THE EVOLUTION OF DESCRIPTIVE ANSWER EVALUATION IN E-LEARNING: A PREDICTIVE ANALYTICS AND MACHINE LEARNING PERSPECTIVE

V. Kavitha
Research Scholar, Department of Computer Applications
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
kavithav.mca@gmail.com

Dr. A. Jaya
Professor, Department of Computer Applications
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
jayavenkat2007@gmail.com
https://doi.org/10.34293/9789361639715.shanlax.ch.013

Abstract

As education progressively transitions to online learning, evaluating students' understanding through descriptive responses continues to be one of the most challenging responsibilities. Unlike multiple-choice or objective formats, descriptive answers require contextual, semantic, and frequently subjective analysis, presenting unique challenges for automated systems. This chapter examines the active development of techniques to evaluate descriptive answers in the e-learning applications, from manually based grading to intelligent automated systems. The chapter describes the evolution in terms of the transition from rule-based systems to machine learning models that predict scores based on student responses. The chapter discusses how predictive analytics, which includes supervised learning, enabled systems to learn from expert-graded answers and assign scores to new student responses. It addresses the recent trends that utilize deep learning models like BERT and GPT, which can make inferences about the context and semantics of student responses. The chapter highlights how developing technology has improved the Accuracy, equity, and scalability of descriptive answer assessment in technology-enabled contexts.

Keywords: *Descriptive Answer Evaluation Learning Assessment, Predictive Analytics, Supervised Learning, Automated Grading Systems, Deep Learning Models.*

Literature Survey

The focus of this section is on the key studies conducted by other authors, specifically in automated assessment, specifically in the field of descriptive answer grading in the e-learning.

Early Research on Manual and Rule-Based Evaluation

Early forms of evaluation for descriptive answers in the e-learning relied on human grading utilizing scoring rubrics or standardized checklists. While manual grading preserved square judgment of the review, human grading methods did not escape the limitations of scoping variation, bias, and time spent (Attali & Burstein, 2006). As a result, further attempts were made to resolve inconsistencies through rule-based systems that utilized keywords and syntactic patterns to score correctness. The Project Essay Grade (PEG) was one of the first rule-based systems developed in the 1960s and was known for automatically scoring essays using surface-level features (Page, 1966). Rule-based systems

are not strong enough to process word or phrasings, synonyms, or inferences, and they are limited to the patterns that were designed by researchers. (Shermis & Burstein, 2013).

Machine Learning Models in the Educational Space

The introduction of machine learning resulted in the transition from hard-wired representations of algorithms to data-based algorithms. With approaches such as supervised learning (e.g. Support Vector Machines (SVM), Naïve Bayes, and, Logistic Regression) trained to predict output values using labeled datasets (e.g. student responses, expert scores (Sukkarieh & Pulman, 2005)), the ML models extracted a range of features (e.g. TF-IDF, length of sentences, patterns of grammar) to predict outputs. Adding to both the number of potential models and the overall flexibility was the development of ensemble and domain-specific classifiers, which increased the generalizability of the models to some extent across previously unseen responses (Burrows, Gurevych, & Stein, 2015). While the introduction of ML models was a positive advancement over more technical rule-based systems (e.g. IF-THEN), traditionally ML models still have a great deal of difficulty capturing deep semantic understanding this limits the performance of these models on high-order cognition or reasoning question type outputs.

Deep Learning and Natural Language Processing in Assessments

The deep learning and natural language processing (NLP) paradigms represent considerable progress towards evaluating more descriptive forms of answers. Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTM) allowed systems to consider sequential dependencies in model student responses (Taghipour & Ng, 2016). More recently, transformer-based models such as BERT (Devlin et al. 2019) and GPT (Brown et al. 2020) have adopted to examine the semantic meaning of answers in a contextual and bidirectional way. These models also appear to provide promising results in grading open-ended responses, and generating feedback student responses (Uto et al. 2021). However, they require appropriately large labeled datasets, and some researchers and developers complain about explainability and fairness issues when they are deployed in real-world educational environments.

Research Gap

This section identifies the possible research gap while evaluating the descriptive answers. Although recent advances in automated grading systems have been impressive, there remain significant gaps in evaluating descriptive aspects of the answers in the e-learning context. Rule-based methods, machine learning models and deep learning methods all have contributed to the field, but none has advanced the required accuracy and precision needed for human-like evaluation of descriptive answers.

