一种基于关键词的技术来评估广泛的问题回答脚本

Tamim Al Mahmud*, Md Gulzar Hussain, Sumaiya Kabir,
Hasnain Ahmad, Mahmudus Sobhan
Department of Computer Science and Engineering
Green University of Bangladesh
Dhaka, Bangladesh

*Corresponding author: Tamim Al Mahmud; tamim@cse.green.edu.bd

Abstract

评估是通过各种技术（如口头或口试测试、主观或客观书面测试）来评估和确定教育系统的方法。本文提出了一种有效的电子化评价主观答题卷的解决方案。在本文中，我们提出并实现了一个综合系统，用于检查和评估书面答题卷。本文重点在于从答题卷中寻找关键词，并将它们与从开放和封闭领域解析出的关键词进行比较。该系统还检查答题卷中的语法和拼写错误。我们的提议系统使用了 100 名学生的答题卷进行了测试，并获得了 0.91 的精确度分数。

关键词：自动评估，主观评估，自动提取。

1 介绍

数字化评估是收集、分析和解释信息以确定学生达到教学目标的程度的过程。每个教育机构都会评估学生的答题卷，以评估其表现。有许多类型的测试来评估学生的表现，如口头或口试、主观或客观测试。目前，多项选择题（MCQ）考试的评分是通过使用光学标记阅读器（OMR）机器进行的。减少资源使用的做法非常有利可图，但这种方法仅适用于评估 MCQ 答题卷。描述性测试有助于了解学生对某一课程的知识深度。就广泛的问题而言，学生需要用一到多个句子来描述特定问题的答案。要手动评估大量的这类答题卷，评分者将面临巨大的工作压力。如果能够电子地评估主观或描述性的答题卷，则将是教育系统的一大成就。这将节省时间、减少资源利用，并避免评分者的偏见错误。在招聘过程中，组织通常会通过笔试来评估候选人的能力。每年每个职业领域的申请人数都在急剧增加。因此，主观性答题卷的数量也在增加。在这篇论文中，我们展示并实现了一个集成软件系统，用于电子地评估答题卷。

我们的主要目标是-

- 构建一个能够电子评估广泛问题答案脚本的高效系统。

- 提供开放和封闭领域问答的功能。
- 形成一个无误的评估系统。

本文剩余部分的结构如下。相关工作在**第 2 节**中讨论。**章节 3** 讨论了方法，并展示了一个样本数据，**章节 4** 显示了结果。最后，**章节 5** 涉及未来工作和结论。

2 背景概述

在本节中，我们简要讨论了一些与我们的工作相关的现有研究。

各种自动评估自由文本或主观答案的技术，如潜在语义分析（LSA）、智能作文评估器、语法增强的 LSA（SELSA）等，在论文 [14] 中被作者回顾，并发现这些方法是基于关键词的，没有考虑到相邻词语。

论文 [7] 的作者提出了一种模型来评估孟加拉语描述性答案试卷，他们发现在这种情况下最小相对误差为 1.8%。

论文 [5] 建议使用维基百科作为独特信息来源，以解决开放领域的问题。

在这篇论文 [6] 中，作者采用了一种使用粗粒度答案类型的方方法。

在本文中，他们提出了一种新颖的自动问答系统，这是首次研究处理消费者电子领域各种类型用户问题的研究 [17]。

在他们的工作中，论文 [3] 的作者为维基百科提出了一个替代关键词搜索方法，设计为一种回答事实性问题的模型解决方案。

本文介绍了一个新的系统，可以评估学生的表现，考虑了通过解析文本和找到学生答案的语义含义来评估广泛类型问题的评估，并最终将其与教师的答案进行比较并分配最终得分 [2]。[9] 的作者提出了一种使用维基百科文章作为信息来源的开放领域信息访问系统来讨论主题。在 [8] 中，作者开发了一个使用广义潜在语义分析 (GLSA) 的老化系统，该系统的性能水平可以超过人类评分者。在论文 [11] 中，作者介绍了一种试图解决封闭和开放领域问题的系统，其中开放领域的问题可以通过搜索引擎来解答，而封闭领域问题的答案必须存储在数据库中。在 [15] 中，作者描述了处理中文和日语多语言问答的 NTCIR。基于维基百科的问答也是一个热门话题。在 [4] 中，他们将维基百科视为知识库。

