International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Employing Transfer Learning for the Automation of Short Answer Grading

Nikunj Chunilal Gamit[1], Jignesh Haren Kumar Vaniya[2], Naimisha Shashikant Bhai Trivedi[3], Chetna Ganesh Chand[4], Nirav Kumar Bharat Kumar Suthar[5]

Submitted:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

Abstract: To accommodate the ever-increasing number of pupils, automated short answer grading (ASAG) has recently attracted interest in the field of education. We take a look at the latest developments in ASAG research as it relates to the impact of recent advances in ML and NLP on the discipline. In this study, we add to the existing literature by giving a thorough evaluation of newly published methods that use deep learning techniques. We focus in on the shift from features that are hand-engineered to representation learning methods, which automatically learn task-specific features from massive data sets. Word embeddings, sequential models, and attention-based strategies frame our examination of deep learning methods. We found that learned representations alone do not help to produce the greatest outcomes, but they rather function in a complimentary fashion with hand-engineered features, which is how deep learning affects ASAG differently from other domains of natural language processing. Combining the strength of the semantic descriptions offered by modern models with the meticulously constructed characteristics, such as in transformer designs, is undoubtedly the key to top performance. We highlight problems and suggest a future research agenda for tackling them.

Keywords: *ASAG, ML, NLP, Language, Processing, Deep Learning, Data Sets*

1. Introduction

The success method to gauge students' mastery of the concepts used to be the written answers that have proved their merit since olden times. Short answer questions challenge students to express their own way of understanding in their own words therefore, these questions give us a glimpse of the depth of knowledge that the students have beyond what the multiple choice assessments can reveal. Nevertheless, the task of moderation of the essay-type answers necessarily laborious, and the inspectors get the enormous job after it. Consequently, it means these things cannot be evaluated as frequently as they should be and nor can they be evaluated widely. The same answer is still not rated the same by different human graders. The advantage of automated grading for short answers is that it would assign precise, former equations and high-speed feedback to the students. This would thus reduce instructors' workload. [1] Previous efforts to automated grading were strongly anchored on knowledge of the surface nature of an answer/response such as word matching, n-gram overlying and manual guidelines provided against a reference textAlthough this can point out the obvious concordance or discordance, it is actually a narrow and inflexible analysis. Semantically valid responses constructed using synonyms, wrong answers that are built from the right elements, and creative interpretations which include nuances would demonstrate instances where such methods would be incapable of generating responses which are coherence and semantically suitable. [2]

Several papers on robotic short answer grading have been published as many researchers investigate how the exceptional linguistic skillsets developed in the areas such like question answering and procedural generation can be adapted to meet the specific requirements of automatic short answer grading. The objective can be viewed in a variety of ways: either the task can be cast as a semantic similarity computation, or multi-task learning where rubrics and grades are manually designed for augmentation, or meta-learning in which graduating evaluating methodology was derived from the data have shown potential. This first part is a summary of the present situation with respect to the usage of transfer learning in the field of automated short answer grading.[3]

[1]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0009-0004-2821-4606
[2]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0009-0002-4121-0074
[3]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0000-0002-8395-1586
[4]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0009-0003-1876-8472
[5]*Computer Engineering Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0009-0004-2820-9786
**CorrespondingAuthorEmail:nikunjgamit@gmail.com*

2.Review of literature

BayerM (2022) [4] When data is few, it becomes much more difficult to construct a categorization model. The problem may be solved using the domain knowledge that the PLMs have gained. After being pretrained with massive datasets spanning several domains, PLMs need fine-tuning using task-specific in-domain datasets. When compared to pretraining from scratch, the data needed to fine-tune PLMs is small ().

Burrows et al., (2022) [5]There have been several advancements in the ASAG challenge, including neural networks and more conventional methods (). Previous research only included students' written comments. Since some subjects like mathematics, physics, and programming include both textual and non-textual characters like symbols, it is necessary to concentrate on these areas when doing the ASAG job. To address ASAG in this area, it is crucial to use new technology.

3. Objectives

- To provide an extensive discussion of the current research developments in automated short answer grading (ASAG), emphasizing mostly deep learning methods and particularly the representation learning techniques, which replace handcrafted features.
- To identify the challenges and the future research directions in ASAG, particularly in the area of combining hand-engineered features with the semantic representations learnt by deep learning models, such as transformer architectures, to achieve the best performance.

