DOI :10.11721/cqnuj20130218

基于计算机辅助评价的主观题自动评测研究*

姜振凤[1], 刘 力[2]
(1. 枣庄学院 信息科学与工程学院, 山东 枣庄 277160 ;2. 沈阳师范大学 教师专业发展学院, 沈阳 110034)

摘要 :对于主观题答案的自动测评研究一直是计算机辅助评价实施中的热点和难点问题。为了更好地发挥现有计算机辅助评价系统的功能,提出了一种新的评估方法来实现对学生提交的主观题答案进行自动评估。评价的执行需要建立一个由课程教师及有关专家编写的参考答案知识库,每个问题包括若干个参考答案。然后重点研究了如何将 BLEU 算法运用到评价中,并通过对 n 元组的处理,针对同义词及拼写错误等问题对算法进行了改进。利用改进的 BLEU 方法来确定与学生给出答案最相似的参考答案,并计算学生答案同选定参考答案的相似度,从而得出对该主观题成绩的评定。最后通过实验证明了该方法的有效性。结果表明,同现有的较为成熟的评价手段相结合,该方法可用于对主观题答案的测评,并能够在计算机辅助评价中发挥更为积极的作用。

关键词 :计算机辅助评价 ;自动评分 ;BLEU 算法 ;主观题
中图分类号 :TP311 **文献标志码** :A **文章编号** :1672-6693(2013)02-0074-05

计算机技术和教育测量与评价思想的融合促使计算机辅助评价(Computer-assisted assessment, CAA)成为考核并评价学习者对知识、技能等掌握情况的有效途径。简单地说,凡是借助于软件和设备进行的测试和评价都属于计算机辅助评价的研究领域[1]。近几年随着计算机技术在教育信息化应用中的不断普及,对于 CAA 的研究也引发了在评价内容、方法和形式上的深刻变革[2]。本文在对现有计算机辅助评价系统进行分析与研究的基础上,重点讨论 CAA 对计算机辅助测试结果所进行的评价和反馈。

现有的计算机辅助评价系统也称为自动阅卷系统,是基于计算机辅助评价理论与技术设计开发的系统[3]。目前,大多数计算机辅助测评系统中,对于试题库维护、智能组卷、在线答题等功能均采用了各自相对比较成熟的解决方案,并且已经能很好地完成对客观试题的自动评阅工作[4],但对于一些主观性很强的问题,并没有很好的解决方法。主观题最重要的特征是题目的解答需要通过语言的表述来完成,标准答案具有多样性、不确定性等特点,因此对于主观题的自动批阅成为 CAA 研究的重点和难点,它涉及到人工智能、模式识别和自然语言理解等方面,越来越受到国内外研究人员的关注。目前,国内外关于自由文本类答案的自动评估方法主要包括 :将基于关键字的方式与文本深度分析相结合 ;使用模式匹配技术 ;将答案分解成概念及其语义依赖 ;将一些机器学习技术进行整合 ;使用潜在语义分析(LSA)来对维度空间进行降维等[5]。这些方法在技术上概括起来可分为 3 类,即关键词分析、完整自然语言处理及信息提取技术。由于很难实现完整文本解析和语义分析,且存在跨语言理解等问题,因此在相关方法类别中,“ 信息提取技术 ”提供了一种更为适合、有效的方式,它利用自然语言处理(NLP)工具对特定内容进行文本搜索而无需作深入的分析[6]。

本文提出了一种基于信息提取的评价方法来对学生提交的主观题答案进行自动评测。利用该方法进行评价,需要建立一个由课程教师及相关专家制定的参考答案知识库,且每个问题包括若干个参考答案,并应用一个改进的 BLEU 算法来实现具体的评价操作。

1 BLEU 方法概述

1.1 BLEU 介绍

BLEU(Bilingual evaluation understudy)是由 IBM 作为翻译系统的自动评估功能而提出的一种度量标准。主要用于计算被测译文与参照译文间相似度及距离的

* 收稿日期 :2012-06-07 修回日期 :2012-09-08 网络出版时间 :2013-03-16 13 :37
资助项目 :辽宁省高等学校科研项目(No. 2008Z197)
作者简介 :姜振凤,女,硕士研究生,研究方向为数字化教育资源、数据挖掘,E-mail :jiangzhenfeng@126.com ;通讯作者 :刘力,E-mail :dlb110@126.com
网络出版地址 :http ://www.cnki.net/kcms/detail/50.1165.N.20130316.1337.201302.74_018.html