在 [13] 中，作者讨论了 NSIR（发音为 “答案”），这是一种正在密歇根大学开发的基于网络的问题回答机制。一旦 NSIR 收到搜索引擎的结果列表，它会处理排名较高的记录并提取一些潜在的回答。[1] 的作者提出了一种系统，该系统将答案作为输入，然后将其与存储的标准答案进行比较，并通过匹配关键词或主要思想与传统回复同义词来评估每个回答。在 [12] 中，本文的目标是探索计算语言学方法在自动标记自由文本答案中的应用。

在 [10] 这项工作中，作者提出了一种用于电子学习文档智能搜索的全自动问答系统。该方法利用自然语言处理技术来定义问题的语义和句法结构。最后，在 [16] 中，本文介绍了建立在问题语义描述和本体论基础上的应用框架和理论。它还介绍了重要技术：问题解析、知识库构建、答案抽取。

3 方法论

在本节中，我们讨论了所提出的系统的运作方法。
我们的系统分为两个步骤来评估考生的答案脚本，这两个步骤如图 1 所示。

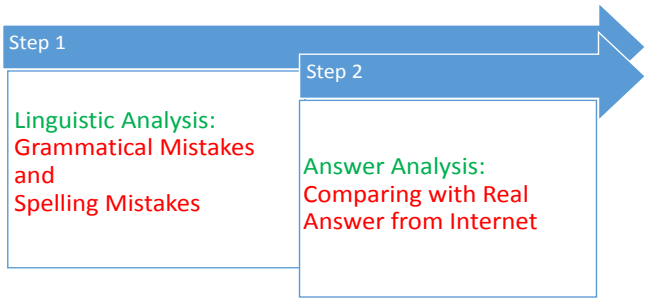

Figure 1: 评估答题卷的步骤

第一步是语言分析，检查答案脚本中的语法错误和拼写错误。对于第二步答案分析，我们提出了两种算法来寻找考生答案脚本以及开放领域和封闭领域结果中单词的频率。这些步骤将在下面进行讨论：

3.1 语言分析

我们可以使用 jlangualetool、Perfect Tense API、Grammar Bot API 等工具进行拼写检查和语法检查。原因是，这些是易于使用且免费的语言工具，在我们的项目中可以用作拼写和语法检查器，同时我们也可以使用现有的项目如 hunspelljna 或 hunspellbird（在 maven 中央仓库）。这些都是强大的语法规正工具，能够准确理解文本的意义和上下文。

使用这些工具，为了给出语言学评分我们提出了算法 1-

Algorithm 1 生成语言分数的算法

字符串答案 = 总学生答案数；
初始 LAScore=0；
初始 SMistake= 拼写错误数量；
初始 GMistake= 语法错误数量；
初始 TWord= 答案中的单词数；
初始 TSentence= 答案中的句子数；
$LAScore = \frac{SMiskate}{TWord} * 100 + \frac{GMiskate}{TSentence} * 100$

3.2 答案分析

答案分析步骤的系统架构如图 2 所示。
本步骤所需的流程如下所述：

3.2.1 从域中删除答案

我们提出的系统将专注于从问题中提取关键词后，在开放领域系统中的数据解析，例如维基百科。Mediawiki 操作 API 是一个网络服务，提供了方便访问 wiki 功