4. Statement of the problem

The automatic grading of short answer responses is a difficult natural language processing problem with significant applications in education. Human graders are able to understand the semantics and intent behind a student's answer and assess whether it demonstrates the required knowledge, even if the wording differs from a reference answer. Nevertheless, implementing these automated systems which can grade short answers with human-like comprehension is still an open problem. The common techniques based on lexical similarity or pattern matching usually are fragile, which fail at meaning extraction. Transfer learning, where models pre-trained on large text corpora are fine-tuned on the target task, has been successful in grading by using representations that encode semantic knowledge.

5. Significance of the study

The fact that short answer responses can be automatically and accurately graded has great implications for education at all levels. Manual grading is labour-consuming, expensive, and vulnerable to mistakes. An efficient grading system that is automated would be an opening to a more frequent assignments giving more rapid feedback loops to boost student engagement and learning. In addition to the conventional multiple-choice tests, short answer questions are more effective in assessing a student's understanding of the concepts and their ability to express them in their own words. Therefore, as a key component of a holistic assessment, devising automated approaches to grade handled responses is a must. Transfer learning employing large pre-trained language models which encode general semantic knowledge that could then be transferred to the grading task via fine tuning in consideration of context specificity shows particular promise. If the project is successful, this could overcome the brittleness of hand-crafted rules and shallow similarity metrics.

6. Research methodology

Three sections make up the survey. In the first section, we go over several key benchmark data sets that are used to test ASAG algorithms. With their unique qualities, the data sets enable us to assess several facets of ASAG systems' generalizability, such as the quality of responses to questions that were not present during training or that pertain to unfamiliar subjects or domains.

In Part 2 of the survey we follow the development and enhancement of methods depending on traditional machine learning techniques and manually constructed feature sets, and in Part 3 we follow the development and improvement of methods based on more contemporary deep learning techniques. Their ability to extract representations of text that are rich in semantic information are of great interest to us. We provide an in-depth analysis of newly published approaches and discuss current advances and trends in ASAG.

Scope

Automated grading of brief replies, as stated by the following criterion, is our primary area of interest:

(1) Rather than just reporting passages from the supplied prompt text, the response should represent the student's knowledge.

(2) The answer should be offered in natural language.

(3) The answer should be roughly 50 words long, although it may include up to around 1,000 words.
(4) The answer's substance, not its quality of writing, is graded;
(5) The question's closed-form constrains the range of viable solutions.

The majority of ASAG methods that have been published in the last few years have been developed to aid in efficient grading in educational settings and are based on supervised learning methodologies. The goal of using unsupervised techniques like ranking or clustering to group comparable student responses was to make grading more consistent; these approaches may be used in conjunction with supervised learning methods.

When it comes to actual exams, the question pool is either somewhat increased or stays the same, with the same questions reused, for courses whose material doesn't vary much over time. Because of this, there are several possible solutions to the same problems, which may serve as inputs for training models that can be optimised during their lives. Here, supervised learning-based automatic grading systems are mostly studied using labelled questions, reference answers, and student replies. As a result, we zero attention on supervised techniques, which are widely used for ASAG, and provide more accurate assessments since they are trained with labelled right responses.

Investigative procedures

The following procedures made up our semi-systematic strategy for the literature review: a) searching scientific databases using keywords; b) concentrating on benchmark data sets and the techniques experimented with on them; and c) doing a search in reverse using the references of relevant articles.

We used the most popular scientific databases—ScienceDirect, Google Scholar, ResearchGate, Semantic Scholar, and arXiv—to find applicable articles. We adjusted our primary search phrases, which included "short answer grading," "digital assessment of students," "automated assessment," and "automatic grading system," based on the findings. After that, we choose the articles that report on the most popular benchmark data sets . The publications that suggest automated grading systems were able to be included in the review, even if they used somewhat different task definitions. Lastly, we collected more articles by perusing the reference lists of the previously chosen ones. Number of citations and publication year, topical agreement and practical applicability, and testing on comparable data sets were the selection criteria.

In order to find the most up-to-date trends and advancements, we paid particular attention to the articles published in the last five years. A rising tide of scholarly interest in ASAG systems and their potential pedagogical applications has been reflected in the recent surge in paper publication counts.

Evaluating Data Sets for Short-Answer Questions

There have been tests of existing approaches on many data sets, including SciEntsBank, Beetle, Delhi, ASAP-SAS, and others. There are a lot of ways in which these datasets vary, including the amount of questions, the kind of questions, the topic, the language, the grading system, and the length of the replies. The findings presented in the primary research articles of the methodologies we examined form the basis of our study. Since their testing often only included portions of the whole data sets, it is not always feasible to draw direct comparisons between their results. The fundamental reason for this is because the public data sets have distinct properties and compositions. We do not include some data sets in our study because they are confidential or because their source is unclear.