机器翻译测评[7]。BLEU 算法的健壮性源于它使用若干个由专业人工翻译的参考文本,来对被测译文进行比较。主要过程概括如下。

1)确定被测译文中对于不同 n 值的 n 元组(*n-grams*)在参考译文中出现的匹配次数。并利用(1)式计算每个 n 元组的匹配精度 p_n。

$$p_n = \frac{\sum_{C \in \{candidates\}} \sum_{n\text{-}gram \in C} Count_{\text{clip}}(n\text{-}gram)}{\sum_{C \in \{candidates\}} \sum_{n\text{-}gram \in C} Count(n\text{-}gram)} \quad (1)$$

2)利用(2)式计算 n 元匹配的 BLEU 分数。

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} (w_n \log p_n)\right) \quad (2)$$

其中,$\{candidates\}$ 表示被测译文句子组,函数 $Count_{\text{clip}}(n\text{-}gram)$ 计算在被测译文和给定的参考译文之间的 n 元组匹配数量。函数 $Count(n\text{-}gram)$ 计算在每个被测译文句子中 n 元组的总数。w_n 代表权重系数($w_n \in [0, 1]$)。需要注意的是,一般对 BLEU 算法的基本处理是将 N 赋值为 4,统一权重 $w_n = 1/N$。BP 是一个简洁处罚系数,由于存在被测译文中包含很多 n 元组但本身并不完整这一问题,因此可利用 BP 来对较短的被测译文进行处理。取 c 为被测译文的长度,r 为有效参考译文的长度,BP 可通过(3)式进行计算。

$$BP = \begin{cases} 1, & \text{if} \quad c > r \\ e^{(1-r/c)}, & \text{if} \quad c \leqslant r \end{cases} \quad (3)$$

利用 BLUE 方法对翻译结果进行评价,BLUE 得分越高就表示译文质量越好。另外,由于使用不同译者翻译的参考文本,能够增强被测译文中某一特定词汇及其相对语序出现在任一参考译文中的可能性,从而提高了评价的精度。

1.2 BLEU 应用于 CAA 的可行性分析

BLEU 算法可用于 CAA 中对带有自由文本性质的主观题答案进行自动评估。它将学生答案同参考答案进行比较。在这里,学生答案可被视为被测译文,其准确性是作者想要评价的;教师给出的参考答案即为专业人工翻译的参考译文[5]。对于每一个问题应该包含若干个参考答案,因此 BLEU 评分可以用于确定与学生答案最接近的参考答案。通过对每个参考答案计算其 BLEU 评分,可以选择具有最高评分的那个参考答案。基于这个参考答案,通过计算相似度分数给出学生答题的成绩。具体实施方法将在下一节详细介绍。

2 BLEU 方法在主观题自动评测中的应用研究

2.1 主观题答案知识库的建立

为了更好地实施评估,提高主观题评价的准确性,需要建立主观题答案知识库。根据主观题的特点,学生在回答问题时可能使用不同的语法、关键词以及不同字数来写答案。因此每个问题应该包含若干个参考答案。这些参考答案可以在语法、关键词、单词计数方面有所区别,但意思应该是相同的,也就是说,参考答案应具有相同的“意译”。

由于通过人工方式给出的参考答案更为准确,因此知识库中对于具体主观题目的参考答案意译应由专家给出。通过为每道题目准备若干个参考答案,可以基于最相似的参考答案对学生答案进行评估,从而得出一个更为准确的评价结果。

2.2 评价实施中 BLEU 算法的改进及应用

2.2.1 BLEU 方法在 CAA 中应用的局限性　如前所述,BLEU 方法在 CAA 中的应用是可行且有效的,它的简洁性和语言无关性使其能够与别的较为成熟的评价方法相结合,在计算机辅助评价中发挥积极的作用。但 BLEU 方法也有一定的局限性,主要体现在:

1)在 BLEU 算法中,对 n 元组的匹配必须是精确的完全字符匹配,忽略了真实语言中存在大量多词同义的事实,无法真实反映语句相似度[8]。

2)在 BLEU 算法中,所有 n 元组都具有相同的权重,没有考虑到主观题答案中不同词的权重问题。

因此将 BLEU 算法应用到主观题答案评估之前需要对该算法进行一些改进处理。

2.2.2 BLEU 方法的改进及应用　1)对同义词及拼写错误的处理。在 BLEU 评分过程中要考虑对同义 n 元组的处理。就是说,如果学生答案中的词与参考答案中的词是同义的,它们将被认定为是相匹配的词。对于一个参考答案和学生答案,作者的目标是找到参考答案的“意译”,这个“意译”同原来参考答案相比在措词上与学生答案更接近。

