# On Deep Learning Approaches to Automated Assessment: Strategies for Short Answer Grading

Abbirah Ahmed[a], Arash Joorabchi[b] and Martin J. Hayes[c]
*Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland*

Keywords: Automatic Short Answer Grading, Deep Learning, Natural Language Processing, Blended Learning, Automated Assessment

Abstract: The recent increase in the number of courses that are delivered in a blended fashion, before the effect of the pandemic has even been considered, has led to a concurrent interest in the question of how appropriate or useful automated assessment can be in such a setting. In this paper, we consider the case of automated short answer grading (ASAG), i.e., the evaluation of student answers that are strictly limited in terms of length using machine learning and in particular deep learning methods. Although ASAG has been studied for over 50 years, it is still one of the most active areas of NLP research as it represents a starting point for the possible consideration of more open ended or conversational answering. The availability of good training data, including inter alia, labelled and domain-specific information is a key challenge for ASAG. This paper reviews deep learning approaches to this question. In particular, deep learning models, dataset curation, and evaluation metrics for ASAG tasks are considered in some detail. Finally, this study considers the development of guidelines for educators to improve the applicability of ASAG research.

## 1 INTRODUCTION

In blended learning environments the combination of large student cohorts, the demand for more detailed and timely feedback coupled with constraints on teaching resources, means that the effort required to accurately grade student assessments is becoming increasingly challenging. In this paper we consider how the workload associated with the grading process and the provision of meaningful feedback to students can be assisted using Natural Language Processing (NLP). The focus of the work is the automated interpretation of student answers so as to reduce inconsistencies in the allocation of marks and to ensure fairness in the overall result that is awarded. Notwithstanding these challenges, free text response-based assessments are considered to be one of the preferred grading tools due to their effectiveness in terms of verifying skills and tacit knowledge demonstration.

One of the tests for automated grading of assessments is that it must ease the burden on instructors. In Science and Engineering particularly, different types of assessment will require specific grading methods to be applied that follow a strict rubric or grading criteria wherein NLP can be posited as a useful analysis tool. The success of any automation effort relies on the application of a grading technique that is well defined, repeatable and where the availability of training data is such that it will take less time for an Educator to employ an automated technique than to assess student effort manually.

Automated grading of natural human responses by computers was first discussed by Page (Page, 1966). His proposed Project Essay Grade (PEG) system used a variety of different natural language processing methods. The system evaluated essays using various features, including the length of the response, number of words, and parts of speech tags and applied multiple linear regression to predict the score. The PEG system performed surface feature analysis using syntactical similarity measures (Page, 1966).Since then, machine grading of natural

[a] https://orcid.org/0000-0001-5541-7290
[b] https://orcid.org/0000-0002-0767-4302
[c] https://orcid.org/0000-0001-6821-5436

Ahmed, A., Joorabchi, A. and Hayes, M.
On Deep Learning Approaches to Automated Assessment: Strategies for Short Answer Grading.
DOI: 10.5220/0011082100003182
In *Proceedings of the 14th International Conference on Computer Supported Education (CSEDU 2022) - Volume 2*, pages 85-94
ISBN: 978-989-758-562-3; ISSN: 2184-5026

responses has attracted a large cadre of NLP researchers. Thus far, it is fair to say that most progress has been observed on the automatic scoring of short human responses.

As essays and short answers fall under the category of descriptive and free text answers, it is necessary to differentiate between the two, so that efficient and accurate solutions can be found.

A Short answer can be defined as a piece of text fulfilling the following criteria (Burrows et al., 2015):

- A student response for a given question must be in natural language.
- A response length must be limited to between one sentence to one paragraph.
- A student response must demonstrate the external knowledge which they gained from their understanding and is not identified within the question.
- A response grade should be based on objective content quality criteria and not on subjective writing quality considerations.
- Natural language responses should be capable of being clearly restricted based on the syntax of the assigned question.

In Automatic Short Answer Grading (ASAG), for a given question, student answers are compared with a reference answer(s) and a mark is assigned using ML to ease the workload of instructors and TAs (Mohler, Bunescu, & Mihalcea, 2011; Mohler & Mihalcea, 2009). Although automatic short answer grading is by no means new, a relatively clear state of the art has now emerged for effective grading of solutions that use natural language answers. In many initial studies, ASAG has been considered as a classification or regression task. in which an answer is either labelled as correct or wrong (classification) and/or assigned a mark (regression). In addition, those studies are largely based on manually created patterns and text similarity algorithms (Mohler et al., 2011; Mohler & Mihalcea, 2009; Sultan et al., 2016). Over the past few years, researchers have started to employ deep learning methods for ASAG due to their proven efficacy in many NLP domains and tasks.