This evaluation primarily focuses on the four most popular data sets used for ASAG technique benchmarking: SciEntsBank, Beetle Delhi, and ASAP-SAS. They are freely available to the public and include a variety of response domains, so we can test automated grading systems in all their glory. In addition, they ensure that the current approaches are compared fairly. What follows is a description of the data sets, together with details on the tasks and applications that inspired their creation. We provide a brief overview of the properties and features of the data sets that were taken into consideration in Table 1.

Table 1: Specific details on the SciEntsBank, Beetle, Delhi2012, and ASAP-SAS reference datasets. "Additional information" specifies whether there is any other textual material for the assignment beyond the question itself.

Characteristics	Sets of training questions and answers	% of accurate responses	Domain count	Question count	The average amount of responses to each question	The typical word count for a response	Word count limit for the response	Required minimum word count for a response	Further details	Label scale	Publicly accessible

SciEntsBank	4,969	40.41 %	12	135	37	13	110	1	No	2-way, 3-way, 5-way classification	Yes
Beetle	17,198	42.49 %	2	50	366	10	80	1	No	2-way, 3-way, 5-way classification	Yes
Delhi2012	2,442	44,22 %	1	90	29	18	173	1	No	Score between 0 and 5	Yes
ASAP-SAS	17,207	21.57 %	4	10	1,721	42	325	1	Yes	Score between 0 and 2 or between 0 and 3	Yes

Data sets of SciEntsBank and Beetle

The goal of the SemEval 2013 competition, which includes the SciEntsBank and Beetle data sets, is to find frequent errors such omissions and incorrect or thematically unrelated statements so that individualised correction techniques may be developed. A variety of short response grading methods are intended to be tested in this challenge.

Models may be trained on 2-, 3-, and 5-way task problems using the three sets of labels included in the datasets. Each response in the 3-way task is marked as either right, contradictory, or wrong, but in the 2-way test it's just right or wrong. The Recognising Textual Entailment (RTE) task comprises the 2-way and 3-way problems. Examining the ability to distinguish between non-domain, accurate, incomplete, contradictory, and irrelevant responses is the goal of the 5-way task. Tutoring conversation systems are the target of this effort.

SciEntsBank.Questions from a standardised test administered to students in Delhi's third through sixth grade are included in the SciEntsBank data collection. See Table 1 for specifics; the dataset comprises 5,000 responses to 135 questions across 12 categories. A 2-way, 3-way, or 5-way categorization may be necessary for the necessary grading, depending on the category of the domain.

Beetle.The Beetle data set is purpose-built to evaluate students' interactions with a genuine tutorial conversation system, in contrast to the SciEntsBank data set. High school physics and electrical and electronics basics are covered in the system. The data set was created by revising the conversations and using the relevant responses to queries, excluding the interaction protocol. Questions might be factual or seek an explanation or clarification. On average, there are 366 student responses to each of the 47 questions in the corpus. Unseen questions and answers are the only kind included in the Beetle collection.

University of Delhi data set

A lot of people utilise the data set from the University of Delhi (Delhi2012) to compare and contrast how well different automated grading systems work. The book includes 32 students' responses to 82 questions, for a total of almost 2400 question-answer pairings. The typical length of a response is fifty word tokens. The questions are compiled from two exams that test students' foundational understanding of computer science and two sets of assignments. A score between zero and five is assigned to each response by two evaluators. Numerical grades are assigned. Since there were no hard and fast criteria for grading, the two evaluators' combined score is taken as gospel.

Data set from ASAP-SAS

The ASAP-SAS data set was published in 2023 as a result of a Kaggle competition. It stands for Automated Student Assessment Prize Short Answer Scoring. The test includes 10 questions covering a variety of subjects, such as science, English, biology, and English language arts. Both the training data set and the test set include a total of 17,207 and 5,224 responses, respectively, with an average of 1,700 answers per question. The typical

length of a response is 50 words, however a tiny percentage of answers (less than 5%) also exceed 100 words. A score between zero and two or three is assigned to each question.

Taxonomy

First, there are the classic ASAG methods that used hand-crafted features and classical machine learning (CML) techniques like logistic regression and support vector machines; second, there are the deep learning (DL) methods that use feature design as a learning problem in conjunction with predictive model training. Word embedding, and sequential models, are the two subcategories of the second category of approaches; these subcategories reflect the stages of development of natural language processing methods.

Word2Vec and other word embedding methods seek to represent words with comparable semantics using nearby vectors in a latent space that has been learnt. The data included in massive text databases may be adequately described by these machine-learned representations. The second set of techniques, which includes RNN and LSTM based systems, took into consideration longer-range linkages in word groups and larger sentences while developing their algorithms. To better depict sentences and paragraphs, these models take into account the longer-distance relationships between words in a phrase.