给定参考答案 $R = r_1, r_2, \cdots, r_m$ 和学生答案 $S = s_1, s_2, \cdots, s_p$,通过将 R 中的词用 S 中的语境等效词替换的方法,生成一个合成的参考答案 S_{RS},S_{RS} 保留了 R 的含义且对 S 的词覆盖率最高。在为每个参考答案生成 S_{RS} 后,再计算其 BLEU 评分。需要注意的是,进行替换操作时应考虑以下假设条件:

①假设参考答案中的词已经在学生答案中出现了,就不必替换了。因此需要关注 $\{r \mid r \in R - S\}$ 和 $\{s \mid s \in S - R\}$ 中不匹配的词。

②如果 $n\text{-}gram_S \in S$,$n'\text{-}gram_R \in R$,且 $n'\text{-}gram_R$ 是 $n\text{-}gram_S$ 的同义词,则 $n'\text{-}gram_R$ 将被替换为 $n\text{-}gram_S$。替换后,$n\text{-}gram_S$ 将不再考虑作进一步替换。为了找到同义词,首先考虑三元词(trigram),然后是二元词(bigram),最后是一元词(unigram)。

③由于每个学习领域都有其独特的专业术语,因此使用一个专业字典来检查 n 元词汇组是否是同义的。专业字典由相关领域专家制定,并对版本进行定期的更新和完善。一种有效的方式是通过从优秀学生答案中找出新的同义 n 元组对词典进行词汇扩展[9]。

此外,如果在学生答案中存在拼写错误问题,也无法利用 BLEU 方法将其同参考答案中相应的词进行匹配。改进 BLEU 评分的一种方法是使用拼写检查脚本(Spell-check scripts)。如果一个学生写入了错误的字,拼写检查脚本可以通知他并建议其修改。

2) n 元组的权重分配处理。当学生答案同参考答案进行比较时,匹配的 n 元组评分过程应考虑词汇在句子中的权重问题。这是因为在自由文本类型的主观题答案中,存在着与答案评估关联度较高的关键词以及关联度较小的一般词汇组,因此不能等同视之,需要在参考答案中对每个 n 元组的权重进行分类处理。

表1 权重分类表

类别(C)	n 元组权重($weight$)
$C1$ 不重要	1
$C2$ 比较重要	2
$C3$ 重要	3
$C4$ 非常重要	4

如表1所示,将 n 元组的权重设置分为4个类别,并对应不同的数值。然后利用(4)式计算 n 元组的精度。其中 ra 表示参考答案,sa 代表学生答案。$p_{ra}(n)$ 是参考答案和学生答案中 n 元组的精度。函数 $Count_{\text{clip}}(n\text{-}gram)$ 用于计算在学生答案和参考答案之间 n 元组匹配的数量,且有

$$Count_{\text{clip}}(n\text{-}gram) = \min(Count_{sa}(n\text{-}gram), Count_{ra}(n\text{-}gram))$$

$Count(n\text{-}gram')$ 用于计算学生答案中 n 元组的总数。

$$p_{ra}(n) = \frac{\sum_{n\text{-}gram \in sa} Count_{\text{clip}}(n\text{-}gram)}{\sum_{n\text{-}gram' \in sa} Count(n\text{-}gram')} \tag{4}$$

若考虑到对 n 元组不同类别的权重设置,将(4)式中的 $\sum_{n\text{-}gram \in sa} Count_{\text{clip}}(n\text{-}gram)$ 部分变为

$$\sum_{n\text{-}gram \in sa} weight_{ra}(n\text{-}gram) \times Count_{\text{clip}}(n\text{-}gram)$$