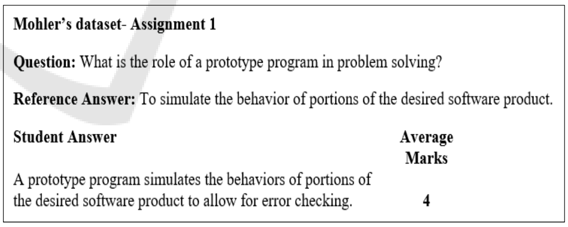
In this paper, we compare a range of deep learning methods used for ASAG, using publicly available datasets that have been widely cited and have compared a variety of ASAG evaluation metrics. In addition to providing a benchmark comparison of existing approaches, we suggest a framework for educators who wish to determine a reliable ASAG assessment strategy for their students in this domain.

# 2 DATASETS

Most NLP tasks (e.g., text classification, named entity recognition, sentiment analysis) have a number of de-facto standard datasets which are used to benchmark the performance of new methods and techniques for these tasks. However, for the task of ASAG, one of the biggest challenges is the lack of appropriate datasets. In literature, there are some publicly available benchmark datasets which have been used to evaluate the performance of different ASAG systems. In this study we only discussed those datasets which are used to benchmark deep learning-based systems. These datasets are diverse in terms of topics, size, and grading scale. It is observed that some well-known datasets were launched through competitions including the ASAP and SemEval-2013 datasets

## 2.1 Mohler's Dataset

This dataset is based on the assignments of an undergraduate course on data structures at University of Texas and it was released in 2009 (Mohler & Mihalcea, 2009). There are three assignments, seven questions each for the class size of 30 students. Thus, the size of dataset is 630 student answers. These answers are graded by two human instructors independently on the scale of 0-5; 0 indicates the completely wrong answer and 5 indicates correct answer. Fig.1 shows an example from Mohler's dataset



**Mohler's dataset- Assignment 1**

**Question:** What is the role of a prototype program in problem solving?

**Reference Answer:** To simulate the behavior of portions of the desired software product.

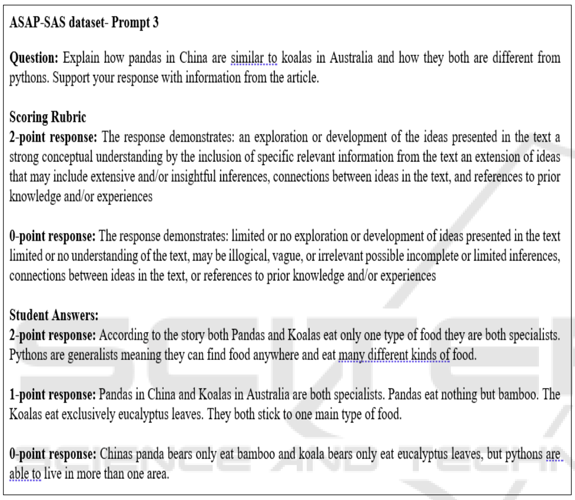
| Student Answer | Average Marks |
|---|---|
| A prototype program simulates the behaviors of portions of the desired software product to allow for error checking. | 4 |

Figure 1: Example from Mohler's dataset.

In 2011, authors released and extended dataset for ASAG task (Mohler et al., 2011). This extended version contains student answers to 10 assignments and two exam papers. Each assignment consists of four to seven questions and each exam paper consists of 10 questions. There are 81 questions and 20 answers per question which sums up to 1620 question-answer pairs in total. Each answer is graded by two independent markers and average of their score is considered as final score. This dataset is publicly available[1].

## 2.2 ASAP-SAS Dataset

Automated Student Assessment Prize- Short Answer Scoring corpus was launched by The Hawlett Foundation on Kaggle[2] . It contains responses from 8th grade to 10th grade students and length of responses are less than 50 words. The dataset consists of 10 prompts, one for each question. There are combined 17204 responses which are marked over two scales 0-2 and 0-3. This dataset also contains the marking rubric for each prompt. Fig.2 represents an example form the ASAP dataset.



**ASAP-SAS dataset- Prompt 3**

**Question:** Explain how pandas in China are similar to koalas in Australia and how they both are different from pythons. Support your response with information from the article.

**Scoring Rubric**
**2-point response:** The response demonstrates: an exploration or development of the ideas presented in the text a strong conceptual understanding by the inclusion of specific relevant information from the text an extension of ideas that may include extensive and/or insightful inferences, connections between ideas in the text, and references to prior knowledge and/or experiences

**0-point response:** The response demonstrates: limited or no exploration or development of ideas presented in the text limited or no understanding of the text, may be illogical, vague, or irrelevant possible incomplete or limited inferences, connections between ideas in the text, or references to prior knowledge and/or experiences

**Student Answers:**
**2-point response:** According to the story both Pandas and Koalas eat only one type of food they are both specialists. Pythons are generalists meaning they can find food anywhere and eat many different kinds of food.

