beijing 100875, China)

Abstract: Automatic short answer grading (ASAG) is a key issue in intelligent tutoring systems. The main challenges in ASAG lie in 1) the reference answer for a given question cannot cover the diverse student answers; and 2) the similarity between student answer and the reference is hard to estimate. This paper applies clustering and maximum similarity to select representative answers, constructing the reference answer set to cover various student answers. Then, this paper employs a deep neural network model based on the attention mechanism to approximate the similarity between the student answer and the reference answer set. Experimental results show that the proposed model effectively improves the accuracy of automatic scoring.

Keywords: automatic short answer grading; representative student answers; reference answer; attention mechanism; neural network

0 引言

人工智能技术在教育领域的应用得到国家和政府的高度重视。2017 年国务院推出的《新一代人工智能发展规划》明确指出：利用智能技术加快推动人才培养模式、教学方法改革，构建包含智能学习、交互式学习的新型教育体系。2018 年教育部印发《教育信息化 2.0 行动计划》，将“智慧教育创新发展行动”列为推动教育信息化 2.0 发展的“八大行动”之一。这充分说明：教育将全面进入智慧教育时代。智慧教育主要包括教育大数据挖掘、教育知识图谱构建、教学过程智慧化、智慧教育平台建设等技术[1]。自动评分是智慧化教学过程中评价教学质量的一个重要环节。该任务通过一定模型与算法对学生答案预测分值，不仅能减轻教师工作量，还能避免因主观性引起的评分不一致问题，是智慧教学中的一个研究热点。

收稿日期：2019-05-14 **定稿日期**：2019-07-01

基金项目：国家自然科学基金(61673248,61772324)；国家社会科学基金(18BYY074)

自动评分主要包括作文自动评分(automatic essay grading，AEG)与短答案自动评分(automatic short answer grading，ASAG)[2]。其中，AEG 侧重质量评价，一般需要从立意、结构、文采等方面对作文形成整体评分；而短答案问题主要考察学生对特定知识(如：科学概念或原理)的理解，因此 ASAG 更注重评价学生答案的正确程度。本文主要研究 ASAG 技术。

ASAG 一般通过一定模型对学生答案与参考答案进行对比，根据匹配程度预测其分值。随着机器学习技术的发展，该任务已取得一定进展，但准确率仍需进一步提高。目前导致 ASAG 不准确的原因主要有两方面：①学生答案多种多样，预先给定的参考答案难以覆盖所有可能答题情况。此外，还存在许多开放程度高、不具有单一明确参考答案的问题。例如，对于表 1 示例 1 的问题，学生可以从故事情节、人物性格、写作特点、语言风格等角度来回答。然而现实中对这类问题要么不提供参考答案，要么提供非常有限的参考答案，因此引起学生答题情况覆盖度不高、自动评分不准确问题；②模型不能准确刻画学生答案与参考答案匹配情况。如表 1 示例 2 中，模型需要对二者进行语义匹配，才能得到学生答案中“身高差不多的人数最多”与参考答案中“人数最为集中，且大家的身高相对接近”语义一致的判断。

表 1　短答案题目示例

序号	题目示例
1	语文题：《海底两万里》是一部纯虚构的小说，这部书最吸引你的地方是什么？(2 分) 参考答案：海底世界充满异国风情和浓厚的浪漫主义色彩以及曲折的情节和对海洋知识的介绍都深深吸引我。 学生答案 1：书中人物鲜明生动，故事情节具逻辑性，但最吸引我的是船长尼摩这人物，性格阴郁，蜇居海底，却又与陆地上的一些人有特殊联系，而且故事里他们在海底历险的几个情节使我感到刺激。(2 分) 学生答案 2：最吸引我的地方就是那惊险又动人的探险过程，他们搁浅，被土著人围攻，与大鲨鱼搏斗，与大章鱼对抗，还有被冰层困住的情节，无一不展示着他们的机智勇敢，让我流连忘返。(2 分)
2	数学题：下表是初二年级 50 名同龄女生身高数据。为了参加广播操比赛，老师打算从以上 50 名女生中挑选出 30 名参赛队员。为了让参赛队员的身高比较整齐，老师应该选择身高在什么范围内的同学呢？请写出答案并简述理由。(2 分) 身高/cm：146 151 153 154 156 157 158 159 160 人数：1 2 2 2 3 4 8 4 4 身高/cm：161 162 163 164 165 166 167 169 人数：2 4 3 2 3 4 1 1 参考答案：老师可以在 155～165 的身高范围内挑选队员. 因为在此范围内，人数最为集中，且大家的身高相对接近。 学生答案：老师应该选身高在 155～165cm 之间的同学。因为他们所占比例最大，频率最高有 38%，身高差不多的人数最多。(2 分)