由 $weight \in [1,4]$ 及(5)式可推出关系表达式(6)。

$$\begin{cases} \max\left(\sum_{n\text{-}gram \in sa} Count_{\text{clip}}(n\text{-}gram)\right) = \\ \sum_{n\text{-}gram' \in sa} Count_{\text{clip}}(n\text{-}gram') \\ \min\left(\sum_{n\text{-}gram \in sa} Count_{\text{clip}}(n\text{-}gram)\right) = 0 \end{cases} \tag{5}$$

$$0 \leqslant \sum_{n\text{-}gram \in sa} weight_{ra}(n\text{-}gram) \times Count_{\text{clip}}(n\text{-}gram) \leqslant 4 \times \sum_{n\text{-}gram' \in sa} Count_{\text{clip}}(n\text{-}gram') \tag{6}$$

$$P'_{ra}(n) = \frac{\sum_{n\text{-}gram \in sa} weight_{ra}(n\text{-}gram)}{4 \times \sum_{n\text{-}gram' \in sa} Count_{\text{clip}}(n\text{-}gram')} \times Count_{\text{clip}}(n\text{-}gram) \tag{7}$$

由于 n 元组的精度在0和1之间,通过变换可以得到加权的 n 元组精度计算(7)式,因此可以得到修改后的 BLEU 评分计算公式(8)。

$$BLEU'_{ra} = \exp\left[\sum_{i=1}^{N} w_n \log(P'_{ra}(n))\right] \tag{8}$$

2.3 成绩评定

本文中,对主观题答案成绩的评定概括起来主要包括以下几个方面:①将学生答案同每个参考答案进行比较,确定与学生答案最接近的参考答案;②基于选定的参考答案,考虑词序相似度的影响;③计算学生答案同选定参考答案的相似度评分并得出评定结果。

2.3.1 选择最相近的参考答案 如前所述,利用改进的 BLEU 方法为每个参考答案计算 BLEU 评分 $BLEU'_{ra}$。然后利用(9)式通过计算评分最大值以选择具有最高评分的参考答案。

$$Max_{\text{Score}} = \text{Max}\{BLEU'_{ra} \mid ra \in RA\} \tag{9}$$

其中,Max_{Score} 为改进后的 BLEU 评分最大值,ra 表示参考答案,RA 表示参考答案集。

2.3.2 确定词序相似度系数 有时,在一个句子中改变词的顺序可能会对句子的含义带来影响,例如下面的例子(表2)。

表2 实例1

参考答案	数据库系统包括数据库管理系统和数据库用户
学生答案1	数据库管理系统包括数据库系统和数据库用户
学生答案2	数据库系统包括数据库用户和数据库管理系统

实例1中,由于次序的不同使得学生答案1同参考答案含义截然不同。因此,在对学生答案评价中应对词序相似度这一问题加以考虑。在这里借鉴了文献[10]中的方法对两个句子中常用词序相似度进行了分析。对于带有 m 和 n 个词的句子 P 和 R,假设 $P = p_1, p_2, \cdots, p_m$, $R = r_1, r_2, \cdots, r_n$ 且 $n \geqslant m$。若 $p_i \in P$, $r_j \in R$,则 $p_i = r_i$ 的含义是在 P 和 R 中能够完全匹配的词,其计数用 δ 标记,且有 $\delta \leqslant m$。把 P 中所含有的 δ 个匹配词按其在原句中相同的顺序放入 X,同样把 R 所含有的 δ 个匹配词按其在原句中相同的顺序放入 Y,因此可

得到两个词集 $X=\{x_1,x_2,\cdots,x_\delta\}$ 和 $Y=\{y_1,y_2,\cdots,y_\delta\}$。使用一个指定的唯一索引号(从 1 到 δ)对 X 中的标记词进行替换,得到 $X=\{1,2,\cdots,\delta\}$,然后将 Y 中的标记词用与 X 中相关联的索引号替换。利用(10)式计算两个句子中的词序相似度 S_0[10]。在本文的评价应用中,将学生答案同参考答案分别看作是算法中的 P 和 R 进行常用词序相似度的计算。

$$S_0=1-\frac{|x_1-y_1|+|x_2-y_2|+\cdots+|x_\delta-y_\delta|}{|x_1-y_\delta|+|x_2-y_{\delta-1}|+\cdots+|x_\delta-y_1|} \quad (10)$$

需要注意的是有些时候词序的改变并不会影响文本内容的中心含义(如实例 1 中的学生答案 2),因此需要定义一个系数 λ 来确定常用词顺序的重要性。

2.3.3 计算相似度评分　基于参考答案集中选定的最相近的参考答案及词序相似度系数,利用(11)式来计算一下学生答案和所选参考答案之间的相似度评分。

$$Sim(s,r)=\lambda\times BP_r\times BLEU'_r+(1-\lambda)\times S_0 \quad (11)$$

其中,$Sim(s,r)$ 是学生答案 s 和选定的参考答案 r 的相似度评分。BP_r 是参考答案 r 的简洁处罚系数,可通过(3)式计算。$BLEU'_r$ 是利用改进的 BLEU 方法得到的参考答案 r 的评分,S_0 是 s 和 r 之间的常用词顺序相似度系数。λ 是常用词顺序重要性的权重系数。