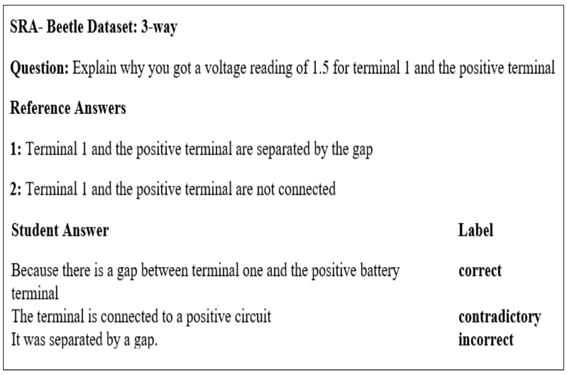
**1-point response:** Pandas in China and Koalas in Australia are both specialists. Pandas eat nothing but bamboo. The Koalas eat exclusively eucalyptus leaves. They both stick to one main type of food.

**0-point response:** Chinas panda bears only eat bamboo and koala bears only eat eucalyptus leaves, but pythons are able to live in more than one area.

Figure 2: Example from ASAP dataset.

## 2.3 SRA Dataset

This dataset was introduced in 2013 by SemEval (Semantic Evaluation) workshop and contains two subsets: SciEntBank (SEB) and Beetle (Dzikovska et al., 2013). The Beetle dataset consists of almost 3000 student responses to 56 questions recorded during interactions with a dialogue system. The SciEntBank dataset comprises of 10,000 student answers to 192 questions covering 16 science subjects of 3rd to 6th grades. The datasets are labelled in 2-way: Correct, Incorrect, 3-way (See Fig.3): Correct, Contradictory, Incorrect and in 5-way: Correct, Partially Correct, or Incomplete, Contradictory, Irrelevant or Not-In-Domain. The SciEntBank test set consists of three subsets: Unseen Questions (UQ), Unseen Answers (UA) and Unseen Domain (UD). UQ dataset consists of in-domain but unseen questions, UA dataset contains answers to questions which are present in training dataset, UD consists of questions and answer which are out of domain. Whereas Beetle test set consists of two subsets: Unseen Questions (UQ) and Unseen Answers (UA).



**SRA- Beetle Dataset: 3-way**

**Question:** Explain why you got a voltage reading of 1.5 for terminal 1 and the positive terminal

**Reference Answers**

**1:** Terminal 1 and the positive terminal are separated by the gap

**2:** Terminal 1 and the positive terminal are not connected

| Student Answer | Label |
|---|---|
| Because there is a gap between terminal one and the positive battery terminal | correct |
| The terminal is connected to a positive circuit | contradictory |
| It was separated by a gap. | incorrect |

Figure 3: Example from SRA dataset.

## 3 EVALUATION METRICS

Based on the design of an ASAG system as a classification or regression model, different evaluation metrics are used. This section provides an overview of commonly used performance metrics used for evaluating ASAG models.

### 3.1 Pearson's r Correlation

This is used to measure the correlation between two numerical variables. The values assigned through this method range from -1 to 1, where 1 indicates positive correlation, 0 indicates no correlation and -1 is for negative correlation. It is calculated by eq.1:

$$r = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y}\ )}{\sqrt{\sum_{1=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}} \qquad (1)$$

Where for two distributions $\boldsymbol{X}$and $\boldsymbol{Y}$, $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ are the $\mathbf{i^{th}}$ value of distributions and $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the mean values for both distributions respectively.

For the task of ASAG, Pearson's correlation is one of the most popular correlation measures which is used to compare the marks assigned by instructors with predicted marks.

[1] http://web.eecs.umich.edu/~mihalcea/downloads/ShortAnswerGrading_v1.0.tar.gz

[2] https://www.kaggle.com/c/asapsas/leaderboard/public

### 3.2 Root Mean Square Error

It is another measure to calculate the error value between predicted and observed values. RMSE score is calculated by eq.2:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_p - x_o)^2} \quad (2)$$

Where $x_p$ is predicted value and $x_o$ is observed value.
For RMSE, lower value indicates better results.

### 3.3 F1 Score

It is an evaluation metric for classification which combines precision and recall. It is the weighted average of precision and recall and ranges between 1(best) and 0(poor).

$$\text{F1} = \frac{2*(Precision*Recall)}{(Precision+recall)} \quad (3)$$

*Precision:* It is the ratio of the correct predictions made by model to the total predictions:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

*Recall:* it is the ration of correct predictions made by the model to the actual labels.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

*Macro-average F1*: It calculates average of F1 score per class by eq.6.

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=0}^{N}(F1\ score)_i \quad (6)$$

*Micro-average F1:* It is the harmonic mean of the precision and recall for each individual class.

$$\text{Micro-F1} = \frac{2*(Micro\ Precision*Micro\ Recall)}{(Micro\ Precision+Micro\ Recall)} \quad (7)$$

$$\text{Micro Precision} = \frac{\sum TP_i}{\sum TP_i + \sum FP_i} \quad (8)$$

$$\text{Micro Recall} = \frac{\sum TP_i}{\sum TP_i + \sum FN_i} \quad (9)$$

### 3.4 Quadratic Weighted Kappa

This is a measure to find the inter-rater agreement for expected and predicted scores and can be calculate by:

$$\text{k} = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \quad (10)$$

Where, $\boldsymbol{p_0}$ is the observed agreement and $\boldsymbol{p_e}$ is the expected agreement. Its value is 1 for complete agreement when the expected and predicted scores are same and 0 for the complete disagreement.