本文采用基于聚类与最大相似度方法选择代表性学生答案重新构建更完备的参考答案，尽可能覆盖学生不同答题情况。在此基础上，本文提出基于注意力机制(attention)的深度神经网络自动评分模型，提升系统对学生答案与参考答案匹配的准确刻画。相关数据集实验结果表明：本文模型有效提升了自动评分准确率。

1　相关研究

1.1　短答案自动评分研究现状

Page[3]于 1966 年开始针对自然语言形式的答案进行自动评分研究。自此，研究者围绕作文或短答案自动评分进行研究并取得了一定进展。其中，ASAG 的方法主要有三类：①基于规则方法[4-6]。例如，Bachman 等[4]根据参考答案生成正则表达式规则，每条规则与一个分数相关联，当学生答案与规则相匹配就获得对应分数。由于规则获取精度与表达能力有限，因此该方法泛化能力较差。②基于传统机器学习方法，利用一定特征基于分类或回归模型预测分数[7-9]。例如，Sultan 等[7]使用基于对齐或嵌入式的文本相似度、词项权重等特征构建了随机森林分类器，在 SemEval-2013 评测数据集 SCIENTSBANK 上获得 55% 的 F 测度值。③基于深度学习方法，无需人工设计特征，通过对数据进行表

示学习，实现端到端的训练与输出[10-11]。例如，Riordan等[11]使用CNN与LSTM构成的神经网络进行自动评分，获得的效果比非神经网络方法好。

参考答案是自动评分的重要依据，对ASAG系统性能有重要影响，但目前对参考答案构建的深入研究还较少。Marvaniya等[12]通过对人工评分的学生答案进行聚类、选择与排序，获得各个分数对应的代表性答案来构成参考答案。实验表明，重新构造的参考答案显著提高了短答案评分的性能。还有研究关注如何处理学生答案以加速或简化评分过程。例如，Lan等[13]对学生答案聚类后，在每个簇中选择代表性样例让专家评分，然后再为同簇其他样例自动评分。

本文构建参考答案后进行自动评分的思想很大程度受到了文献[12]与文献[13]的启发。但本文与这些文献的主要不同是：①选择代表性答案的方法不同。文献[12]认为参考答案应该具有长度较短、句法结构良好的特点，并按照该特点选择代表性答案。由于判断学生答案是否正确的依据不是长度或句法结构，而是看其是否包含标准答案所需的关键概念，因此本文没有对候选样例进行显式建模，而是采用基于最大相似度方法选择代表性答案；②目的不同。文献[13]选择代表性答案目的是减少专家评分工作量，而本文目的是构造更完备参考答案以覆盖更多的学生答题情况；③自动评分模型不同。文献[12]基于参考答案与学生答案的比较特征训练多元逻辑回归分类器进行自动评分。文献[13]依据每个簇中的人工评分样例，采取同簇同分数的策略对未评分样例进行评分(或利用样例属于该簇概率调整分数)。本文是通过神经网络模型引入注意力机制来捕获参考答案与学生答案的匹配信息进行打分。

1.2 注意力机制研究现状

Bahdanau等人最早将注意力机制(attention mechanism，AM)引入基于编码器-解码器框架的神经机器翻译系统[14]，解决输入与输出不能对齐的问题。从此研究者针对各种NLP任务提出不同的AM方式，并取得很好的效果。因此，AM目前已成为神经网络架构中一个重要概念。

AM可以从以下几个视角进行分类[15]：①按照AM是否捕获多个输入之间的关系，分为互注意力与自注意力(self-attention)机制。前者用来捕捉多个输入之间关系；后者用来学习同一输入序列中词语之间关系；②按照AM包含的层次，分为单层注意力(single-level attention)与多层注意力(multi-level attention)机制。其中多层AM用来获取输入的层次结构信息。如：文本存在词语、句子、篇章层次结构；③按照AM计算上下文向量所需信息量，分为全局注意力(global-attention)与局部注意力(local-attention)机制。全局AM是使用输入序列所有隐藏状态的加权平均值来构建上下文向量；局部AM是在输入序列的关注点周围选择一个窗口来创建上下文向量。