7. Results and Discussion

- **Hands-On Design And Automated Learning**

A combination of a standard machine learning classifier, such Logistic Regressor, Support Vector Machine, Random Forest, or Naïve Bayes, and feature vectors taken from the raw text was used in the majority of the studies that were examined. Ensemble techniques, which combine the predictions of several classifiers, formed the basis of some of the examined systems. Here we take a look at various techniques, organise them into groups based on the characteristics used to describe the text (lexical, syntactic, and semantic elements), and provide further details on how they work in current methods. Keep in mind that very few methods really used feature sets that included a mix of various kinds of characteristics. This complicates matters when trying to determine which feature or set of characteristics is inherently better. The findings shown in Table 3 on their performance on benchmark data sets.

Lexical characteristics

An essential part of the first ASAG systems were algorithms for calculating word overlap. By estimating the degree to which two or more sample sentences overlap either in terms of words or characters, overlap-based features are able to quantify the number of words that occur in both the student's and the reference's answers. These approaches were often used in conjunction with other pre-processing techniques, such stemming or lemmatization, to provide even better results. The authors examined the effects of three different word overlap calculation techniques on ASAG system performance and compared them in. The methods were the dice coefficient, the jaccard coefficient, and the cosine coefficient. Using these techniques, we may determine how similar two phrases are by counting the number of words that overlap in them. According to the authors, the cosine coefficient helped get the greatest results when estimating sentence similarity. In order to further enhance the performance of an ASAG system, a weighted cosine coefficient was used. Different methods, such as cosine and Lesk similarity, determined the raw amount of overlapping words and computed various string similarity scores. The authors also used lexical characteristics to calculate sub-tree matching, which included tallying the number of overlapping words and word stems. Features based on word overlap between the reference response and the student's answer were used by the writers in.

Characteristics of syntax

In order to quantify crucial information about a sentence's meaning, syntactic characteristics identify and characterise the functions and interdependencies of the words inside it. To deduce the meaning of a textual response, one must be able to characterise the link between words. Parse trees and dependency n-grams or part-of-speech tagging (POS tags) are basic ways to extract syntactic information from text. By classifying words according to their syntactic relationship, such as verb and subject, dependency n-grams are formed. In order to calculate syntactic characteristics, n-grams consisting of combinations of POS tags were generated in. After that, we compared each response to a reference answer and graded it based on the content using the reliance of words in sequences.

Semanticfeatures

Sentence semantics are not captured by lexical features, while syntactic features capture them to a lesser degree. Consequently, in order to more effectively compute similarity across sentences, more advanced characteristics were developed by using knowledge-bases to identify the meaning of words. Various similarity measures were utilised in conjunction with computational methodologies based on Latent Semantic Analysis (LSA) and Explicit Semantic Analysis (ESA) , as well as knowledge sources such as WordNet . To represent the semantic link between words that induce hyponyms and synonyms, WordNet uses word models. To improve the incorporation of words' semantic

meaning, several strategies used Wordnet in conjunction with similarity measures. As a corpus-based similarity approach, LSA uses a multi-dimensional semantic space to represent words as vectors. After showing greater performance than word and n-gram vectors, this approach became widespread.In order to improve the relevance of individual words, LSA was used to measure their similarity and then coupled with a word-weighting factor in. It was shown that ESA, which was built to employ Wikipedia knowledge extraction, performed as well as LSA, or even better, in certain instances. Using WordNet-based preset context vectors, the authors of included semantic similarity.

By averaging or adding up the embedding for individual words. Word and sentence embedding outperform earlier features that were hand-engineered when it came to capturing semantic information in textual data.In the second group of approaches, recurrent neural networks (RNNs) are used to represent the sequential properties of textual data; LSTM-based RNNs are among the most prominent.

Table 2: Unless otherwise noted, we guarantee accuracy in our reports. The variables Fˆ, FM, and Fm represent different averages of the F1 score. QW-K is a measure of quadratic weighted kappa. RMSE is the root mean square error. ρ is the Pearson's correlation coefficient.

