Automatic Short Answer Grading: A Survey

Dishank Aggarwal, Pushpak Bhattacharyya, Bhaskaran Raman
Department of Computer Science and Engineering, IIT Bombay, India
{dishankaggarwal, pb, br}@cse.iitb.ac.in

Abstract

As education systems increasingly transition to digital platforms, the demand for efficient and accurate assessment methods has grown significantly. This paper provides a comprehensive overview of Automatic Short Answer Grading (ASAG), a technology designed to automatically evaluate student responses. We explore two main types of ASAG systems: traditional ASAG, which focuses solely on grading, and ASAG with feedback, which also provides students with explanations and constructive feedback. The paper reviews the various datasets utilized in ASAG research, including traditional and feedback-based datasets, and examines the range of methods employed, from basic machine learning algorithms to advanced deep learning techniques. We also discuss key challenges such as the need for multimodality, multilingual support, and personalization, and present future directions to address these challenges. Our goal is to present the progress, challenges, and future possibilities in ASAG, making it easier for researchers and educators to understand and leverage this technology in enhancing educational outcomes.

1 Introduction

Technology integration in education has resulted in transformative changes, redefining traditional pedagogical approaches and assessment methodologies. Effective education relies on both feedback and explanations provided during assessments to ensure quality learning outcomes (Shute, 2008). Grading questions in tests and examinations have proven to be a good measure to assess student learning and understanding of a topic or a subject. An exam could include various question types, such as multiple choice, fill-in-the-blank, short answers, essays, etc. Among these question types, short answers and essays are more complicated to analyze than multiple-choice or fill-in-the-blank type questions due to flexibility and natural language in the response. Automating the grading process becomes crucial, especially in countries with extremely high student-to-teacher ratios, as it can significantly reduce teachers' workloads and improve the assessment process. Significant advancements have been achieved in this field in recent years, primarily attributed to the SemEval Semantic Textual Similarity (STS) task (Agirre et al., 2012, 2013, 2014, 2015).

This challenge can be approached as a machine learning issue, where the objective is to calculate the grade of a student's response based on how similar it is to the reference answers. However, simply assigning a score or label to a learner's response is often inadequate in practical educational contexts. Nevertheless, the grading process's accuracy depends on various factors, like the features used to depict the student's answer, the similarity metric used, and the quantity and quality of the data used to train the grading model.

1.1 Problem Statement

The problem statement for the task of ASAG is further divided into two subcategories

1. Traditional Automatic Short Answer Grading

2. Automatic Short Answer Grading with Feedback

1.1.1 Traditional Automatic Short Answer Grading

Given a question, a reference answer, and a student's answer, the goal is to assign a label indicating the degree of correctness in the student's answer compared to the reference answer. This task involves evaluating the alignment between the student's and reference answers and assigning an output label or grade for that particular answer. Figure 1 illustrates the problem statement, showing an input and output sample.