## 4 DEEP LEARNING APPROACHES FOR ASAG

For a long time, traditional machine learning approaches were widely used for measuring text similarity and provided promising results in various NLP tasks including Machine translation, text summarization and automatic scoring. Earlier ASAG models used traditional text similarity methods such as corpus-based and string-based similarity measures combined with feature engineering and simple text vectorization methods. Nevertheless, manual generation of features using regular expressions is a time-consuming process and requires domain expertise. Besides, these systems are trained on high dimensional sparse vectors which makes them computationally expensive. Since 2016, Deep learning architectures gained popularity over traditional machine learning approaches for text similarity-based tasks, in particular Automatic Essay Scoring and Automatic Short Answer Grading (Camus & Filighera, 2020; Ghavidel et al., 2020; Gomaa & Fahmy, 2020; Hassan et al., 2018; Kumar et al., 2017; Prabhudesai & Duong, 2019; Saha et al., 2018; Sasi et al., 2020; Sung et al., 2019; Surya et al., 2019; Tulu et al., 2021). Multiple authors also studied word embedding models with deep learning approaches. In this paper, we have categorized the work in this domain into three categories: Early Neural Network-based, Attention based, and Transformer-based ASAG architectures.

### 4.1 Early Neural Network Approaches

Earlier ASAG systems were based on features engineering while many recent ASAG systems are neural network-based eliminating the need of feature engineering. In 2017, a novel framework was proposed which combined Siamese bi-LSTM for student and model answers, Earth-Mover distance-based pooling layer for all hidden states, and a support vector output layer for scoring. Additionally, they used task-specific data augmentation for the training process. This model was tested on Mohler's and SemEval's dataset and outperformed various baselines (Kumar et al., 2017). In 2017, researchers experimented with neural architectures for ASAG task, which were previously used for Automatic Essay Scoring (AES) (Riordan et al., 2017). The experiments were performed on three publicly available datasets: ASAP-SAS, Powergrading (Basu et al., 2013), and SemEval. They investigated multiple research questions through their work: (a) how pretrained embeddings with fine-tuning perform; (b) whether the convolution layer will be able to generate minimum features for ASAG task; (c) is it possible to use smaller hidden layers; (d) the role of bi-directional LSTM and attention mechanism. According to their findings, a basic neural architecture with tuned pre-trained embeddings and LSTMs is a relatively effective architecture for ASAG. In 2018, a bi-LSTM based deep learning model was proposed utilizing paragraph embeddings for the vector representation of answers (Hassan et al., 2018). In their work, authors investigated multiple word embeddings including Word2Vec, GLoVe, fastText and paragraph embeddings such as Doc2Vec, Infersent and Skip-thought. With word embedding models, they used sum of word vectors to generate paragraph embedding models for student and model answers. However, among all the embedding models, highest Pearson's correlation score of 0.569 was achieved by the Doc2Vec paragraph embedding model. In 2019, another neural model was developed using Siamese bidirectional LSTM (bi-LSTM) and hand-crafted features (Prabhudesai & Duong, 2019). These features included the length of student response, length ratio of student and model answers, total number of words and unique words in student response. Researchers used GloVe embeddings at the embedding layer and used data augmentation of reference answers in training as novel features of their work. To evaluate the performance, they used Mohler's dataset. In another study, Saha et al. proposed a joint multi domain neural architecture (JMD-ASAG) for ASAG (Saha et al., 2019) . This architecture is based on bi-LSTMs, generic scorer, and domain specific scorers. The domain adaptation is performed by learning generic and domain specific characteristics using limited task and domain-specific training data. In 2020, another deep learning-based method for automatic short answer grading called Ans2Vec has also been proposed (Gomaa & Fahmy, 2020). This approach has based on Skip-thought vectorization method to convert model and student answers into semantic vectors. The model was tested for Mohler's dataset, SciEntBank dataset, and Cairo university dataset (Gomaa & Fahmy, 2014). Recently in 2021, one more deep neural network architecture approach was proposed by combining Manhattan LSTMs and SemSpace vectors (Tulu et al., 2021). These vectors were derived from the WordNet database to predict student grades from the ASAG datasets. The system consists of two identical LSTM networks, where each sentence pair of student and model answer is fed to the system as sense vectors and vectorial similarity is calculated by Manhattan distance at output. To evaluate the system performance, they used two datasets: Mohler's dataset and Cukurova University-NLP (CU-NLP) dataset which was specifically prepared for this study. For experimentation, they prepared separate CSV file for each question containing corresponding student and reference answers. They achieved 0.95 Pearson's correlation for the majority files.