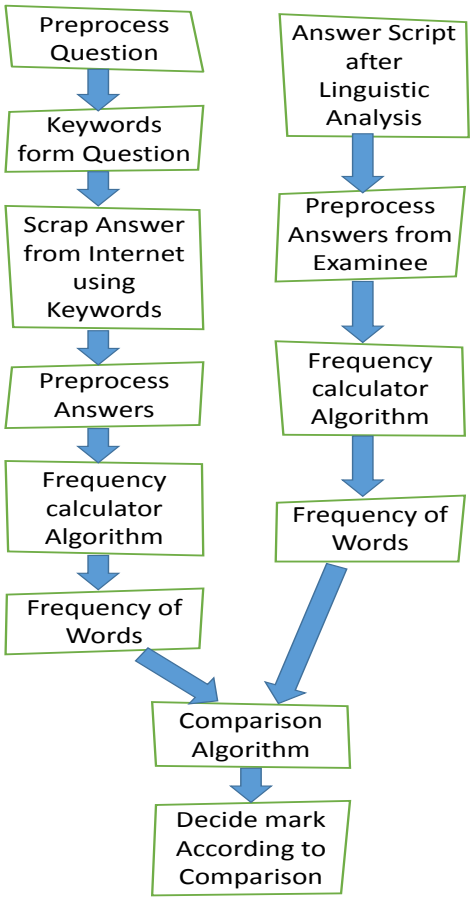

Figure 2: 答案分析的系统架构

能的途径。入口点：“https://en.wikipedia.org/w/api.php”，或任何其他维基。参数通过查询字符串传递。空值传递会给我们自动生成文档的帮助页面。我们需要选择一个输出格式。MediaWiki 提供的数据输出形式有 JSON、jsonfm、php（序列化格式）、phpfm、wddx、wddxfm、XML、xmlfm、yaml、yamlfm 和 rawfm。后缀为”fm”的格式是以 html 的美化方式呈现。本文考虑了使用维基百科在开放领域环境中回答问题的问题。维基百科包含了人类感兴趣的更新知识。该系统的主要目标是检索用户提出查询的正确答案。返回的答案形式为相关文档。数据源可以是全球网络或本地域系统，它会接受一组关键词构成的问题，例如，什么是开放领域系统？它将接受关于这个关键词的所有来自维基百科的数据，然后通过预处理减少解析数据中的不必要部分。

3.2.2 预处理

在此过程中，每当文本被渲染时，特定的词语会被提取。像 a、and、the、is 等停用词在这个过程中会被移除。规范化过程也会发生，比如去除不必要的字符和特殊字符如’#’、’,’ 等。大写会被去掉，例如’Hello’ 会转换成’hello’。但有些例子如’US’ 不会被转换为’us’ 因为它改变了意思。分词也被用来将段落分割成句子，并进一步将句子分割成单个词语。这些预处理是使用各种正则表达式 (RE) 完成的。

3.2.3 单词频率计算器

我们提出了一种算法，在预处理答案后计算每个单词的频率。此过程同时适用于考生的答案脚本和从领域中获取的结果。为此，提出了算法 2-

Algorithm 2 单词频率计算器算法

字符串答案 = 总预处理后答案;
初始数组频率 = 空值;
for *Every words in answer* **do**
    if *Word is not in the Array* **then**
        Add the word as index in the Array
        frequency[word] = 1
    else
        frequency[word] = frequency[word] + 1
    end
end

3.2.4 两个结果的比较

在对来自开放和封闭领域的解析数据运行词频计算器算法后，我们将得到两个频率集。为了评估学生的答案，我们需要比较这两个频率集，并根据比较结果设定分数。为此，我们提出了算法 3-

Algorithm 3 比较算法

初始 AAScore=0

$SWFrequency = FrequencyResultOfStudentAnswer$

$RWFrequency = FrequencyResultOfParsedAnswer$

初始 lengthSWF=Length of SWFrequency

初始 lengthRWF= $\sum AllValuesOfRWFrequency$

初始 WeithtRWF= 空数组

for *Every words in RWFrequency* **do**
  Add 'word' as index in WeightRWF
  $WeightRWF[word] = \frac{RWFrequency[word}{lenghtRWF} * 100$
end

初始长度 WRWF= 权重长度 RWF

for *Every words in SWFrequency* **do**
  if *word is in WeightRWF* **then**
    $AAScore = AAScore + \frac{WeightRWF[word]}{lenghtWRWF}$
  else
    Do nothing and Continue
  end
end

for *Every words in SWFrequency* **do**
  if *word is in RWFrequency* **then**
    $AAScore = AAScore + \frac{SWFrequency[word]}{RWFrequency[word]} * 100 + \frac{SWFrequency[word]}{lengthSWF}$
    $RWFrequency[word] = 0$
  else
    Do nothing and Continue
  end
end

for *Every words in RWFrequency* **do**
  if *RWFrequency[word] is not 0* **then**
    $AAScore = AAScore - \frac{RWFrequency[word]}{lengthSWF} * 100$
    $RWFrequency[word] = 0$
  else
    Do nothing and Continue
  end
end

3.3 最终评分学生的答案

经过比较和语言分析后，它将给出分数。分数将根据表 I 分配，该表可以由系统管理员修改。

使用表 I 和下列公式给出学生的答题卷最终得分。

$FinalScore = TotalMark*0.7*(AAScore/100)+TotalMark*0.3*(LAScore/100)$

Table 1: 分数分布取决于不同步骤

Section	Scores
Frequency of Word	70%
Linguistic Mistakes	30%

3.4 简单的示例与样本数据

首先，问题将从维基百科中提取并进行搜索，这个过程是为了明确问题。例如：你对达卡大学了解多少?