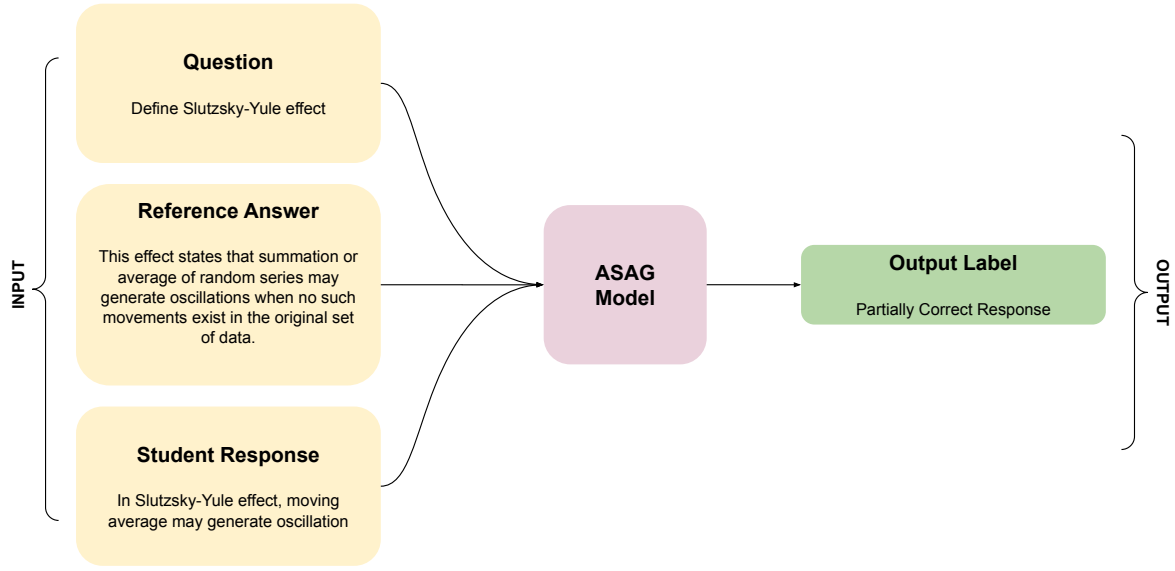

Figure 1: Traditional ASAG accepts a question, reference answer, and student answer as input and outputs an output label from 'correct,' 'partially correct,' or 'incorrect.'

1.2 Automatic Short Answer Grading with Feedback

Given a question, a reference answer, and a student's answer, the aim is to provide content-focused elaborated feedback and assign a label indicating the degree of correctness in the student's answer compared to the reference answer. Here, we focus on questions where the answer type is a sentence or a short paragraph. This task involves evaluating the alignment between the student's and reference answers. Figure 2 illustrates the problem statement, showing an input and output sample.

1.3 Motivation

The increasing demand for technology in education has led to a need for more efficient and effective methods of grading short-answer assessments. Due to very high student-to-teacher ratio in the educational field, traditional manual grading methods frequently consume a significant amount of educator's time, limiting their ability to provide students with timely and constructive feedback. Manual grading of short answers is also prone to human error and lacks consistency. To support this argument, reference to **2011** data reveals that in the realm of secondary education, **India** exhibits a students-per-teacher ratio of **25.92**. In contrast, for countries like **Central African Republic (CAR)** and **Croatia**, the corresponding student-per-teacher ratios are **66.82** and **8**, respectively. In this section, data values have been taken from the nationmaster website [1].

As a result, the task of manually grading student answers places a significant burden on the teacher or instructor. Over that, real-world tests and evaluations include questions from multiple domains. Grading models trained on such broad datasets can better simulate student response complexity and variety, yielding more accurate and relevant assessment outcomes. Given the varying student-to-teacher ratios within India and worldwide, which can be seen in Figure 3, a critical need arises for innovative solutions to manage assessment and feedback processes effectively. Here's how the introduction of a new ASAG dataset and, hence, improving the ASAG system can address these challenges:

- **Scalability**: Automating the grading process becomes crucial in countries where student-to-teacher ratio is extremely high. An ASAG system can efficiently handle a large volume of student responses.
- **Timely Feedback**: With larger class sizes, providing timely and constructive feedback to each student becomes challenging. An automated grading system can offer immediate feedback.
- **Standardization**: Automated grading ensures

[1] https://www.nationmaster.com/country-info/stats/Education/Pupil--teacher-ratio,-secondary#2011

Question	What is the difference between basin order and channel order?
Reference Answer	Basin order is highest order of any stream in that basin whereas Channel order is order of stream which denotes that in what order of streams has joined the channel."
Student Answer 1	Highest order channel is the basin order whereas channel order is the order of channel from tributaries to reaches to main river stream.
Label	**2 (Correct response)**
Feedback	Excellent! You have a clear understanding of the distinction between basin order and channel order.
Student Answer 2	Channel order reflects to the number of streams coming together to form a channel.
Label	**1 (Partially Correct response)**
Feedback	Your answer includes a part of the distinction. Channel order indeed indicates the number of streams joining together to make a channel, but the difference between basin order and channel order is not mentioned.
Student Answer 3	Channel order is the Order of the highest order streams. For example, two first order streams (or more) will make a second order stream and similarly for highest orders.
Label	**0 (Incorrect response)**
Feedback	The student answer confuses basin order with channel order. Basin order refers to the highest order of streams within a basin, while channel order refers to the order of streams based on the sequence of junctions.