## 4.2 Attention-based Approaches

In the Deep Learning domain, greatest achievements in the last decade have been the advent of the attention mechanism. Since its inception, many NLP advances have been made including Transformers such as Google's BERT (Bidirectional Encoder Representations from Transformers). Attention's main objective is to construct the context vectors required by the decoders by considering all the states of intermediate encoders. In 2019, an attention-based framework was presented which extracts semantic information from student and model answer without the need of feature engineering (Liu et al., 2019). Based on transformer layers and multiway attention mechanism, this model provides enhanced semantic relationship between words in a sentence. This framework was evaluated on K-12 dataset. In their paper, published in 2019, authors (Gong & Yao, 2019) presented another attention-based deep model for ASAG. To learn sentence vector representation of student and model answers, this model utilizes pre-trained word embeddings and BiRNN with LSTM and attention-mechanism. This system is tested on K-12 dataset and achieved 10% increased performance compared to baseline models. Similarly, another attention-based neural architecture was designed for ASAG and evaluated on ASAP-SAS dataset (Xia et al., 2020). In this architecture, researchers combined Google Word Vector (GWV)[3] and attention mechanisms using a BiLSTM neural network. In contrast to several

[3] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing

baseline models, this model provides achieved average QWK value of 0.70 for short answer scoring.

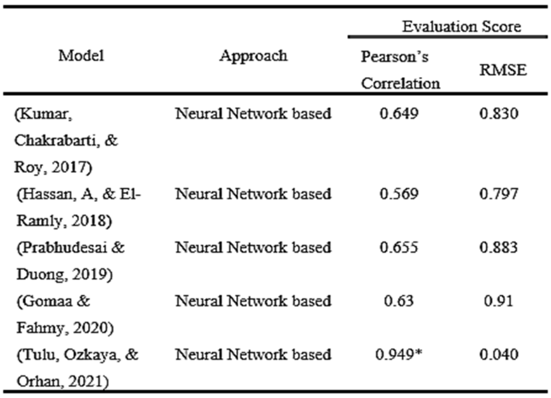
### 4.3 Transformer-based Approaches

In 2017, a novel architecture called Transformers was presented in the paper "Attention Is All You Need" (Vaswani et al., 2017) which also leverages the attention mechanism. Later, in 2018, a new transformer-based language representation model called BERT (Devlin et al., 2018) was introduced. By conditioning on both left and right context simultaneously in all layers, it aims to learn deep bidirectional representations from unlabelled text. Since then, NLP researchers began to investigate its effectiveness in different downstream NLP tasks including ASAG. In 2019, BERT was used to improve the automatic short answer scoring task (Sung et al., 2019). The performance of BERT for ASAG was evaluated using the SciEntBank and Psychology domain datasets from SemEval-2013 and found that pretrained transfer learning models performed 10% better than the classical methods. It was further observed that a model fine-tuned on single domain data was not suitable for other domains, however, a single model can be fine-tuned on multiple domains. In 2020, the above work was extended by analysing different transformer-based architectures such as BERT and its variants on SemEval dataset (Camus & Filighera, 2020). It was observed that models trained on different datasets can be used for ASAG task using the transfer learning approach. In addition, it was discovered that training models using multiple languages can improve their performance. Moreover, transformer-based models demonstrate better performance and generalization capabilities. In another study presented in 2020, multiple data augmentation strategies were combined with BERT for ASAG (Lun et al., 2020). By employing three data augmentation strategies including back translation, correct answer as reference answer and swap content, authors enhanced the sentence representation and further improved the performance of the model with fine-tuned BERT. For the system evaluation they used SemEval dataset. An autoregressive pre-training architecture that uses BERT and XLNET (Extra Long-Network) was proposed in 2020 as another reference answer-based model (Ghavidel et al., 2020). Unlike previously described works, this approach does not utilize any manually crafted features neither use questions for training or as input to the system. The system's performance was evaluated on SciEntBank dataset provided by SemEval-2013.According to the observations, both models performed similarly by providing semantic relationship between student answer and model answer yet outperformed previous state of the art approaches.