The following gaps have identified in the current literature and technologies:

Domain Dependency

The majority of current systems are domain-specific, requiring new training on the subject matter. A model that has been trained in biology or literature may struggle with

programming or engineering questions because of differences in vocabulary, answer structures, and concepts. This absence of domain adaptability inhibits scalability across educational platforms.

Limited Generalization

Automated grading models also struggle to generalize to different question types, phrases, or student writing styles, even in a given domain. Models that rely mainly on surface-level features or shallow context tend to misclassify answers that were paraphrased, but are actually considered correct, and often are incapable of distinguishing partly correct answers from incorrect ones.

Absence of Multilingual and Cultural Adaptability

Most models are built and evaluated on English datasets, with limited investigation into non-English or multilingual settings. As a result, the ability for generalizability in global contexts is compromised, especially in countries or areas where students provide or write vernacular or code-mixed responses.

Static Learning

Most existing systems use fixed models based on static datasets. The systems do not update or learn from student interactions over time or, if they do, it will require retraining from the end-user. This can lead to stale feedback patterns and a lack of personalization. Thus, there is a need for adaptive systems that adapt over time with new data.

Real-Time Assessment and Feedback

A significant challenge in the e-learning contexts (e.g., MOOCs, synchronous assessments) is that feedback is delivered after the learner has completed a task, which is not in real-time. Most AI programmed systems evaluate asynchronously and do not integrate into interactive learning environments.

Objective

The objective of this Chapter is to provide researchers with an integrated, evolutionary perspective on descriptive answer evaluation. By transitioning from manual and rule-based systems to predictive analytics, and ultimately to deep methods through natural language processing models, this work demonstrates both how far we have come and where we need to go next. Our goal in this contribution is to development of future, scalable, domain-independent, explainable, and adaptive evaluation systems in the e-learning.

Methodology

This section outlines the dataset acquisition and Evolution of Models in the Descriptive Answer Evaluation System.

Dataset:

This research utilizes a custom-developed dataset that researchers specifically built to evaluate descriptive answer grading systems in a domain-specific educational context. The dataset includes open-ended student responses written in a classroom setting, and experts assigned scores to those responses.

Overview and Evolution of Models

A number of models are developed over the years for automating the grading of descriptive answers. The following figure depicts the evolution of the answer evaluation models.

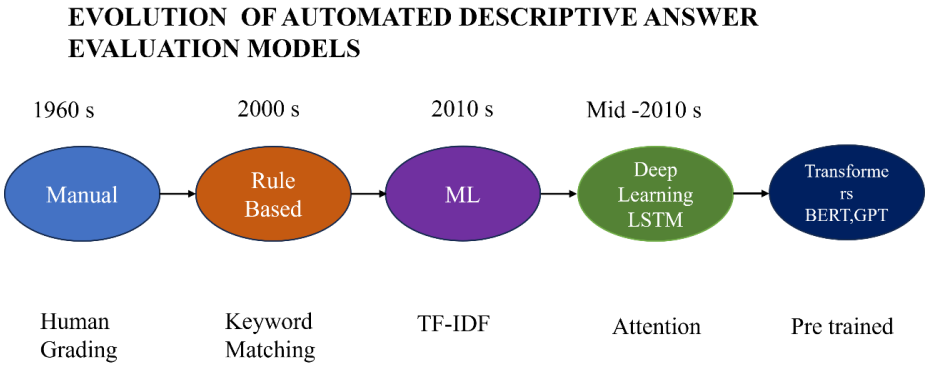

Fig. 1 Evolution of Automated Descriptive Answer Evaluation Models

The Fig.1 refers to the timeline that highlights key developments in these models, along with their main techniques, strengths, and weaknesses.

Human Grading

Human assessment was the foundation of descriptive answer assessment in education. This meant that evaluators were interpreting student responses based on established rubrics or their own interpretations. While human assessment inherently offers tremendous depth of context, it remains time-consuming, inconsistent, and can be biased based on perception and other factors, including fatigue or understanding of context.

Rule-Based Systems

Early attempts to automate grading are based on rule-based systems that compared keywords and phrases in students' answers to pre-specified correct answers. The systems used rules created by designers combined with regular expressions. This method uses a simple and straightforward approach to grading answers, but was inflexible and did not acknowledge that an assignment could be semantically correct even though answered in a different manner.

System with Traditional Machine Learning Models

The next stage of evolution introduced traditional machine learning models, such as Support Vector Machines (SVM), Logistic Regression, and Random Forests, that used engineered features (i.e., "TF-IDF" vectors, length of answers, grammar measures, etc.) to learn from labeled datasets of proximal symbols. While machine learning models improved the generalizability of predictions, it is still limited by an inability to understand profoundly and with context.