Table 1: An example showing a question, reference answer, and three student answers (Student#1, Student#2, and Student#3) alongside their corresponding labels and synthetically generated Feedback/Explanation for the assigned label from the EngSAF Dataset

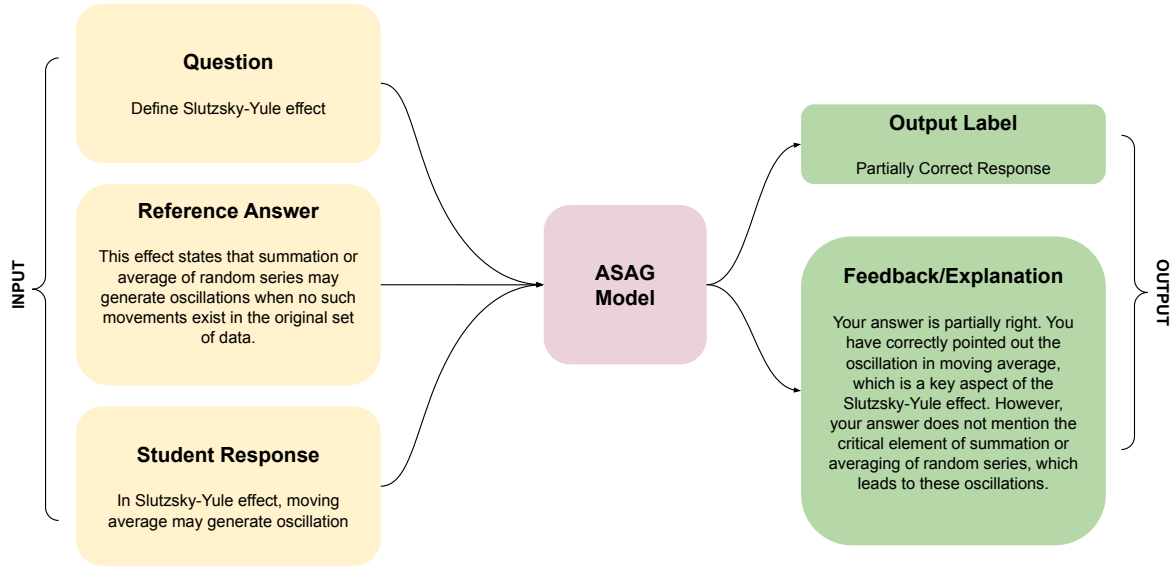

Figure 2: ASAG with feedback model which accepts a question, reference answer, and student answer as input and outputs an output label from 'correct,' 'partially correct,' or 'incorrect' along with the feedback/explanation of the assigned grade.

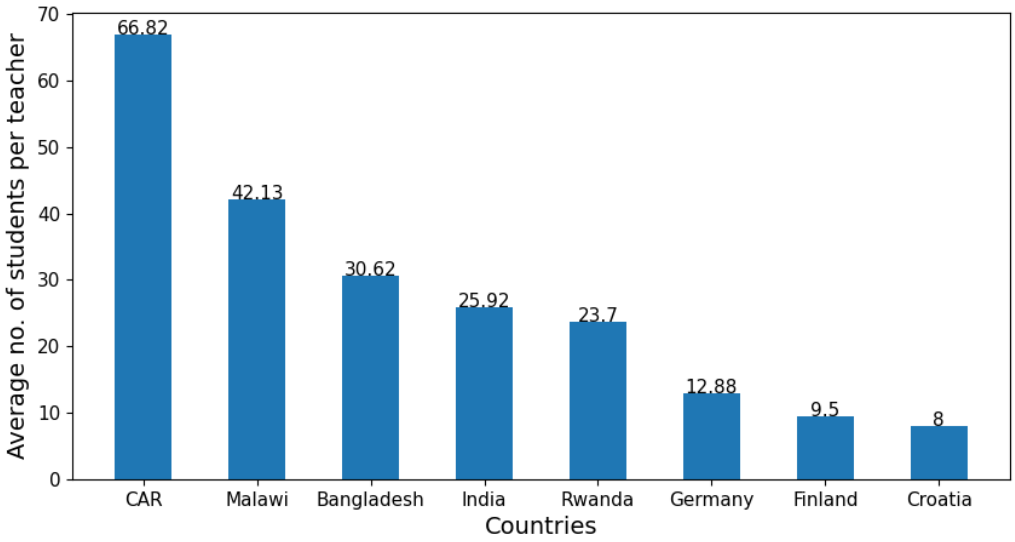

Figure 3: 2011 Student-teacher ratio for secondary education among different countries.

a consistent and objective evaluation process. This is especially important in cases where differences in educator expertise or subjectivity may result in uneven grading standards.