3 实验

为了对提出的方法进行验证,本文设置了一个虚拟评价环境对该评价方法的准确性及有效性进行了测试。其中用于测试使用的答案知识库包含了 45 道名词解释题,涉及 300 个参考答案,分别由 3 位不同的专业教师给出。这 3 位教师还负责对 45 名学生的答案进行人工评测,且每道题给定的成绩范围在 0 到 1 之间。参加测试的学生首先对相关题目及答案进行自主学习,然后参加测试。从测试结果中抽取了 50 个学生答案分别对其进行人工测评和基于改进的 BLEU 相似度计算测评,并对这两种评价结果进行了比较,如图 1。另外以 3 位教师给出的人工评价结果的平均值为标准,对 BLEU 算法改进前后的评价结果相关性进行了比较,见表 3。

表 3　BLEU 改进前后的评价结果相关性对比

评价方法	相似度
BLEU	0.36
改进后的 BLEU	0.84

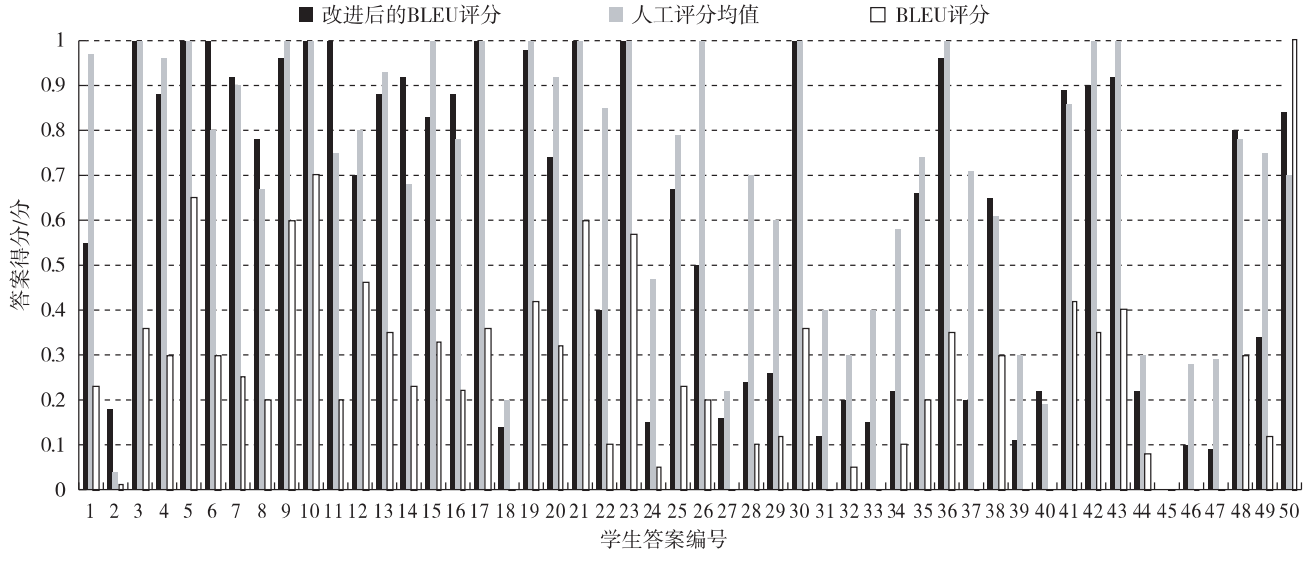

图 1　人工测评与改进的 BLEU 测评结果对照图

通过对评价结果数据的处理,平均人工评分和相似度计算评分之间的相关度为 0.84,且近似一致性至少达到了 75%。其中由于某些学生提交的答案比较抽象,或部分同义词未能正常识别等问题使得 2 种评价方法得到的结果差异较大,这个问题可以通过对答案知识库及专业字典的扩充而改善。此外,同未改进的 BLEU 评价方法相比,改进后的 BLEU 方法由于考虑了对同义词的匹配及对不同词权重系数的设置,在评价结果的相对准确性上有了显著的提高。这意味着本文提出的方法是可行的,且可用于对主观题答案进行自动评分。