## 5 DISCUSSION

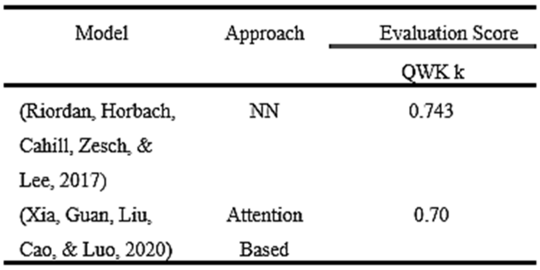
Most ASAG systems require a 'deep' learning model that consists of a number of training layers and incorporating a training phase on a corpus that is sufficiently representative of a broad cross section of good and bad sample answers. There are a variety of factors that will determine a system's ASAG effectiveness including, inter alia, the efficiency of the system training stage, processing time allowed to infer the final scores, the incorporation of deep layer fine-tuning on specific questions in the training phase, scalability, and the ability to regenerate the results on similar, yet different, datasets (Bonthu et al., 2021). In contrast to conventional feature engineering based ASAG systems, deep learning models have been shown to provide better results in terms of accuracy, semantic similarity, computational cost, and generalizability. From the models reviewed in this study, it is observed that Attention-based and BERT based models outperform alternatives for the ASAG task on Mohler, ASAP-SAS and SRA datasets. Furthermore, recent advancements within the transformer and pre trained language model literature can provide excellent performance in terms of efficacy. For instance, Transfer learning mechanism have been shown to reduce the requirement for broad, yet still domain specific training data and excellent adaptations in relation to generalization have been reported. Based on the results presented in Table 1, it is clear that rudimentary LSTM-based models provided good results using Mohler's dataset and are computationally efficient. Using a separate training file for each question, it was shown to achieve the highest Pearson score (with this corpus) of 0.94 (Tulu et al., 2021). Conversely, when the system was trained with a single training file that includes all questions, answers, and reference answers, the Pearson correlation reduces markedly to 0.15. Evidently, an extension to a large dataset, with an attendant growth in context training set, causes an ASAG system to train far more slowly and also exhibits a much lower success rate when a large amount of Out-Of-Vocabulary (OOV) words appear in the dataset. This has clear negative ramifications for the consideration of LSTM methods in this use case. Other models that employ word or sentence embeddings e.g., Gomaa et. al (Gomaa & Fahmy, 2020) have been shown to perform well in this use case. Such approaches, that utilize embeddings of student and reference answers preserve the semantic and syntactic relation among words. In Table 2 it can be seen that an attention-based model outperforms an LSTM using the ASAP-SAS dataset. Moreover, most of the attention-based models that have been evaluated on the K-12 dataset have demonstrated even better results. This provides significant actionable feedback in relation to the curation of data for future blended learning trials. In general, further improvements in performance for the ASAG use case has been achieved using transformer-based models that have been trained and tested on the SemEval dataset and as shown in Table 3. Most of these models have used the finetuning of BERT and its variants. The system presented by Lun et. al (Lun et al., 2020)outperformed other state-of-the-art models using a combination of the classical BERT multi-layer deep contextualized language model introduced in 2018, combined with (multiple) training data augmentation strategies. It is pretrained on Wikipedia and a large Book Corpus that learns the contextual relationship between sentences which plays a significant role in text similarity-based task such as ASAG. While most of these systems are capable of producing good results, the generalized performance of these systems is still questionable for the case of completely new student cohorts who may provide answers on new question banks that are significantly different to the (limited) datasets that have been considered in this work.

Table 1: Performance scores for Mohler's Dataset.



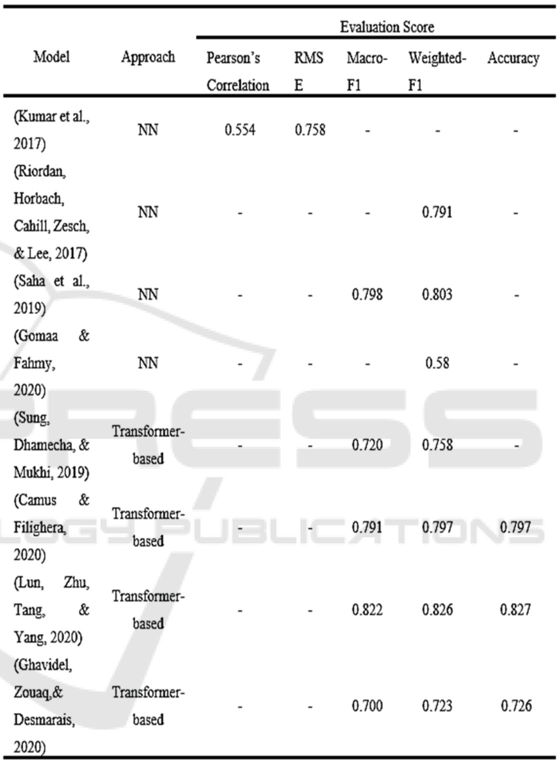
| Model | Approach | Evaluation Score | |
|---|---|---|---|
| | | Pearson's Correlation | RMSE |
| (Kumar, Chakrabarti, & Roy, 2017) | Neural Network based | 0.649 | 0.830 |
| (Hassan, A, & El-Ramly, 2018) | Neural Network based | 0.569 | 0.797 |
| (Prabhudesai & Duong, 2019) | Neural Network based | 0.655 | 0.883 |
| (Gomaa & Fahmy, 2020) | Neural Network based | 0.63 | 0.91 |
| (Tulu, Ozkaya, & Orhan, 2021) | Neural Network based | 0.949* | 0.040 |

Table 2: Performance scores for ASAP-SAS Dataset.



| Model | Approach | Evaluation Score |
|---|---|---|
| | | QWK k |
| (Riordan, Horbach, Cahill, Zesch, & Lee, 2017) | NN | 0.743 |
| (Xia, Guan, Liu, Cao, & Luo, 2020) | Attention Based | 0.70 |

Table 3: Performance scores for SemEval's Dataset.