Deep Learning Models

Deep learning brought different forms of architectures, including Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, that could understand sequential dependencies better and inform approaches to language recognition. However, these models also required more training data than traditional NLP approaches and struggled with long-range dependencies and complex semantic structures.

Transformer - Based Models

The field experienced a radical transformation with the introduction of transformer architectures, such as BERT and GPT. Transformers use attention mechanisms to understand the context of words in a sentence, allowing for more sophisticated interpretations of student responses. These models can be pre-trained on large corpora, then fine-tuned for tasks such as grading in a fair and contextual way. They can also be directed to provide feedback that is more comment-heavy. Despite their power, they require enormous computational power and generate additional questions about interpretability.

The advances in automated assessment systems, from rule-based approaches to transformers reflects a growing sophistication in the models. Each generational advance addressed the limitations of previous models while continually improving the accuracy, fairness, and contextual sensitivity with which to evaluate descriptive and student - generated responses.

Evaluation Metrics

This section describes the Evaluation metrics used for grading the descriptive answers. There are multiple evaluation metrics to evaluate the performance of automated grading systems for descriptive answers with variation in the evaluation when the task involves score prediction, classification, or feedback generation.

Accuracy

The most frequently used evaluation metric is Accuracy, which describes the number of student responses that were assigned the accurate score by the model, as a proportion. Although publicly available, accurate and intuitive, accuracy might not capture partial correctness or near-miss responses, particularly when the grading involves a descriptive answer.

Cohen's Kappa

Cohen's Kappa evaluates the inter-rater agreement between the model's assigned scores and the human - assigned scores. As Cohen's Kappa takes into account chance agreement, it is less susceptible to raw accuracy. A larger Kappa score indicates more consistency between human judgment, which is necessary to have buy-in in the educational space.

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE)

For continuous score predictions, we have Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Both RMSE and MAE report how different predicted scores are from expert-assigned scores on average. RMSE treats larger errors more harshly than MAE, making it a better metric when larger errors are relatively more important.

BLEU and ROUGE

When generating textual feedback or summaries of performance is the focus of a model, metrics including BLEU (Bilingual Evaluation Understudy) and (Recall-Oriented Understudy for Gisting Evaluation) may be important. BLEU and ROUGE describe the quality of the generated feedback by measuring whether the predicted feedback is similar (or overlapping) with expert feedback, typically based on some count of words in common and ordered similarity.

The choice of evaluation metric, often depends on the model which evaluates and the type of output. For example, traditional ML models report accuracy or RMSE metrics when relevant, while deep learning models, which can create feedback often report BLEU or ROUGE metrics along with scoring metrics.

Different classes of models employ their evaluation metrics based on their capabilities and their outputs:

- Rule-Based Models typically utilize Accuracy as a measure due to their deterministic and inflexible nature.
- Machine Learning Models (e.g., Support Vector Machines, Logistic Regression) uses Accuracy, Cohen's Kappa (to measure agreement with human raters), and Root Mean Square Error (RMSE) for forecasting continuous scores.
- Deep Learning Models (e.g., LSTM, BiLSTM) often evaluate performance on Accuracy, RMSE, and Mean Absolute Error (MAE) that involve sequence context.
- Transformer-Based Models (e.g., BERT, GPT) integrate traditional scoring metrics such as Accuracy and Cohen's Kappa along with feedback metrics like BLEU and ROUGE, especially when used to generate evaluative comments or summaries.

Comparative Summary of Model Capabilities

This section compares the various models and the metrics used for evaluation. Over time, the field of automated descriptive answer evaluation has grown, and ever-so- slightly different modelling paradigms have emerged which can allow for somewhat different abilities, and specialize in different aspects. In this section, it summarizes key characteristics

of rule-based systems, traditional machine learning models, deep learning models, and transformer-based models, in terms of dimensions such as contextual awareness, flexibility, feedback generation, and scalability.

Rule-based systems are relatively simple, can produce results rapidly, and rely on local knowledge of the domain, but these systems typically suffer from limited ability to address variation in language and styles of semantic meaning.

Traditional machine learning models have improved in flexibility and will learn from data, however, they still rely on the engineered features. Although these models theoretically have deep contextual knowledge, their performance may not be evident until it tests.

Deep learning, particularly RNNs or LSTMs, allow for further delineation in meaning from the sequence of the data and deeper understanding of context which can afford greater dimensionality for meaning.