4 结语

本文对计算机辅助评价中主观题的评价方法进行了研究,探讨了 BLEU 算法在评价实施中的应用。针对主观题答案在词的选择、词序及语法方面存在的多

样性等问题,将 BLEU 算法进行改进以满足对学生主观题答案进行自动评测的需求。根据本文提出的方法,每个问题要包括若干个由专家提供的参考答案,将学生提供的答案同参考答案进行比较,利用改进的 BLEU 算法来选择与学生答案最为相近的参考答案。选定某个参考答案后,基于相关测量值,如改进的 BLEU 评分、学生答案和选定参考答案之间的常见词序相似度等,来计算学生答案同选定参考答案的相似度评分,并通过实验验证了方法的有效性。在后续的研究中作者将要借助具体的计算机辅助评价系统,基于更为完善的答案知识库对该算法做进一步的实践研究及完善。

参考文献:

[1] 马光仲,蔡旻君. 计算机辅助评价发展的回顾与思考[J]. 电化教育研究,2010(4):49-53.
Ma G Z, Cai M J. Reviewing and thinking about the development of computer assisted assessment[J]. E-education Research, 2010(4): 49-53.

[2] 邹显春,黄敏,李莉. 基于 WEB 的教学辅助平台研究[J]. 西南师范大学学报:自然科学版,2009,34(6):200-203.
Zou X C, Huang M, Li L. Research on Web-based teaching assisted platform[J]. Journal of Southwest China Normal University: Natural Science Edition, 2009, 34(6): 200-203.

[3] 冯立,张景韶,周利平. 基于 B/S 模式下的网络题库平台研究与实践[J]. 重庆师范大学学报:自然科学版,2012,29(4):77-81.
Feng L, Zhang J S, Zhou L P. Research and practice on network item bank platform based on B/S mode [J]. Journal of Chongqing Normal University: Natural Science, 2012, 29(4): 77-81.

[4] 王立君. 运用齐次马尔可夫链分析法进行教学质量评价[J]. 江西师范大学学报:自然科学版,2002,26(2):146-149.
Wang L J. The combination method for dependent Evidence [J]. Journal of Jiangxi Normal University: Natural Science, 2002, 26(2): 146-149.

[5] P′erez D, Alfonseca, E, Rodr′lguez P. Application of the BLEU method for evaluating free text answers in an e-learning environment[C]//[s. n.]. Proceedings of the language resources and evaluation conference (LREC-2004). Portugal: Lisbon, 2004.

[6] Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. [C]//[s. n.]. Proceedings of the American association for artificial intelligence (AAAI 2006). USA: Washington, 2006.

[7] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation. [C]//[s. n.]. Proceedings of the 40th annual meeting of the association for computational linguistics (ACL 2002). USA: Philadelphia, 2002.

[8] Noorbehbahani F, Kardan A A. The automatic assessment of free text answers using a modified BLEU algorithm[J]. Computers & Education, 2011, 56(7): 337-345.

[9] He Y, Hui S H, Quan T T. Automatic summary assessment for intelligent tutoring systems[J]. Computers & Education, 2009, 53(3): 890-899.

[10] Islam A, Zaiu Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity[J]. Transactions on Knowledge Discovery from Data, 2008, 2(2): 1-25.

Probe into the Automatic Score of Subjective Items Based Upon Computer-Assisted Assessment

JIANG Zhen-feng [1], LIU Li[2]

(1. College of Information Science and Engineering, Zaozhuang University, Zaozhuang Shandong 277160;
(2. School of Teacher Education, Shenyang Normal University, Shenyang 110034, China)

Abstract: In allusion to the hot and difficult problem of computer-assisted assessment, in order to improve the function of computer-assisted assessment system, in this study, a new assessment method is presented to realize the automatic score of subjective items. To perform the assessment, it is necessary to establish a repository of reference answers written by course instructors or related experts. Then focused on how to use the BLEU algorithm to assessment, and through the *n-gram* processing, this paper improved the algorithm in the synonyms and spelling errors, etc. Applied the modified version of the BLEU to identify the most similar reference answer to the student answer and then the student answer is scored on its similarity with the selected reference answer. At last, through one empirical research we proved the availability of the new evaluation method. The results showed that, combined with the existing relatively mature evaluation method, this method can be used to for subjective answers assessment, and play a more active role in CAA.

Key words: computer-assisted assessment; automatic score; BLEU algorithm; subjective item

(责任编辑 游中胜)