本文针对ASAG任务需要比较学生答案与参考答案的特点，使用互注意力机制来捕获两者之间的关系。

2 基于代表性答案构建参考答案

由于学生答案具有多样性，预先给定的参考答案难以覆盖学生所有可能答题情况。针对该问题，本文尝试构建更完备的参考答案，增强对答题情况的覆盖能力。

如图1所示，本文构建参考答案包括两个步骤：(1)基于聚类获取学生可能的答题情况。我们认为：聚类后得到的每一个簇代表学生的一种答题情况。(2)在每个簇中选择一个或多个代表性答案作为本簇代表构建参考答案。

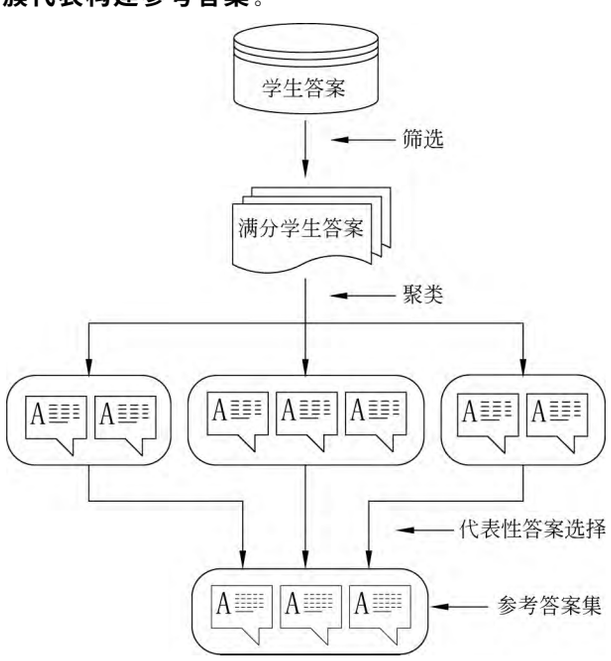

图1 参考答案构建示意图

2.1　学生答案聚类

本文采用 k-means 聚类方法对学生答案聚类。其核心思想是：对数据集 $D=\{x_1,x_2,\cdots,x_m\}$，考虑所有可能的 k 个簇集合，目标是找到一个簇集合 $\{C_1,C_2,\cdots,C_k\}$，使得每一个样本到其对应簇的中心的距离的平方和 E 最小，具体如式(1)所示。

$$E=\sum_{i=1}^{k}\sum_{x_j\in C_i}|x_j-\mu_i|^2 \tag{1}$$

其中，$\mu_i=\frac{1}{|C_i|}\sum_{x_j\in C_i}x_j$ 是簇 C_i 的中心。平方距离刻画了簇内样本与簇中心向量的相似程度。平方距离越小，簇内样本相似程度越高。

聚类质量的评价指标有外部指标和内部指标。外部指标是计算聚类结果与已有标准分类结果的吻合程度。内部指标是利用数据集的固有特征来评价一个聚类质量。本文没有对满分答案进行预先分类，因此采用内部指标评价聚类效果。轮廓系数(silhouette coefficient)是一种常用的内部评价指标，一般按照式(2)计算。

$$s_i=\frac{b_i-a_i}{\max\{a_i,b_i\}} \tag{2}$$

其中，s_i 表示簇中第 i 个样本的轮廓系数，a_i 表示簇中第 i 个样例到其他样例的平均距离，b_i 表示簇中第 i 个样例到其他样例的最小距离。轮廓系数取值范围为$[-1,1]$。当簇内样本距离越近及簇间样本距离越远，其值越大，聚类效果越好。

2.2　代表性答案的选择

代表性答案指能够代表学生答题情况的答案。本文选择簇内与其他样例相似度最高的样例作为代表性答案。该策略基于的假设是：与其他样例越相似的样例越能代表簇成员。关键问题是如何计算相似性。

由于簇内样本相似度已经较大，继续使用聚类过程中的相似度(距离)计算公式，将不能很好区分簇内样本之间的差异。本文从词语及句子长度特征出发，使用词重叠度、句子长度相似度来计算簇内样本 x_1 与 x_2 的相似度，按照式(3)进行计算：

$$\mathrm{Sim}(x_1,x_2)=\alpha_1\frac{2L_{\mathrm{overlap}}}{(L_1+L_2)}+\alpha_2\left(1-\left|\frac{L_1-L_2}{L_1+L_2}\right|\right) \tag{3}$$

其中，α_1、α_2 为权重参数，二者之和为 1，本文具体取值 0.5；L_{overlap} 表示学生答案 x_1 与 x_2 之间的重叠词个数；L_1 和 L_2 分别表示 x_1 和 x_2 的词数。