Table below summarizes standard metrics used across model types:

Table 1: Metrics and Model Comparison

Model Type	Context Understanding	Flexibility	Feedback	Scalability
Rule-Based	Low	Low	No	High
ML (e.g., SVM)	Medium	Medium	Basic	Medium
LSTM/BiLSTM	High	Medium	Yes	Medium
BERT/GPT	Very High	High	Advanced	Low

The Table 1 refers to the metrics used for different ML Models.

Finally, transformer-based models such as BERT and GPT are the latest rendering in this evolution which contain the ability to register boundaries and relationships of meaning, and also have a degree of generativity which can permit generation of natural language textual feedback, but take a lot of resources. The limits of interpretability and fairness are concerns with these models.

Conclusion

Although automated grading systems for descriptive answers have improved significantly, there are several significant challenges and limitations which still impede the technology from broad adoption and functioning efficiently. The limitations that exist are technical, practical and ethical particularly noted with the implementation of more advanced models such as deep learning and transformers.

Lack of Interpretability

Despite the strong performance of deep learning and transformer-based models, many of them operate as 'black box' models. Educators and stakeholders may have difficulty in seeing how models arrive at particular scores or feedback, and this lack of transparency can raise questions of fairness, trust, and accountability, especially in high-stakes testing.

Domain Specificity of ML Models.

Machine learning models have typically been trained on datasets from particular subject domains and will not generalize to all other domains. For example, a model trained on the specific patterns of English literature answers will not give accurate predictions on pressing questions in a computer science domain. This domain specificity reduces the scalability of the models and requires additional data collection and model tuning.

Language Bias and Multilingual Structures

The majority of grading models installed into practice constructs from large language models. In its, much of the text the model has been trained on will be of English-language and thus may have an inherent performance bias to any output from other languages or more likely still in mixed form, code-mixed multilingual structures. Therefore, an effective multilingual grading model has not been appropriately developed and remains an unresolved problem.

Dependence on Large-Scale Labelled Data

Supervised learning models need to have access to larger datasets that are sufficiently labeled otherwise accuracy will decline. Collecting labeled data, and sufficiently training and annotating high quality response sets of descriptive answers with expert assigned scores can be complex and require significant resource commitments. The scarcity of data undermines the viability and long-term sustainability of training and evaluating models that are both reliable and valid.

Real-Time Application Constraints

The deployment of assessments in real-time educational contexts (e.g. a live assessment context) will be much more challenging and excepting (e.g. interactive, interactive box, spaces, etc) your design, models may also be bound by time (e.g. completion time), access computational resources (particularly all large models of BERT and GPT for example), and the complexity of the full integration for feedback will add constraints around limitations, bottlenecks of measures and noticeable latency that will require additional consideration to resolve.

References

1. Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. The Journal of Technology, Learning and Assessment, 4(3).
2. Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, 25(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

4. Page, E. B. (1966). The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5), 238–243.
5. Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.
6. Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1882–1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Curran Associates, Inc.
8. Sukkarieh, J. Z., & Pulman, S. G. (2005). Automatic short answer marking. In Proceedings of the Second Workshop on Building Educational Applications Using NLP (pp. 197–206). Association for Computational Linguistics.
9. Yin, D., Lam, P., & Sinha, T. (2020). Multilingual short answer grading using pretrained language models. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 3298–3307). International Committee on Computational Linguistics.
10. Zhang, M., & Litman, D. (2018). Co-training for automated short answer scoring. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 620–626). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1100
11. Kulkarni, S. A., Joshi, R., Ghayal, S., Deshpande, S., & Shinde, R. (2024). A study on evaluation of theoretical answers using machine learning and generation of model answer using generative artificial intelligence. International Journal of Creative Multimedia.
12. Bahel, V., & Thomas, A. (2021). Text similarity analysis for evaluation of descriptive answers. arXiv Preprint arXiv:2105.02935.
13. Moholkar, K., Chaturvedi, M., Jain, A., & Parkhe, A. (2023). Machine learning techniques for descriptive answer evaluation: A comprehensive survey. International Journal of Advanced Computer Science and Applications, 14(9), 123–132. https://doi.org/10.14569/IJACSA.2023.0140917
14. Deepthi Rani, D., Kumar, A., & Rajasekaran, V. (2023). Intelligent descriptive answer evaluation system. Computer Science, Engineering and Technology Journal, 13(4), 45–52.
15. Susnjak, T. (2024). Beyond predictive learning analytics modelling and onto prescriptive analytics. International Journal of Artificial Intelligence in Education, 34(2), 345–367. https://doi.org/10.1007/s40593-023-00336-3