Ref.	Year	Classifier	SciEntsBank			Beetle			Delhi2012	Other
			2-way	3-way	5-way	2-way	3-way	5-way		
[6]	2012	SVM	-	-	-	-	-	-	0.518 (p) 0.998 (RMSE)	-
[7]	2012	kNN	-	-	-	-	-	-	-	0.79 Eng-lish Dev. Corpus
[8]	2012	Decision Tree	-	-	0.29 (Fx) 0.42 (Fm)	-	-	-	-	-
[9]	2013	SVM, Naive Bayes	0.612	0.55	0.421	0.648	0.523	0.464	-	-
[10]	2013	Naive Bayes	0.696 (F)	0.606 (F)	0.464 (F)	-	-	-	-	-
[11]	2013	Logistic Regression	-	-	0.524 (F)	-	-	0.659 (F)	-	-
[12]	2013	SVM, Logistic Regression	0.684	0.612	0.486	0.77	0.624	0.588	-	-
[13]	2013	Bagged Decision Tree	0.726	0.649	0.527	0.724	0.538	0.513	-	-
[14]	2015	Random Forest Regressors	-	-	-	-	-	-	0.61 (p) 0.86 (RMSE)	0.78 (QW-K) ASAP-SAS
[15]	2016	SVM	0.605 (F)	-	0.48 (F)	-	-	-	-	-
[16]	2017	Random Forest	-	-	0.56 (F)	-	-	-	0.85 (p) 0.63 (RMSE)	-
[17]	2018	Logistic Regression	-	-	0.565 (F)	-	-	-	0.82 (RMSE)	-
[18]	2019	Random Forest, Extreme Gradient Boosting	0.775	0.719	0.59	0.81	0.643	0.644	-	-

[19]	2020	Random Forest	-	-	-	-	-	-	-	0.791 (QW-K) ASAP-SAS

By accounting for word sentences of varying durations and longer-range links between words in sentences, these approaches may capture the text's semantic features. Because of this, the prediction models were able to draw stronger and more accurate conclusions from the responses that were provided.

Wordembeddings

ASAG deep learning approaches using word embeddings essentially map words with comparable semantic meanings to nearby locations in a latent space. Word embeddings have been so effective because they capture all the text's rich semantic properties. While these techniques performed a better job of evaluating word similarity, they weren't noticeably better than prior approaches when it came to representing whole sentences in ASAG systems. For example, pre-trained embedding models like Word2Vec andGloVe were used to produce word vector representations of text, which the authors examined using various similarity metrics in. When tested on the SemEval 5-way challenge, they discovered that embeddings of words and sentences performed below par. On ASAG challenges, models with hand-engineered features performed better than those using purely embedding-based techniques.

Table 3:which are based on deep learning. Unless otherwise noted, we guarantee accuracy in our reports. The following variables are defined: Fˆ as the F1 score, QW-K as the quadratic weighted kappa measure, C-K as the Cohen's Kappa, RMSE as the root mean square error, and ρ as the Pearson's correlation coefficient.

Ref.	Cat.	Year	SciEntsBank			Beetle			Delhi2012	Other
			2-way	3-way	5-way	2-way	3-way	5-way		
[20]	DL3	2017	0.712		0.533	0.79		0.633	0.818 (p) 0.993 (RMSE)	0.721 (QW-K) ASAP-SAS
[21]	DL2	2017		0.634 (MAE) 0.904 (RMSE)	0.34 (p)			0.61 (p) 0.77 (RMSE)		
[22]	DL1	2018	0.752	0.654	0.540	-			0.57 (p) 0.902 (RMSE)	
[23]	DL1	2019	-	-	-	-	-	-	-	0.791 (QW-K) ASAP-SAS
[24]	DL3	2019	-	-	-	-	-	-	-	0.77 (QW-K) ASAP-SAS
[25]	DL2	2020	0.803 (F)	0.744 (F)	0.656 (F)	-	-	-		0.724 (F) Large Scale Industry Dataset
[26]	DL1	2020	-		0.503 (F)	-	-	-	0.63 (p) 0.91 (RMSE)	
[27]	DL3	2020	-	0.68 (F)	-	-	-	-	-	-
[28]	DL3	2021	-	-	-	-	-	-	-	0.969 (Acc) 0.999 (F) Chinese data

[29]	DL3	2022	-	-	-	-	-	-	-	0.889 (Acc) 0.943 (AUC) Real world K-12
[30]	DL2	20223	-	-	-	-	-	-	0.850 (Acc) 0.830 (F) Core fitters	

When word-embeddings and hand-engineered features are used together, ASAG systems function well, but when word-embeddings are used alone, they don't always provide excellent results. Building a big feature set using a combination of learnt and hand-engineered features, for example, yielded excellent performance results. Features that capture the variety and style of formulation of the student replies are combined with regularly used text representations such as Word2Vec, Doc2Vec, POS tagging, and n-gram overlaps in the authors' suggested technique.

Sequence-based models

To strengthen the word and sentence representation and enhance the quality of the learnt features, ASAG was trained using sequential machine learning models. The writers checked the student's response against the reference answer by comparing the pairwise distance of their latent vectors. They demonstrated that automated grading systems and learnt representation quality are both enhanced by extracting semantic textual aspects using sequence-based models.