对簇内任意两个答案计算相似度后得到相似度矩阵 $M\in R^{n\times n}$，其中 m_{ij} 表示第 i 个答案 x_i 与第 j 个答案 x_j 的相似度。本文用 v_i 表示答案 x_i 对簇代表的程度，$v_i\in[0,1]$，v_i 越大表示答案 x_i 的代表性越强。按照式(4)计算 v_i：

$$v_i=\frac{1}{n}\sum_{j=1}^{n}m_{ij} \tag{4}$$

3　模型

自动评分任务可形式化为：给定参考答案 r，学生答案 s，按照式(5)预测评分结果 g。

$$g=\mathrm{argmax}P(g\mid r,s) \tag{5}$$

本文采用基于注意力机制的神经网络模型(Att-Grader)进行自动评分。模型结构如图 2 所示，由编码层、注意力层、输出层三部分构成。编码层的输入为学生答案与第 i 个参考答案 r^i(下文简写为 r)，该层对两者进行编码，生成包含语义信息的向量集合；注意力层负责获取参考答案与学生答案之间的匹配信息；输出层利用 CNN 进一步获取局部特征并经过计算后，输出学生答案预测分值。

3.1　编码层

编码层负责对输入的学生答案 $s=\{s_t\}_{t=1}^{N}$ 和参考答案序列 $r=\{r_t\}_{t=1}^{M}$，中的单词进行编码。r_t、s_t 分别表示参考答案与学生答案中第 t 个词，M、N 分别表示 r、s 词数。

将输入序列中的单词用词向量表示后得到：$R=\{w_t^r\}_{t=1}^{M}$，$S=\{w_t^s\}_{t=1}^{N}$，其中 $w\in\mathrm{R}^d$，d 表示词向量的维度。之后，将这些词向量序列输入到双向 LSTM 中进行编码，具体如式(6)、式(7)所示。

$$u_t^r=\mathrm{BiLSTM}(u_{t-1}^r,w_t^r) \tag{6}$$

$$u_t^s=\mathrm{BiLSTM}(u_{t-1}^s,w_t^s) \tag{7}$$

其中，u_t^r 表示参考答案中第 t 个词经过编码之后的表示，u_t^s 表示学生答案中第 t 个词经过编码之后的表示。

3.2　注意力层

这里使用双向注意力机制获取学生答案和参考答案匹配信息。该层输入是编码层的输出，即学生答案和参考答案上下文向量表示 $U^s=\{u_t^s\}_{t=1}^{N}$，$U^r=\{u_t^r\}_{t=1}^{M}$。