现在系统将从问题中识别出主要和独特的关键词。例如：”大学，达卡”。然后将在维基百科中搜索数据。示例如下图 3 所示：

Meeting at Amtala on Ekushey February.JPGthumbMeeting on the University of Dhaka premises on 21 February 1952313x313px Established in 1921 under the Dacca University Act 1920 of the Indian Legislative Council, it is modelled after British universities. Honorable Chancellor Lord Lawrence John Lumley Dundas, 2nd Marquess of ZetlandRonaldshay was the Governor of Bengal between 1917 and 1922. He designated Nawaab Syed Shamsul Huda as a life member. On his rmendation, Lord Ronaldshay designated Ahmad Fazlur RahmanSir Ahmad Fazlur Rahman as a provost, earlier he was in Aligarh Muslim University.The Muslim Heritage of Bengal-by Muhammad Mojlum Khan-Kube Publishing Ltd.,UK- ISBN978-1-84774-059-5 Academic activities started on 1 July in 1921 with 847 Studentsname Dhakauniversity>citationepaper.thedailystar.netDU- Prospectus-2008.pdf archivedate14 November 2012 accessdate16 July 2016

Figure 3: 从维基百科解析达卡大学的数据

然后通过对预处理后的维基百科数据运行频率计算器算法，我们将获得频率表，并将其存储为图 4 中给出的标准答案。

Ramna=1, through=1, Faculties=1, Institutes=1, Centres=1, inception=1, main=1, areas=1, research=1, number=1, character=1, contributions=1, than=1, 877=1, create=1, July=1, up=1, 51=1, day=1, Halls=1, knowledge=3, 11=1, new=1, 13=1, making=1, having=1, enriched=1, pool=1, opened=1, doors=1, risen=1, 3=3, University=7, 600=1, 60=1, dormitories=1, acres=1, scholars=1, present=1, fields=1, Dhaka=1, Sir=1, halls=1, 20=1, Faculties12=1, city=1, purpose=1, part=1, students=5, distinct=1, respectively=1

Figure 4: 模型频率分析数据解析

之后系统将把学生的答案作为文本，并将下图 5 视为样本。

预处理并运行频率计算器算法后，我们将得到频率表，并将其存储为图 6。

系统将比较这些频率表，并根据比较算法对答案进行评分。最后，学生的答案得分将根据表 I 进行评估。

The University started its activities with 3 Faculties,12 Departments, 60 teachers, 877 students and 3 dormitories (Halls of Residence) for the students. At present the University consists of 13 Faculties, 77 Departments, 11 Institutes, 20 residential halls, 3 hostels and more than 51 Research Centres. The number of students and teachers has risen to about 37,064 and 1,885 respectively.
The main purpose of the University was to create new areas of knowledge and disseminate this knowledge to the society through its students. Since its inception the University has a distinct character of having distinguished scholars as faculties who have enriched the global pool of knowledge.

Figure 5: 学生样答

University=7, Faculties12=1, city=1, purpose=1, part=1, students=5, distinct=1, respectively=1, disseminate=1, global=1, notable=1, teaching=1, Research=1, society=1, At=1, known=1, teachers=2, land=1, Since=1, Hartog=1, picturesque=1, 77=1, Departments=2, 37064=1, set=1, more=1, started=1, Residence=1, distinguished=1, faculties=1, 1921=1, residential=1, 1885=1, activities=1, hostels=1, consists=1, PJ=1, first=2, On=1

Figure 6: 学生回答的频率

4 结果分析

本研究的目的是自动评估描述性答案脚本并为其打分。为此，我们把学生的答案作为文本，并运行了我们的方法进行测试。我们从 100 名学生那里收集了答案并用我们的方法进行了测试。我们也让本研究所的 10 位教师手动测试了这些答案脚本。然后我们计算了所提方法的精准率、召回率和 F 值，这些得分如表 2 所示。

Table 2: 精确率、召回率和 F 值对于我们提出的方法

	Score of our proposed system
Precision	0.91
Recall	0.81
F-score	0.87

从表 2 我们可以看出，我们提出的方法提供了可接受的精度、召回率和 F 值。

5 结论与未来工作

自动评分脚本的简便性在现代是一个很大的挑战。学生和求职者的人数不断增加，教师和组织每天都在评估答题纸时面临更大的困难。在这种情况下，借助自动答案脚本评估提供了新的解决方案。我们的系统将评估主观答案。我们的系统根据关键词来评估学生的答案。基于标准答案和学生的答案进行判断后，分数会被分配