The development of sequential NLP models and how to apply them to ASAG issues received more attention. In particular, we investigated the possibility of using transfer learning approaches to fine-tune pre-trained models for analysis of text sequences from other, more broad domains to the ASAG problem. With the help of transfer learning, it is possible to take use of characteristics with superior semantic representation capabilities that have been learnt from massive text data corpora." Word representations were often extracted in transfer learning situations using the Universal Sentence Representation model in this context.

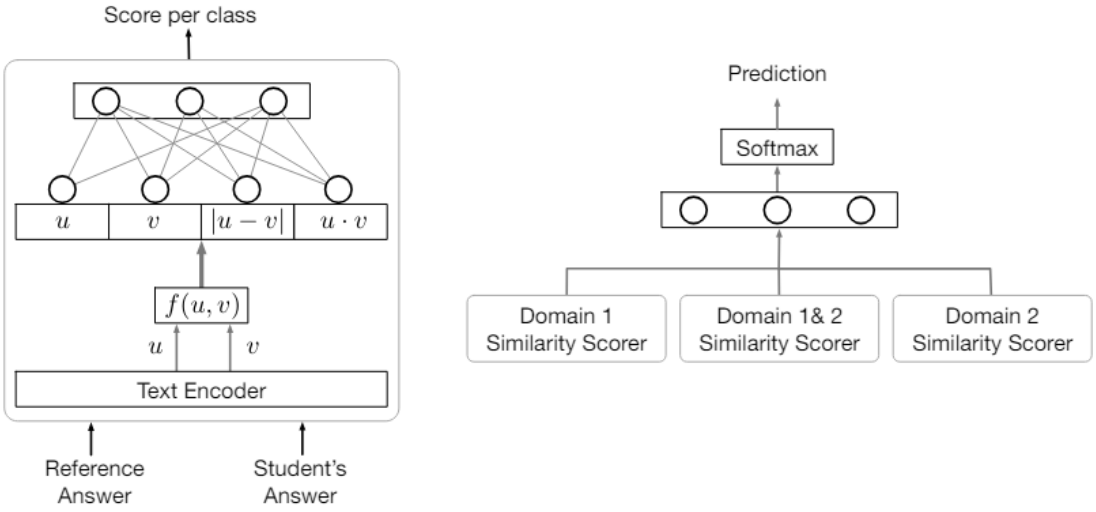

Fig. 1. The disclosed method's architecture. Different similarity scores were trained on domain-specific subsets of responses and a domain-independent data set as part of the authors' domain adaption.

8. Conclusion

We took a look back at the latest developments in automated short answer grading (ASAG) and gave a rundown of what has been accomplished utilising deep learning techniques. By expanding upon earlier literature reviews, we were able to determine which architectural decisions and critical aspects affected the performance of ASAG systems throughout the Deep Learning period. We connected the outcomes that new approaches got on benchmark data sets to the methodological advancements.

From traditional Machine Learning techniques to more recent Deep Learning approaches, this study covers it all, including current and future research trends and a taxonomy of methodologies. Adapted to ASAG tasks using transfer learning and domain adaption approaches, Deep Learning architectures for natural language processing are insufficient to handle the demands and obstacles of this domain. It is challenging for deep learning methods to reliably grasp the meaning of brief responses in order to compare them with reference replies. This was addressed by using hybrid models that integrate deep representation learning with feature engineering, as well as by using an ensemble of classifiers and stacked models. When combined with pre-existing lexical, syntactic, and semantic characteristics, the attention-based analysis of Transformers and the embedding capabilities of Deep Learning models may significantly improve ASAG system performance.

8.1 Findings of the study

The survey evaluated the status of Automated Short Answer Grading (ASAG) using deep learning. It pinpointed the architectural decisions that affect ASAG performance and linked the methodological improvements to the benchmark results. Comprehending nuances of the subject matter remains a challenge for a pure deep learning based grading. In order to substitute these approaches, ensemble/stacked/hybrid models combining engineered features and deep representation learning were developed. Transformer embeddings and attention complete the lexical/syntactic/semantic features, which, in turn, improve the performance. Nonetheless, a thorough benchmark dataset that evaluates the methods equally and fosters progress is the need of the hour.

8.2 Scope for further research

It has been shown that deep learning techniques may augment the text representation capabilities of methods that rely on hand-engineered features, and that these techniques have helped to increase the ASAG system's performance. It is essential to take action to resolve the aforementioned issues if we want to broaden the use of deep learning techniques for short answer grading and increase their potential. There is a lack of diversity in the questions and brief reference responses across areas in the existing benchmark data sets, which makes them insufficient. Because of this, learning-based approaches are more likely to be overfit, which limits their ability to generalise. Because of this, activities such as expanding current data sets, augmenting data, and creating synthetic data using generative models are very pertinent to advancing the discipline in the future. Important areas that require more study are the explainability and resilience of models based on deep learning.