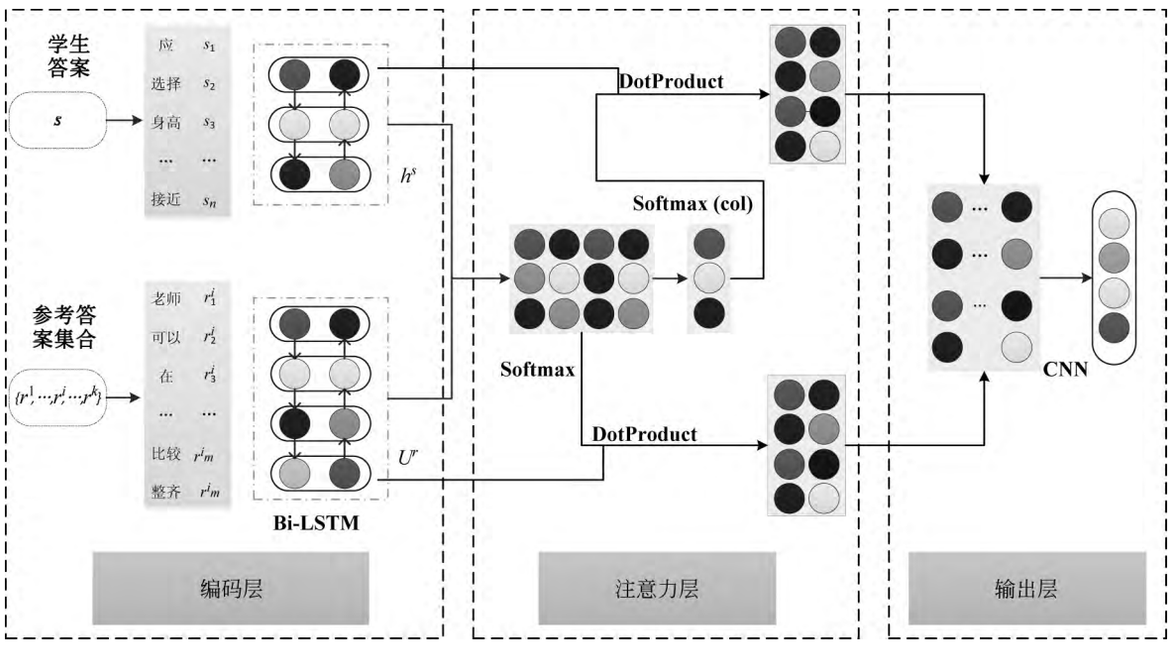

图2 Att-Grader模型

计算注意力权重时需要共享由学生答案(U^s)与参考答案(U^r)导出的相似度矩阵$C\in \mathrm{R}^{M\times N}$,其中$C_{ij}$表示参考答案中第$i$个词与学生答案中第$j$个词之间的相似度值,具体计算如式(8)所示。

$$C_{ij}=\alpha(u_i^r,u_j^s) \tag{8}$$

其中,α是一个可训练的标量函数,对两个输入向量之间的相似性进行编码;u_i^r是U^r的第i个列向量,u_j^s是U^s的第j个列向量。

在相似矩阵C的基础上计算双向注意力权重。

(1) 学生答案到参考答案的注意力权重(student-to-reference attention,S2R)。首先计算学生答案中的词与参考答案中第i个词的注意力权值为$a_i\in \mathrm{R}^M$,计算公式如式(9)所示。

$$a_i=\mathrm{softmax}(C_{i,})\in \mathrm{R}^M \tag{9}$$

然后计算学生答案中每个注意力向量,如式(10)所示。

$$\overline{U}^s_{:,t}=\sum\nolimits_j a_{tj}U^r_{:,j}。其中\ \overline{U}^s\in \mathrm{R}^{d\times N} \tag{10}$$

其中,包含所有学生答案注意力向量的矩阵。

(2) 参考答案到学生答案注意力权重(reference-to-student attention,R2S),表示参考答案中的词与学生答案中每个词的相关度。计算方法与S2R类似,最后得到$\overline{U}^r$。

将上述两个注意力向量拼接得到$G=\beta(\overline{U}^r,\overline{U}^s)$,其中每个列向量表示学生答案中的词与参考答案匹配的信息,β是将$\overline{U}^r$和$\overline{U}^s$按列拼接的函数。

3.3 输出层

经过编码层和注意力层之后,初始化输入r、s被转换成矩阵G,$G=\{g_t\}_{t=1}^N$,其中g_t表示学生答案第t个词与参考答案的匹配信息。输出层通过卷积神经网络(CNN)[16]进一步捕捉局部信息,并生成最终预测向量F,具体计算如式(11)所示。

$$F=\begin{bmatrix}P(y=0)\\P(y=1)\\\cdots\\P(y=K-1)\end{bmatrix} \tag{11}$$

其中,K表示分值种类,$P(y=k)$表示学生答案在对应分值上的概率,其中$k=0,1,\cdots,K-1$。

本文使用Adam优化算法[17]来最小化训练数据上的交叉熵损失函数[18]。损失函数如式(12)所示。

$$C(g',g)=-\frac{1}{n}\sum_{i=1}^{n}[g_i\ln g'_i+(1-g_i)\ln(1-g'_i)] \tag{12}$$

其中,n表示训练集中学生答案的数量,g'_i和g_i分别为真实分值和预测分值,然后计算出交叉熵。

4 实证研究

4.1 实验数据

本文数据集来自某中学八年级期末考试试题及学生答卷,涉及数学、语文两门课程。数据集中有2道数学题、3道语文题,分别对应表2中的

MATH1、MATH2 与 CRCC1 到 CRCC3。其中，数学题是针对特定知识点的问答题，语文题为阅读理解问答题，两者相比，语文题对应的学生答案多样性程度更高。数据集具体信息如表 2 所示，学生答案均经过两位教师人工打分，QWKappa（QWK）值反映了两个评分者评分一致性。

实验中利用 80％数据作为训练集，20％作为测试集。

表 2　数据集信息表

	Subject	Num. of Samples	Avg Word Num.	Score	QWK
MATH1	数学	4 250	22.3	0-2	*
MATH2	数学	4 913	9.4	0-2	*
CRCC1	语文	2 579	39	0-2	0.984 7
CRCC2	语文	2 571	33	0-2	0.972 3
CRCC3	语文	2 382	226	0-3	0.942 7