给学生。对学生人数过多的考试来说成本很高，并且评估他们的答题纸耗时很长。提议的系统提供了一个无烦恼的评分系统，在这个系统中时间和成本被最小化，并且也减少了人工评分错误。

5.1 未来工作

由于我们的系统无法提供更高的效率，我们提出了一些未来工作的想法。

- 通过图表、表格和数学表达式来评估答案。实现这些功能后，该系统将对每次评估都有用。
- 关注直接指示问题类型的关键词。例如：“什么”，“何时”，“谁”，“描述”，“定义”。我们知道这些关键词指示与时间、地点、人物等相关答案。如果系统能自动捕捉到这些关键词，处理速度将会提高。
- 可以应用于远程学习系统。
- 学生可以从偏远地区参加考试。
- 当前系统仅评估用英语书写的答案。进一步地，它也可以扩展到评估用其他语言书写的答案。
- 实现一个封闭域将使其对敏感考试有用。

References

[1] Kamlesh Koyande Aditi Tulaskar, Aishwarya Thengal. Subjective answer evaluation system. *International Journal of Engineering Science and Computing,*, 10459, 2017.

[2] Sraboni Barua and Tamim Al Mahmud. Future of social network with collaboration of cloud computing and brain computer interface. *International Journal of Computer Applications*, 145(9):34–37, 2016.

[3] Adam Brzeski and Tomasz Boiński. Relation-based wikipedia search system for factoid questions answering. 2014.

[4] Davide Buscaldi and Paolo Rosso. Mining knowledge from wikipedia for the question answering task. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 727–730, 2006.

[5] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

[6] Po-Chun Chen, Meng-Jie Zhuang, and Chuan-Jie Lin. Using wikipedia and semantic resources to find answer types and appropriate answer candidate sets in question answering. In *Proceedings of the Open Knowledge Base and Question Answering Workshop (OKBQA 2016)*, pages 1–10, 2016.

[7] Md Gulzar Hussain, Sumaiya Kabir, Tamim Al Mahmud, Ayesha Khatun, and Md Jahidul Islam. Assessment of bangla descriptive answer script digitally. In *International Conference on Bangla Speech and Language Processing (ICBSLP)*, volume 27, page 28, 2019.

[8] Md Monjurul Islam and ASM Latiful Hoque. Automated essay scoring using generalized latent semantic analysis. In *2010 13th International Conference on Computer and Information Technology (ICCIT)*, pages 358–363. IEEE, 2010.

[9] Graham WI Lcock. Wikitalk: A spoken wikipedia-based open-domain knowledge access system. In *24th International Conference on Computational Linguistics*, page 57. Citeseer, 2012.

[10] Ankush Mittal, Sumit Gupta, Praveen Kumar, and Shrikant Kashyap. A fully automatic question-answering system for intelligent search in e-learning documents. *International Journal on E-Learning*, 4(1):149–166, 2005.

[11] Abhay Mone, Ishwar Mete, Priyanka Gangarde, and Malhari Kharad. Automatic answering system for english language questions.

[12] Stephen G Pulman and Jana Z Sukkarieh. Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 9–16. Association for Computational Linguistics, 2005.

[13] Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. Evaluating web-based question answering systems. *Ann Arbor*, 1001:48109, 2002.

[14] Ms Paden Rinchen. Comparative study of techniques used for automatic evaluation of free text answer. *International Journal on Recent and Innovation Trends in Computing and Communication*, 2(12), 2014.

[15] Yutaka Sasaki, Hsin-Hsi Chen, Kuang-hua Chen, and Chuan-Jie Lin. Overview of the ntcir-5 cross-lingual question answering task (clqa1). In *NTCIR*, 2005.

[16] Jinzhong Xu, Keliang Jia, and Jibin Fu. Research of automatic question answering system in network teaching. In *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for*, pages 2556–2560. IEEE, 2008.

[17] Seunghyun Yoon, Mohan Sundar, Abhishek Gupta, and Kyomin Jung. Automatic question answering system for consumer products. In *Proceedings of SAI Intelligent Systems Conference*, pages 1012–1016. Springer, 2016.