References

[1] Michael Mohler, Razvan C. Bunescu, and RadaMihalcea, 'Learning to grade short answer questions using semantic similarity measures and dependency graph alignments.', in ACL, pp. 752–762, (2021).

[2] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, 'Bleu: a method for automatic evaluation of machine translation', Technical report, IBM Research Report, (2021).

[3] DaiX. et al.(2020) An analysis of simple data augmentation for named entity recognition.

[4] BayerM. et al. (2022) A survey on data augmentation for text classificationACMComput. Surv.

[5] BurrowsS. et al. (2022) The eras and trends of automatic short answer grading Int. J. Artif. Intell. Educ.

[6] MichaelMohler,RazvanBunescu,andRadaMihalcea. 2012.LearningtoGradeShortAnswerQuestionsusing SemanticSimilarityMeasuresandDependencyGraph hAlignments.In*Proceedingsofthe49thAnnualMeetingoftheAssociationforComputationalLinguistics:HumanLanguageTechnologies*.AssociationforComputationalLinguistics,Portland,Oregon, USA, 752–762.

[7] DetmarMeurers,RamonZiai,NielsOtt,andStaceyM Bailey.2012.Integratingparallelanalysismodulesto evaluatethe meaning of answers to reading comprehension questions.*International journal of continuing engineering educationandlife-longlearning*21(2012),355–369.

[8] Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012.Towards Effective Tutorial Feedback for ExplanationQuestions:ADatasetandBaselines.In*Proceedingsofthe2012ConferenceoftheNorthAmericanChapteroftheAssociationforComputationalLinguistics:HumanLanguageTechnologies*.AssociationforComputationalLinguistics,Montréal, Canada,200–210.

[9] MilenKouylekov, Luca Dini, AlessioBosca, and Marco Trevisan. 2013.Celi: EDITS and Generic Text Pair Classification.In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the SeventhInternationalWorkshoponSemanticEvaluation(SemEval2013)*.AssociationforComputationalLinguistics,Atlanta,Georgia,USA,592–597.

[10] OmerLevy,TorstenZesch,IdoDagan,andIrynaGurevych.2013.In*SecondJointConferenceonLexicalandComputa-tional Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, 285–289.

[11] MichaelHeilmanandNitinMadnani.2013.ETS:DomainAdaptationandStackingforShortAnswerScoring.In*SecondJoint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh InternationalWorkshoponSemanticEvaluation(SemEval2013)*.AssociationforComputationalLinguistics,Atlanta,Georgia,USA,275–279.

[12] NielsOtt, Ramon Ziai, Michael Hahn, and DetmarMeurers. 2013.CoMeT: Integrating different levels of linguisticmodelingformeaningassessment.In*SecondJointConferenceonLexicalandComputationalSemantics(*SEM),Volume2:ProceedingsoftheSeventhInternationalWorkshoponSemanticEvaluation(SemEval2013)*.AssociationforComputationalLinguistics, Atlanta, Georgia, USA, 608–616.

[13] SergioJimenez,ClaudiaBecerra,andAlexanderGelbukh.2013.SOFTCARDINALITY:HierarchicalTextOverlapforStudentResponseAnalysis.In*SecondJointConferenceonLexicalandComputationalSemantics(*SEM),Volume2:ProceedingsoftheSeventhInternationalWorkshoponSemanticEvaluation(SemEval2013)*.AssociationforComputationalLinguistics, Atlanta, Georgia, USA, 280–284.

[14] LakshmiRamachandran,JianCheng,andPeterFoltz.2015.IdentifyingPatternsForShortAnswerScoringUsing Graph-based Lexico-Semantic Text Matching. In *Proceedings of the Tenth Workshop on Innovative Use of NLP forBuildingEducationalApplications*.AssociationforComputationalLinguistics,Denver,Colorado,97–106.

[15] Ahmed EzzatMagooda, Mohamed A. Zahran, Mohsen A. Rashwan, Hazem M. Raafat, and Magda B. Fayek. 2016.Vector Based Techniques for Short Answer Grading. In *Proceedings of the Twenty-Ninth International Florida ArtificialIntelligence Research Society Conference, FLAIRS 2016, Key Largo, Florida, USA, May 16-18, 2016*, Zdravko Markov andIngrid Russell (Eds.). AAAI Press, 238–243

[16] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2017.Fast and Easy Short Answer Grading with HighAccuracy.In*Proceedingsofthe2016ConferenceoftheNorthAmericanChapteroftheAssociationforComputationalLinguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 1070–1075.