4.2　实验设置

预处理。利用 jieba 分词工具包进行分词并去除停用词。针对低频词（词频＜2）使用字符＜UNK＞代替。

聚类算法选择。本文比较了 Birch（balanced iterative reducing and clustering using hierarchies）聚类方法和 k-means 聚类方法。两者都通过机器学习工具包 sklearn 实现。Birch 算法是一个综合的层次聚类算法，采用聚类特征和聚类特征树进行聚类描述。我们在聚类簇数 K 分别为[3，6，9，12，15，18，21，24，27]的情况下比较了这两种方法在语文数据集上的聚类效果，采用轮廓系数的平均值为评价指标。具体如图 3 所示。

由图 3 可知，k-means 聚类算法在语文数据集上（CRCC1-CRCC3）的轮廓系数均值远远高于 Birch 算法，因此本文选择 k-means 作为聚类算法对学生答案进行聚类。

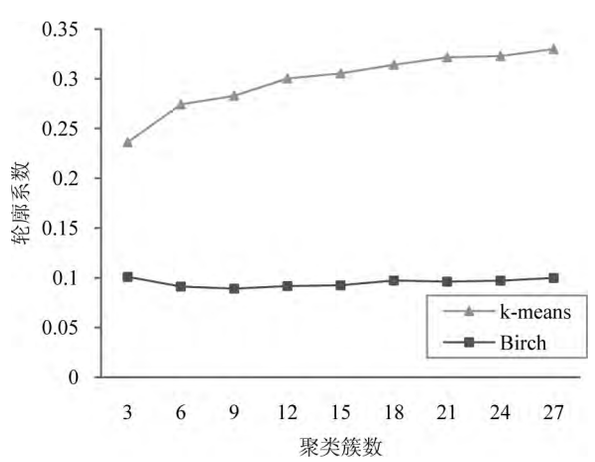

图 3　聚类方法的比较

聚类簇数的选择。由图 4 可知，k-means 算法在数据集 CRCC 1、CRCC2、CRCC3 上的聚类簇数 K 分别为 6、12、3 时，轮廓系数值较高，表明聚类效果较好。

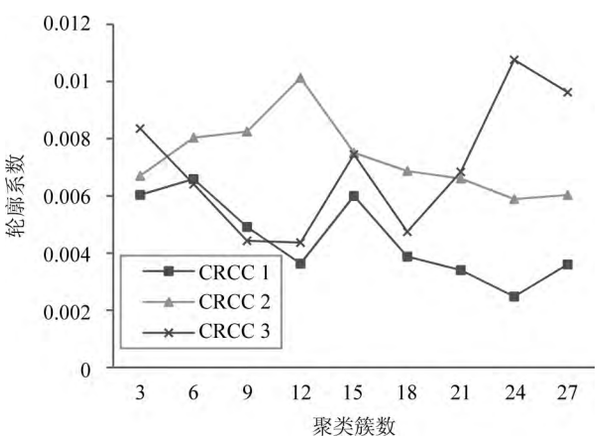

图 4　不同聚类簇数下的聚类效果

相关参数。实验中采用的长短时记忆模型 LSTM、卷积神经网络 CNN 均利用 Tensorflow 深度学习框架实现，Mini-batch 值为 32，学习率为 0.001。每次迭代过程中损失函数为交叉熵损失函数，优化器选择 Adam-optimizer。词向量利用 Gensim 工具包中 Word2Vec 工具构建，词向量维度设置为 400。

4.3　实验结果分析

4.3.1　自动评分结果分析

本文选择以下模型作为 baseline：

KAGrader　该模型由 Yang 等提出[19]，使用了连续词袋模型（CBOW）与长短期记忆模型（LSTM），在汉语短答案自动评分任务上取得了很

好的效果。该模型采用的数据集与本文有重叠。

CNN CNN[16]是目前应用范围最广的神经网络模型之一，许多学者利用其解决多种NLP任务并获得很好效果[20-22]。

LSTM＋CNN LSTM属于递归神经网络(RNN)，也是一种主要的深度神经网络结构[23-24]。对于CNN与LSTM已有研究表明[25]：CNN擅长提取位置不变特征，LSTM擅长对序列中的单元建模，两种模型可以为NLP任务提供相互补充的信息。因此本文将LSTM与CNN结合作为基线系统，该模型恰好是本文Att-grader模型没有使用attention的版本。实验中将学生答案向量作为输入，经过LSTM处理后，再经过CNN处理，完成自动评分任务。