2 Background

2.1 Definitions and Terminologies

- **Question** - The question refers to the string of text that represents the query being asked in the examination.
- **Context/Support Document/Document** - Support Document (also referred to as a document) denotes the corpus the user wants the system to consult as the source of the question being asked.
- **Correct answer** - It refers to the string of text that denotes the gold/correct answer to the query/question being asked. It could be provided as input by the instructor or derived using the question-answering model given the question and the context.

- **Student response**: It refers to the string of text provided by the student against a question that needs to be graded.
- **Feedback/ Explanation**: This feedback typically includes a judgment on the correctness of the response, specific guidance on areas of improvement, and explanations of errors or misconceptions. This feedback aims to enhance student learning by offering constructive and informative insights that help students understand their mistakes and learn from them.
- **Automatic Short Answer Grading (ASAG)** - ASAG is a task to assign grades to the student's answer against the reference answer based on the degree of correctness of the response.
- **Extractive Summarization** - This form of summarization produces a summary using segments of text from the original support document(s).
- **Abstractive Summarization** - This form of summarization produces the summary by generating text that is relevant to the support document.
- **Language Modeling**: A variety of statistical and probabilistic techniques are used in language modeling (LM) to estimate the likelihood that a given string of words will appear in a sentence. Language models examine corpora of text data to establish a foundation for their word predictions.
- **Cosine Similarity**: Cosine similarity is a metric that can assess how similar two vectors are. Its value lies between 0 (least similar) and 1 (most similar).
- **Large Language Models (LLMs)**: Designed to comprehend, create, and manipulate human language at a highly proficient level of knowledge, large language models (LLMs) are a type of advanced artificial intelligence system. These models use vast quantities of text data and sophisticated machine learning methods to perform various natural language processing (NLP) tasks, which include text completion, translation, summarization, and question-answering.

Key functionalities of LLMs include:

1. **Text Generation**: Creating new text based on a given prompt. This includes completing sentences, writing essays, and generating dialogue.
2. **Text Summarization**: Condensing long pieces of text into concise summaries while retaining the main ideas and important information.
3. **Translation**: Converting text from one language to another.
4. **Question Answering**: Providing precise answers to questions posed in natural language by understanding and retrieving relevant information.
5. **Sentiment Analysis**: Analyzing and determining the sentiment or emotional tone of a piece of text.
6. **Conversational Agents**: Powering chatbots and virtual assistants that can engage in natural, context-aware conversations with users.

2.2 Natural Language Inference

Natural Language Inference (NLI) is a fundamental task in natural language processing (NLP) that involves determining the relationship between a pair of sentences. Specifically, given two sentences—typically referred to as the "premise" and the "hypothesis"—the goal is to classify their relationship into one of three categories: entailment, contradiction, or neutral.

- **Entailment**: The hypothesis logically follows from the premise or can be entailed from the premise.
- **Contradiction**: The hypothesis directly contradicts the premise.
- **Neutral**: The Hypothesis can not be determined or inferred from the premise.

2.3 Prompting Large Language Models

We are leveraging Large Language Models (LLMs) and prompting techniques for the task of Automatic Short Answer Grading (ASAG). By employing LLMs, we can prompt the model with specific questions and reference answers to evaluate the accuracy and relevance of a student's response.

Prompting is a technique used to interact with large language models (LLMs) by providing specific input text, or “prompts," to guide their output generation. This method leverages the model's pretrained knowledge to perform various tasks based on the context and structure of the prompt provided.
Types of Prompts:

- **Instruction Prompts**: These prompts explicitly instruct the model to perform a task. For example, "Translate the following English sentence to Japanese: 'Hello, how are you?'"
- **Completion Prompts**: These prompts provide an initial text that the model completes. For instance, starting a story with "Once upon a time, in the Roman Empire, there was a king..."
- **Question Prompts**: These prompts ask questions that the model is expected to answer. For example, "Who is the president of India?"
- **Contextual Prompts**: These prompts include contextual information that the model uses to generate a relevant response. For example, providing background on a topic before asking for a summary or opinion.

3 Datasets for ASAG

ASAG is an essential area of research that has garnered significant attention in recent years. Several publicly accessible datasets have been curated, each designed to facilitate research and benchmarking in the task of ASAG. Despite many advancements that have been made in the field of NLP, short answer grading has received an insignificant amount of attention. This is primarily due to the unavailability of good quality and freely accessible datasets for this field. The following corpora were utilized extensively throughout the work on ASAG.