| Model | Approach | Evaluation Score | | | | |
|---|---|---|---|---|---|---|
| | | Pearson's Correlation | RMSE | Macro-F1 | Weighted-F1 | Accuracy |
| (Kumar et al., 2017) | NN | 0.554 | 0.758 | - | - | - |
| (Riordan, Horbach, Cahill, Zesch, & Lee, 2017) | NN | - | - | - | 0.791 | - |
| (Saha et al., 2019) | NN | - | - | 0.798 | 0.803 | - |
| (Gomaa & Fahmy, 2020) | NN | - | - | - | 0.58 | - |
| (Sung, Dhamecha, & Mukhi, 2019) | Transformer-based | - | - | 0.720 | 0.758 | - |
| (Camus & Filighera, 2020) | Transformer-based | - | - | 0.791 | 0.797 | 0.797 |
| (Lun, Zhu, Tang, & Yang, 2020) | Transformer-based | - | - | 0.822 | 0.826 | 0.827 |
| (Ghavidel, Zouaq,& Desmarais, 2020) | Transformer-based | - | - | 0.700 | 0.723 | 0.726 |

## 6 FUTURE DIRECTIONS

Through this study, several conclusions have been drawn for the ASAG use case. These observations not only concern the main features of system evaluation, but also consider future research directions which go beyond the 'nuts and bolts' issues of datasets, models, and evaluation metrics.

### 6.1 Datasets

It is observed that currently few and limited datasets are available for the task of ASAG, and those available datasets cover very a handful of domains such as, science (SemEval) and computer science (Mohler). Also, these datasets are collected from different education levels which limits their generalization potential. Furthermore, depending on the number of times that a course is offered and the number of enrolled students per offering, amount of available data can be limited which could create further problems in training a model. Moreover, most of the online education and distance learning programs are based on university level courses which involve programming and numerical solutions. Therefore, it is important to develop new, domain specific datasets for various levels which cover programming and numerical type data.

## 6.2 Model Building

In recent years, various state of the art models has been introduced following the introduction of transformer architecture, such as Reformer (Kitaev, Kaiser, & Levskaya, 2020), pre-trained language models such as GPT (Radford et al., 2018) and its variants, BERT (Devlin et al., 2018), XLNET (Yang et al., 2019), T5 (Raffel et al., 2019) and ELECTRA (Clark et al., 2020). The potential of using these models for the task of ASAG should be investigated.

## 6.3 Perspective of Stakeholders

In an educational environment, multiple entities are affected by the implementation and deployment of ASAG systems. The performance of an ASAG system must meet the expectation of these entities also called stakeholders (Madnani & Cahill, 2018), which involves Students/Test-takers, Teachers/Examiners, Subject-experts, NLP Researchers, and Educational Technology (EdTech) Companies. Assessments and grades significantly impact the future of the students. For instance, when considering matriculation to the next grade level or the achievement of base levels of performance for the completion of a particular learning outcome a floor level or 'pass' mark needs to be established. All stakeholders demand reliable and accurate scoring systems. Further, instant student feedback on scores as well as deeper feedback identifying the reasons for the award of marks are now required. For ASAG systems to be adopted they must be trusted by teachers, particularly in the case of formative assessment. Subject experts must assume responsibility for the careful design of possibly fewer rich assessments and scoring rubrics that tightly describe the marking criteria so that automated grading can be deployed successfully. There is a clear trade-off between the rubrics that are necessary for complex open-ended responses, the design time for a (potentially) more limited form of assignment to be graded automatically and the manual assessment of student material. Therefore, clear guidelines need to develop for teachers so that automated methods can be deployed successfully. Additionally, for longer form answers existing benchmark datasets will need to be labelled i.e., graded, typically by two human graders marking the responses and taking the average of both scores as a 'golden' score. Hence when considering extensions to ASAG use cases, it is essential 12 to quantify the level of human intervention and accuracy measurements required to deliver acceptable system performance. For NLP engineers, it is important to build tools that will be adopted widely with a minimum of local configuration. The development of a clear roadmap for adopters is critical in this regard.

## 6.4 Integration of ASAG System with Learning Management Systems (LMS)

Most LMS offered by educational institutions now provide various features such as online discussion forums, study progress tracking, creation of online assessments, and user feedback. Integrating automated grading support integrated within commercial LMSs poses multiple challenges. The integrate of new tools into such systems may impose significant extra costs. Moreover, although some of these systems already provide automated assessments these are generally limited to MCQs and filling in tightly constrained blank spaces. Grading of free-text responses is still a significant open challenge in these applications. To address these problems, an open-source LMS such as Moodle, modified to include an ASAG component, is the obvious starting point. How limiting is such a starting point in terms of reducing the available marker of adopters? The consideration of common ethical issues regarding fairness, validity and plagiarism when developing an ASAG system that can be integrated within a standalone LMS also raises liability questions when any mistakes are made.

## 6.5 Creating a Roadmap for ASAG Tool Deployment

It is necessary for NLP researchers and subject experts to work together to develop custom ASAG systems that incorporate the appropriate features for widescale deployment. An ASAG system should be user-friendly enough so that instructors from a wide range of technical disciplines feel confident that they can adopt it without the aid of developers. It must be capable of being deployed as a web application or as a tool without an explicit programming support environment. Thirdly, it must be a fast, cost-effective, trustworthy, scalable, and generalizable so that the system keeps pace with technological advances.