本文使用准确率(Acc)、QWKappa(QWK)作为评价指标，采用5重交叉验证的均值作为最终结果。具体自动评分结果如表3所示。

表3 自动评分实验结果表

	Math 1		Math 2		CRCC 1		CRCC 2		CRCC 3	
	Acc/%	QWK	Acc/%	QWK	Acc/%	QWK	Acc/%	QWK	Acc/%	QWK
KAGrader	73.60	0.517 0	91.24	0.871 0	*	0.452 0	*	0.498 3	*	**0.869 4**
CNN	86.52	0.658 1	92.13	0.868 2	64.38	0.310 1	74.59	0.399 6	83.55	0.717 0
LSTM+CNN	86.7	0.670 0	92.43	0.872 4	64.39	0.471 3	74.60	0.478 3	83.90	0.834 8
Att-Grader	**88.3**	**0.710 0**	**93.06**	**0.883 3**	**69.87**	**0.513 2**	**75.92**	**0.512 4**	**84.90**	0.844 2

从表3中可以看出，Att-Grader模型在数学数据集上表现明显优于其他三个baseline，这表明：系统加入注意力机制与新构建的参考答案后，不仅更好地捕获了学生答案与参考答案的匹配情况，而且通过参考答案覆盖了更多学生答题情况。同时也可看出LSTM+CNN表现优于CNN，可能的原因为：LSTM与CNN结构特点不同，为NLP任务提供互补的信息。这与研究者已经得到的结论一致[20]。

从表3中还可以看出，在语文数据集上，Att-Grader模型的表现大部分时候优于CNN与LSTM+CNN，只是在问题3上的QWKappa指标不如KAGrader。可能的原因是：问题3开放程度更高，利用满分答案拓展的参考答案不能覆盖学生的各种答题情况。因此需要进一步针对开放程度高的问题研究如何选择代表性答案来形成参考答案。

4.3.2 不同注意力机制对自动评分的影响

本文在语文数据集上探讨了不同注意力机制对系统性能的影响，具体结果如表4所示。其中，No-Attention表示Att-Grader模型没有使用attention层；SelfAttention表示模型仅使用自注意力机制；CoAttention表示模型仅使用双向互注意力机制；SelfCoAttention表示模型中既有自注意力机制又有双向互注意力机制，即学生答案与参考答案分别先通过自注意力机制获取内部关键特征后，再通过互注意力获得两者的匹配信息。

表4 不同注意力机制下的自动评分结果

	CRCC 1		CRCC 2		CRCC 3	
	Acc/%	QWK	Acc/%	QWK	Acc/%	QWK
NoAttention	64.39	0.471 3	74.60	0.478 3	83.90	**0.834 8**
SelfAttention	66.18	0.491 6	74.10	**0.503 2**	83.60	0.808 5
CoAttention	**69.05**	**0.508 8**	**74.79**	0.489 5	**84.03**	0.831 6
SelfCoAttention	66.29	0.449 2	73.97	0.480 1	83.68	0.818 9

由表4可以看出：模型加入注意力机制后效果更好，表明注意力机制能有效提升自动评分的性能。还可看出：CoAttention比SelfAttention以及Self-CoAttention效果都要好，可能的原因是目前数据规模较小，答案长度较短且答题方式多样，模型不能很好地学习出自身的重要概念。因此，在Att-Grader模型中，本文选择效果更好的双向互注意力机制。

4.3.3 代表性答案选择对自动评分的影响

本文以数据集CRCC 1作为测试样例集，与随

机选择满分答案构建的参考答案集进行对比，具体实验结果如表 5 所示。

表 5　不同参考答案下的自动评分结果

	Acc/%	QWK
Att-Grader-1	69.05	0.5088
Att-Grader-6(Random)	69.24	0.5051
Att-Grader-6	**69.87**	**0.5132**

其中 Att-Grader-1 表示评分中仅使用预先提供的参考答案；Att-Grader-6(Random)表示使用随机选择方式构建参考答案集来进行评分；Att-Grader-6 表示通过聚类及代表性答案选择的方式构建的参考答案集来进行评分。

可以看出：Att-Grader-6 的评分效果在两个评价指标上均高于其他两个。表明通过选择代表性答案扩展参考答案对自动评分任务非常有效。

4　结论与展望

本文采用基于聚类与最大相似度方法选择代表性答案构建更完备的参考答案，更多地覆盖了学生答题情况。此外还提出基于互注意力机制的神经网络模型，刻画参考答案与学生答案的匹配情况。实验结果表明：本文所提方法有效提升了自动评分效果。但是短答案自动评分的准确率，尤其是开放程度高的短答案问题的评分还有很大提升空间。未来我们将研究不同分值下代表性答案的选择，旨在进一步扩充参考答案；同时还将从错误发现等角度探索可解释分值的实现策略。

参考文献

[1] 郑庆华，董博，钱步月等，智慧教育研究现状与发展趋势[J]. 计算机研究与发展，2019，56(1)：209-224.