This chapter presents and analyzes the datasets relevant to our task. This chapter is divided into two sections: the first section covers datasets used for the traditional ASAG task, while the second section focuses on datasets specifically designed for incorporating feedback into ASAG.

3.1 Traditional ASAG datsets

The datasets in this section are used for traditional Automated Short Answer Grading (ASAG), which involves assigning a label or grade to a student's answer based on its correctness relative to a reference answer. Figure 1 illustrates the problem statement, showing an input and output sample.

3.1.1 SCIENTSBANK dataset

This dataset is a filtered subset of the SCIENTSBANK Extra corpus (Nielsen et al., 2008), used for traditional Automated Short Answer Grading (ASAG). Problematic questions and certain types of student answer facets were removed to simplify the dataset. Specifically, only facets labeled as 'Expressed' or 'Unaddressed' were retained, and complex inter-propositional and relational facets were excluded. The dataset is split into training and test sets, with the training set containing 13,145 reference answer facets (5,939 'Expressed' and 7,206 'Unaddressed') and the test set containing 16,263 facets (5,945 'Expressed' and 10,318 'Unaddressed'). The test set is further divided into subsets for unseen answers, questions, and domains, aligning with the splits used in the main task.

Given a question, a known correct 'reference answer,' and a 1- or 2-sentence 'student answer,' each student answer in the corpus is labeled with one of the following judgments:

- **Correct**: The student's answer is a complete and correct paraphrase of the reference answer.
- **Partially Correct Incomplete**: The student's answer contains some but not all information from the reference answer.
- **Contradictory**: The student's answer contradicts the reference answer.
- **Irrelevant**: The student's answer discusses domain content but does not provide the necessary information.
- **Non-Domain:** The student answer does not include domain content, e.g., "I don't know,"

3.1.2 Basic Electricity and Electronics Tutorial Learning Environment (BEETLE)

The BEETLE corpus (**?**) comprises 56 basic electricity and electronics questions, each requiring 1- or 2-sentence answers and approximately 3,000 student responses to these questions. Each student's response is annotated with labels indicating the correctness of the answer. These labels include categories as:

- **Correct**: The student's answer is a complete and correct paraphrase of the reference answer.
- **Partially Correct Incomplete**: The student's answer contains some but not all information from the reference answer.
- **Contradictory**: The student's answer contradicts the reference answer.
- **Irrelevant**: The student's answer discusses domain content but does not provide the necessary information.
- **Non-Domain:** The student answer does not include domain content, e.g., "I don't know,"

The BEETLE dataset is designed to facilitate the development and evaluation of ASAG systems. It provides a benchmark for comparing different ASAG approaches and techniques.

3.1.3 University of North Texas dataset

Mohler et al. (2011) released the dataset containing 80 undergraduate Data structure questions and 2,273 student responses from an exam of the University of North Texas graded by two human judges. These questions are spread across ten assignments and two tests, each on a related set of topics (e.g., programming basics, sorting algorithms). A reference answer is provided for each question. Inter-annotator agreement was 58.6% (Pearson's ρ) and .659 (RMSE on a 5-point scale). The average of the two human scores is used as the final gold score for each student's answer.

The task is to assign a real-valued score between 0 and 5 to a student response against a correct reference response. Zero (0) means the student's answer is completely incorrect, while Five (5) means the student's answer is fully correct.

3.1.4 Other ASAG Datasets

Another dataset that is publicly available on Kaggle and could be used for the task of ASAG is provided by Hewlett Foundation named ASAP-AES[2] (Automated Assessment Prize Competition for Essay Scoring). Researchers face challenges when using ASAP, such as shortlisting samples for ASAG based on response length and managing non-uniformity in grading scales. AR-ASAG by Ouahrani and Bennouar (2020) and Cairo by Gomaa and Fahmy (2014) are Arabic datasets containing 2133 and 610 student responses, respectively. There are 61 questions, each answered by ten students in the Cairo dataset. AR-ASAG is publicly available, whereas Cairo is not. SPRAG is a recently released dataset in 2022 containing around 4k student responses in the Python programming domain (Bonthu et al., 2022).

[2] https://www.kaggle.com/c/asap-aes/data

3.2 Feedback-based ASAG datasets

The datasets in this section are used for Automated Short Answer Grading (ASAG) along with Feedback/Explanation, which involves assigning a label or grade to a student's answer based on its correctness relative to a reference answer and the explanation of the assigned output label. Figure 2 illustrates the problem statement, showing an input and output sample.