[17] ShouryaRoy,HimanshuS.Bhatt,andY.Narahari.2018.AnIterativeTransferLearningBasedEnsembleTechniqueforAutomatic Short Answer Grading.*ArXiv*abs/1609.04909 (2018).

[18] Lucas B. Galhardi, Helen Senefonte, Rodrigo de Souza, and Jacques Brancher. 2019.Exploring Distinct Features forAutomaticShortAnswerGrading.In*AnaisdoXVEncontroNacionaldeInteligênciaArtificialeComputacional*.SBC,Porto Alegre, RS, Brasil, 1–12.

[19] YamanKumar,SwatiAggarwal,DebanjanMahata,RajivRatnShah,PonnurangamKumaraguru,andRogerZimmer-mann. 2020.Get IT Scored Using AutoSAS - An Automated System for Scoring Short Answers. In *AAAI*.

[20] BrianRiordan,AndreaHorbach,AoifeCahill,TorstenZesch,andChongMinLee.2017.Investigatingneural architectures for short answer scoring. In *BEA@EMNLP*.

[21] Sachin Kumar, SoumenChakrabarti, and Shourya Roy. 2017.Earth Mover's Distance Pooling over Siamese LSTMsforAutomaticShortAnswerGrading.In*ProceedingsoftheTwenty-SixthInternationalJointConferenceonArtificialIntelligence,IJCAI-17*.2046–2052.

[22] SwarnadeepSaha,TejasI.Dhamecha,SmitMarvaniya,RenukaSindhgatta,andBikramSengupta.2018.SentenceLevelorTokenLevelFeaturesforAutomaticShortAnswerGrading?:UseBoth.In*ArtificialIntelligenceinEducation*,CarolynPensteinRosé,RobertoMartínez-Maldonado,H.UlrichHoppe,RoseLuckin,ManolisMavrikis,KaskaPorayska-Pomsta,Bruce McLaren, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 503–517

[23] Tianqi Wang, Naoya Inoue,Hiroki Ouchi, TomoyaMizumoto,and KentaroInui. 2019.InjectRubrics into ShortAnswerGradingSystem.In*Proceedingsofthe2ndWorkshoponDeepLearningApproachesforLow-ResourceNLP(DeepLo2019)*.AssociationforComputationalLinguistics,HongKong,China,175–182.

[24] WaelHassanGomaaandAlyAlyFahmy.2019.Ans2vec:AScoringSystemforShortAnswers.In*TheInternationalConferenceonAdvancedMachineLearningTe*

chnologiesandApplications(AMLTA2019),AboulEllaHassanien,Ah-mad TaherAzar, TarekGaber, RoheetBhatnagar, and Mohamed F. Tolba (Eds.). Springer International Publishing,Cham,586–595.

[25] ChulSung,TejasIndulalDhamecha,andNirmalMukhi.2020.ImprovingShortAnswerGradingUsingTransformer-BasedPre-training.In*ArtificialIntelligenceinEducation*,SeijiIsotani,EvaMillán,AmyOgan,PeterHastings,BruceMcLaren, and Rose Luckin (Eds.). Springer International Publishing, Cham, 469–481.

[26] HuiQi,YueWang,JinyuDai,JinqingLi,andXiaoqiangDi.2020.Attention-BasedHybridModelforAutomaticShortAnswer Scoring. In *Simulation Tools and Techniques*, Houbing Song and Dingde Jiang (Eds.). Springer InternationalPublishing, Cham, 385–394.

[27] Tianqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019.Automatic ShortAnswer Grading via Multiway Attention Networks.*ArXiv*abs/1909.10166 (2020).

[28] ChulSung,Tejas I. Dhamecha, SwarnadeepSaha, Tengfei Ma, V. Pulla Reddy, and Rishi Arora. 2021.Pre-TrainingBERT on Domain Resources for Short Answer Grading. In *EMNLP/IJCNLP*

[29] LishanZhang,YuweiHuang,XiYang,ShengquanYu,andFuzhenZhuang.2022.Anautomaticshort-answergradingmodelforsemi-open-endedquestions.*InteractiveLearningEnvironments* 0,0(2019),1–14.

[30] YuanZhang,ChenLin,andMinChi.2023.Goingdeeper:Automaticshort-answergradingbycombiningstudentandquestionmodels.*UserModelingandUser-AdaptedInteraction*30,1(01Mar2020),51–80.https://doi.org/10.1007/s11257-019-09251-6.