[2] Madnani N，Loukina A. Cahill A. A large scale quantitative exploration of modeling strategies for content scoring[C]//Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications，2017：457-467.

[3] Page E B. The imminence of grading essays by computer[J]. Phi Delta Kappan，1966，47(5)：238-243.

[4] Bachman L F，Carr N，Kamei G，et al. A reliable approach to automatic assessment of short answer free responses[C]//Proceedings of the 19th International Conference on Computational Linguistics，2002：1-4.

[5] Tom M，Nicola A，Peter B. Computerized marking of short-answer free-text responses[C]//Proceedings of the 29th Annual Conference of the International Association for Educational Assessment，2003.

[6] Pulman S，Sukkarieh J. Automatic short answer marking[C]//Proceedings of the 2nd Workshop on Building Educational Applications Using Natural Language Processing，2008：29.

[7] Sultan M A，Salazar C，Sumner T. Fast and easy short answer grading with high accuracy[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2016：1070-1075.

[8] Saha S，Dhamecha T，Marvaniya S，et al. Sentence level or token level features for automatic short answer grading?: use both[C]//Proceedings of Artificial Intelligence in Education，2018：503-517.

[9] Bailey S，Meurers D. Diagnosing meaning errors in short answers to reading comprehension questions [C]//Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications at ACL，2008：107-115.

[10] Zhang Y，Shah R，ChiM. Deep learning + student modeling + clustering：a recipe for effective automatic short answer grading[C]//Proceedings of the 9th International Conference on Educational Data Mining，2016：562-567.

[11] Riordan B，Horbach A，Cahill A，Zesch T，et al. Investigating neural architectures for short answer scoring[C]//Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications，2017：159-168.

[12] Marvaniya S，Saha S，Dhamecha T，et al. Creating scoring rubric from representative student answers for improved short answer grading[C]//Proceedings of Conference on Information and Knowledge Management，2018：993-1002.

[13] Lan A，Vats D，Waters A，et al. Mathematical language processing：automatic grading and feedback for open response mathematical questions[C]//Proceedings of the 2nd ACM Corfernce on Learning at Scale，2015：167-176.

[14] Bahdanau D，Cho K，Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv：1409.0473，2014.

[15] Chaudhari S，Polatkan G，Ramanath R，et al. An attentive survey of attention models[J]. arXiv preprint arXiv：1904.02874，2019.

[16] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv：1408.5802，2014.

[17] Kingma D, Ba J. Adam. A Method for Stochastic Optimization[C]//Proceedings of the 3rd International Conference for Learning Representations, San Diego, 2015.

[18] Boer P, Kroese D, Mannor S, et al. A tutorial on the cross-entropy method[J]. Annals of Operations Research, 2005, 134(1): 19-67.

[19] Huang Y, Yang X, Zhuang F, et al. Automatic chinese reading comprehension grading by lstm with knowledge adaptation[C]//Proceedings of Pacific-asia Conference on Knowledge Discovery & Data Mining, 2018: 118-129.

[20] **彭敏，姚亚兰，谢倩倩，等. 基于带注意力机制CNN的联合知识表示模型[J]. 中文信息学报**, 2019, 33(2): 51-58.

[21] **谭咏梅，刘姝雯，吕学强. 基于CNN与双向LSTM的中文文本蕴含识别方法[J]. 中文信息学报**, 2018, 32(7): 11-1.

[22] **包乌格德勒，赵小兵. 基于RNN和CNN的蒙汉神经机器翻译研究[J]. 中文信息学报**, 2018, 32(8): 60-67.

[23] **沈龙骧，邹博伟，叶静，等. 基于双向LSTM与CRF融合模型的否定聚焦点识别[J]. 中文信息学报**, 2019, 33(1): 25-34.

[24] **彭敏，杨绍雄，朱佳晖. 基于双向LSTM语义强化的主题建模[J]. 中文信息学报**, 2018, 32(4): 44-49.

[25] Yin W, Kann K, Yu M, et al. Comparative study of CNN and RNN for natural language processing[J]. arXiv preprint arXiv: 1702.01923, 2017.

谭红叶(1971—)，**副教授，硕士生导师，主要研究领域为数据挖掘与人工智能。**
E-mail: hytan_2006@126.com

卢宇(1982—)，**副教授，硕士生导师，主要研究领域为教育数据挖掘、学习分析、普适计算、人工智能及其教育应用。**
E-mail: luyu@bnu.edu.cn

午泽鹏(1993—)，**硕士研究生，主要研究领域为自然语言处理及应用。**
E-mail: wuzepeng_sxu@126.com