3.2.1 Short Answer Feedback (SAF) dataset

Filighera et al. (2022), introduces an inaugural dataset for short-answer feedback comprising bilingual responses in English and German. Unlike traditional datasets that only provide labels or scores for answers, SAF includes elaborate feedback explaining the given score. This dataset enables the training of models that grade answers and explain where and why mistakes were made. The SAF dataset comprises 4,519 submissions to German and English questions, demonstrating high inter-annotator agreements. However, the dataset contains questions from only one domain (physics), which lacks generability on multiple domains.

3.3 Dataset comparison

A fundamental question emerges in the context of ASAG: **W**hat prompts the necessity for a new dataset when publicly available datasets are already present for the ASAG task? Table 2 compares SAFEAA with other existing ASAG datasets. The SAFEAA dataset contains 104 questions from multiple domains compared to other datasets, which contain questions from only a particular domain.

Here are some of the key advantages of datasets that contain questions from multiple domains over the dataset that contain questions from a particular domain:

- **Generalizability**: The developed grading model can generalize its learning across different subjects and topics with a multi-domain dataset.

Dataset	Answers	Language	Domain	Availability	IAA
Texas	630	English	Data Structures	Yes	0.644
Texas Extended	2273	English	Data Structures	Yes	0.79
ASAP	2200	English	Science, Biology	Yes	
AR-ASAG	2133	Arabic	Cybercrimes	Yes	0.838
Cairo	610	Arabic	Environmental Science	No	0.86
Beetle	3000	English	Basic Electronics and Electricity	Yes	
SPRAG	4039	English	Python Programming	Yes	0.779
SAFEAA	3704	English	Multiple Engineering Domains	Yes	

Table 2: A Comparison of the SAFEAA Dataset with other popular Short Answer Grading Datasets

- **Versatility**: Grading models trained on a multi-domain dataset can handle short answers from different academic areas, including science, mathematics, literature, and more. This versatility is particularly important for educational technology applications that aim to cater to a broad spectrum of subjects.
- **Real-World Relevance**: In many real-world educational scenarios, exams, and assessments contain questions from multiple domains. Grading models trained on such diverse datasets can better emulate the complexity and variety of student responses, providing more accurate and relevant assessment results.
- **Cross-Domain Insights**: Working with a multi-domain dataset exposes researchers and developers to insights and challenges that arise when assessing short answers across different subjects.
- **Educational Value**: For educational researchers and practitioners, a multi-domain dataset can offer insights into how students respond to different types of questions across subjects.

4 Automatic Short Answer Grading Methods

ASAG is an essential area of research that has garnered significant attention in recent years. Despite many advancements that have been made in the field of NLP, short answer grading has received an insignificant amount of attention. Several approaches have been proposed for the task of ASAG, ranging from rule-based methods to more sophisticated machine-learning techniques. The survey paper by Bonthu et al. (2021) gives an in-depth view of all the methods mentioned here.

One early approach for ASAG was based on keyword or pattern matching, where the presence or absence of certain keywords in the student's answer was used to determine its accuracy. These methods, however, were limited in their ability to handle synonyms and variations in student responses (Mitchell et al., 2002; Sukkarieh et al., 2004; Nielsen et al., 2009).

To overcome these limitations, researchers have developed more sophisticated methods that use natural language processing (NLP) techniques. One such method is based on Latent Semantic Analysis (LSA), which represents texts as high-dimensional vectors and compares them to the reference answers using cosine similarity (LaVoie et al., 2020). In a related study, the task of ASAG is addressed by incorporating features such as answer length, grammatical correctness, and semantic similarity in comparison to reference answers (Sultan et al., 2016).