## 7 CONCLUSIONS

In this paper we provide an overview of the recent deep learning-based solutions for the task of Automatic Short Answer Grading and its relevance in the educational setting. We also reviewed available benchmark datasets and evaluation metrics and discussed their major shortcomings. We showed that recently adopted transfer learning and transformer-based models outperform earlier neural network-based models used for ASAG. Nonetheless, the application of the latest transformer-based models such as GPT-2, GPT-3, T5, and XLNET still remain to be explored in this context. In this paper, several possible future directions that can present a barrier to widescale deployment have been identified. The interests of various stakeholders, the need for new dataset curation guidelines and the pressing need for more user-friendly interfaces to enable the adoption of ASAG systems have been highlighted.

## REFERENCES

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics, 1*, 391-402.

Bonthu, S., Rama Sree, S., & Krishna Prasad, M. (2021). *Automated Short Answer Grading Using Deep Learning: A Survey.* Paper presented at the International Cross-Domain Conference for Machine Learning and Knowledge Extraction.

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.

Camus, L., & Filighera, A. (2020). *Investigating transformers for automatic short answer grading.* Paper presented at the International Conference on Artificial Intelligence in Education.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., et al. (2013). *Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge*: NORTH TEXAS STATE UNIV DENTON.

Ghavidel, H. A., Zouaq, A., & Desmarais, M. C. (2020). *Using BERT and XLNET for the Automatic Short Answer Grading Task.* Paper presented at the CSEDU (1).

Gomaa, W. H., & Fahmy, A. A. (2014). Arabic short answer scoring with effective feedback for students. *International Journal of Computer Applications, 86*(2), 35-41.

Gomaa, W. H., & Fahmy, A. A. (2020). Ans2vec: A Scoring System for Short Answers (pp. 586-595): Springer International Publishing.

Gong, T., & Yao, X. (2019). An attention-based deep model for automatic short answer score. *International Journal of Computer Science and Software Engineering, 8*(6), 127-132.

Hassan, S., A, A., & El-Ramly, M. (2018). Automatic Short Answer Scoring based on Paragraph Embeddings. *International Journal of Advanced Computer Science and Applications, 9*(10).

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Kumar, S., Chakrabarti, S., & Roy, S. (2017). *Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading.* Paper presented at the IJCAI.

Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G. Y., & Liu, Z. (2019). *Automatic short answer grading via multiway attention networks.* Paper presented at the International conference on artificial intelligence in education.

Lun, J., Zhu, J., Tang, Y., & Yang, M. (2020). *Multiple data augmentation strategies for improving performance on automatic short answer scoring.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Madnani, N., & Cahill, A. (2018). *Automated scoring: Beyond natural language processing.* Paper presented at the Proceedings of the 27th International Conference on Computational Linguistics.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). *Learning to grade short answer questions using semantic similarity measures and dependency graph alignments.* Paper presented at the Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies.

Mohler, M., & Mihalcea, R. (2009). *Text-to-text semantic similarity for automatic short answer grading.* Paper presented at the Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009).

Page, E. B. (1966). The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan, 47*(5), 238-243.

Prabhudesai, A., & Duong, T. N. B. (2019, 10-13 Dec. 2019). *Automatic Short Answer Grading using Siamese Bidirectional LSTM Based Regression.* Paper presented at the 2019 IEEE International Conference on Engineering, Technology and Education (TALE).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. (2017). *Investigating neural architectures for short answer scoring.* Paper presented at the Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications.

Saha, S., Dhamecha, T. I., Marvaniya, S., Foltz, P., Sindhgatta, R., & Sengupta, B. (2019). Joint multi-domain learning for automatic short answer grading. *arXiv preprint arXiv:1902.09183*.

Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., & Sengupta, B. (2018). *Sentence level or token level features for automatic short answer grading?: Use both.* Paper presented at the International conference on artificial intelligence in education.

Sasi, Nair, D., & Paul. (2020). Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading. *arXiv pre-print server*.

Sultan, M. A., Salazar, C., & Sumner, T. (2016). *Fast and easy short answer grading with high accuracy.* Paper presented at the Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). *Improving short answer grading using transformer-based pre-training.* Paper presented at the International Conference on Artificial Intelligence in Education.

Surya, K., Gayakwad, E., & Nallakaruppan, M. (2019). Deep learning for short answer scoring. *Int. J. Recent. Technol. Eng.(IJRTE), 7*(6).

Tulu, C. N., Ozkaya, O., & Orhan, U. (2021). Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM. *IEEE Access, 9*, 19270-19280.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Xia, L., Guan, M., Liu, J., Cao, X., & Luo, D. (2020). *Attention-Based Bidirectional Long Short-Term Memory Neural Network for Short Answer Scoring.* Paper presented at the International Conference on Machine Learning and Intelligent Communications.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.