The researchers used different methods like Transfer Learning, Siamese LSTM, clustering, Latent Semantic Analysis, Bidirectional Transformers, Paragraph Embeddings, Deep Autoencoders, and Attention Networks, and Transformer-based pretraining in recent years. New progress in deep learning for NLP shows that deep learning tools like the Attention mechanism and Transformer are useful for handling more complex NLP tasks. More recently, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied to the task of ASAG. These models are trained on large amounts of annotated data and can capture the semantic relationships between words in a student's answer and the reference answers (Surya et al., 2019; Zhang et al., 2022). Pre-trained Language Models (PLMs), such as BERT (Devlin et al., 2018), GPT (Radford et al., 2019), RoBERTa (Liu

et al., 2019), DistillBERT (Sanh et al., 2019), and ALBERT (Lan et al., 2019) have performed exceptionally well for various tasks in NLP, also used for the task of ASAG. Sentence-BERT or SBERT (Reimers and Gurevych, 2019), is a cutting-edge approach to natural language processing that focuses on creating meaningful sentence embeddings. An architecture of Siamese networks (Koch et al., 2015) sits at the core of SBERT. SBERT performs exceptionally well on current publicly available ASAG datasets (Condor et al., 2021).

5 Future Directions

The future of Automatic Short Answer Grading (ASAG) offers several promising research and development avenues to enhance its capabilities and applicability. Key directions include:

- **Multimodal Data Integration:** Incorporate text, audio, images, and videos for comprehensive evaluation.
- **Multilingual Support:** Develop ASAG systems to cater to diverse linguistic backgrounds.
- **Personalization and Adaptivity:** Implement tailored feedback and adaptive grading based on individual student profiles.
- **Explainability and Transparency:** Ensure grading processes are clear and understandable.
- **Real-time Feedback:** Provide immediate feedback for interactive learning.
- **Ethical Considerations and Bias Mitigation:** Address biases and ensure fairness and equity.
- **Scalability and Integration:** Focus on scalable solutions that integrate seamlessly with existing educational technologies.
- **Cross-disciplinary Applications and Longitudinal Studies:** Explore diverse applications and assess long-term impacts.

By addressing these areas, ASAG systems can become more advanced, equitable, and effective, better serving the diverse needs of modern education.

6 Summary and Conclusion

In this paper, we provided an extensive overview of Automatic Short Answer Grading (ASAG), exploring its development, current methodologies, and future directions. ASAG systems have evolved from basic machine learning models to sophisticated deep learning techniques capable of grading short answers with increasing accuracy. We discussed traditional ASAG approaches that focus solely on grading, as well as feedback-based ASAG systems that provide valuable insights and explanations to students, enhancing their learning experience.

We also reviewed various datasets used in ASAG research, highlighting their significance and comparing their features. Our examination of different ASAG methods showcased the progress in the field, while also identifying challenges such as the need for multimodality, multilinguality, and personalization. Future directions emphasize the importance of explainability, real-time feedback, robustness to diverse answer styles, and ethical considerations. Addressing these areas will be crucial for the continued advancement and widespread adoption of ASAG technology.

In conclusion, ASAG presents a promising solution for efficient and accurate assessment in digital education environments. By integrating advanced techniques and addressing current limitations, ASAG systems can significantly enhance educational outcomes, making learning more personalized, inclusive, and effective. As research continues to evolve, ASAG will play an increasingly vital role in shaping the future of education.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A

pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Sridevi Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In *Machine Learning and Knowledge Extraction*, pages 61–78, Cham. Springer International Publishing.

Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. 2022. Sprag: Building and benchmarking a short programming related answer grading dataset.

Aubrey Condor, Max Litster, and Zachary Pardos. 2021. Automatic short answer grading with sbert on out-of-sample questions. *International Educational Data Mining Society*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. 2022. Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8577–8591.

Wael Hassan Gomaa and Aly Aly Fahmy. 2014. Arabic short answer scoring with effective feedback for students. *International Journal of Computer Applications*, 86(2):35–41.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Noelle LaVoie, James Parker, Peter J Legree, Sharon Ardison, and Robert N Kilcullen. 2020. Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, 80(2):399–414.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 752–762.

Rodney D Nielsen, Wayne Ward, and James H Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501.

Rodney D Nielsen, Wayne H Ward, James H Martin, and Martha Palmer. 2008. Annotating students' understanding of science concepts. In *LREC*. Citeseer.

Leila Ouahrani and Djamal Bennouar. 2020. Ar-asag an arabic dataset for automatic short answer grading evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2634–2643.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Jana Z Sukkarieh, Stephen G Pulman, and Nicholas Raikes. 2004. Auto-marking 2: An update on the ucles-oxford university research into using computational linguistics to score short, free text responses. *International Association of Educational Assessment, Philadephia*.

Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075.

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

Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu, and Fuzhen Zhuang. 2022. An automatic short-answer grading model for semi-open-ended questions. *Interactive learning environments*, 30(1):177–190.