Deep Learning Techniques for Automatic Short Answer Grading: Predicting Scores for English and German Answers

Jörg Sawatzki, Tim Schlippe, Marian Benner-Wickner

IU International University of Applied Sciences
tim.schlippe@iu.org

Abstract. We investigate and compare state-of-the-art deep learning techniques for *Automatic Short Answer Grading*. Our experiments demonstrate that systems based on the *Bidirectional Encoder Representations from Transformers* (BERT) [1] performed best for English and German. Our system achieves a Pearson correlation coefficient of 0.73 and a Mean Absolute Error of 0.4 points on the Short Answer Grading data set of the University of North Texas [2]. On our German data set we report a Pearson correlation coefficient of 0.78 and a Mean Absolute Error of 1.2 points. Our approach has the potential to greatly simplify the life of proofreaders and to be used for learning systems that prepare students for exams: 31% of the student answers are correctly graded and in 40% the system deviates on average by only 1 point out of 6, 8 and 10 points.

Keywords: automatic short answer grading, artificial intelligence in education, natural language processing, deep learning.

1 Introduction

The research area "AI in Education" addresses the application and evaluation of Artificial Intelligence (AI) methods in the context of education and training [3]. One of the main focuses of this research is to analyze and improve teaching and learning processes. Many educational institutions–public and private–already conduct their courses and examinations online. This means that student examinations and their assessments are already available in digital, machine readable form, offering a wide range of analysis options. An exam typically consists of multiple choice and free text questions. While answers to multiple choice questions can easily be evaluated by machines, the evaluation of free text answers still requires tedious manual work by the examiners.

The focus of our paper is on Automatic Short Answer Grading (ASAG) with deep learning, i.e., the evaluation and further development of various deep learning state-of-the-art approaches to automatically evaluate free text answers. We investigate the following two architectures to compare a student answer with a given model answer and to predict an evaluation in the form of a score:

1. Our *feature extraction architecture* uses general pre-trained sentence embeddings in a high-dimensional semantic vector space as input values and predicts a score using a linear classifier.
2. Our *fine-tuning architecture* is based on a pretrained deep learning model, which–supplemented by a linear layer–is adapted to the specific task of ASAG. In contrast to our *feature extraction architecture*, the parameters of the embeddings are tuned as well.

A data set with manually graded exams and sample solutions from the area *Business Administration* of a German bachelor's program serves to optimize models and evaluate quality. To evaluate and classify the findings in the context of current research, an English data set from the University of North Texas with questions from the undergraduate studies of *Computer Science* is also used.

In the next section, we present the latest approaches of other researchers for ASAG. Section 3 describes our experimental setup. Section 4 characterizes the models which we evaluate and compare. Our experiments and results are outlined in Section 5. We conclude our work in Section 6 and suggest further steps.

2 Related Work

A good overview of rule-based and statistical-based approaches in ASAG before the deep learning era is given in [5]. Newer publications are based on *bag-of-words*, a procedure based on term frequencies [5,6]. The latest trend which has proven to outperform traditional approaches is to use neural network-based embeddings, such as *Word2vec* [7]. [8] have developed *Ans2vec*, a *feature extraction architecture*-based approach. It is based on combine-skip sentence embedding [9] and logistic regression. They evaluated their concept with a non-public data set from the University of Cairo, the SciEntsBank data set [10], and the English data set of the University of North Texas [2]. This English data set of the University of North Texas–like the German data set of our university–contains scored student answers. Consequently, it is a comparative data set which is also evaluated in the experiments of this paper. Like us, [8] use the Pearson correlation coefficient for evaluation. They report a best value of 0.63 on the data of the University of North Texas. [11] use a data set from the Hewlett Foundation[1] to compare different approaches based on deep learning models, including a *fine-tuning architecture* based on the *Bidirectional Encoder Representations from Transformers* (*BERT*) [1]. [12] and [13] deal in their work exclusively with

[1] https://www.kaggle.com/c/asap-sas

BERT fine-tuning architectures. In their work, answers are categorized into 3 classes–there is no point-based grading. *BERT* also provides the basis for our *fine-tuning architecture*, but we focus on point-based grading. [14] and [15] developed systems which classify German student answers as "correct" and "wrong"–based on traditional features such as lemmas.

Our contributions are the analysis of deep learning architectures for transfer learning on an English and a German data set. This includes the investigation of multilingual deep learning models. We are the first to examine point based ASAG of questions with variable maximum score.

3 Experimental Setup

In this section we describe our evaluation metrics and corpora.

3.1 Evaluation Metrics

As in related literature, we evaluate our results with the *Pearson correlation coefficient* [16], the *Mean Absolute Error* and the *Root Mean Square Error (RMSE)*.

Pearson Correlation Coefficient. The Pearson correlation coefficient (*Pearson*) indicates how strong the linear relationship between predictions and target values is. If the value is close to 0, there is no correlation; if it is close to 1, there is a strong correlation. The Pearson correlation coefficient is the normalized covariance. Therefore, it is independent of the scaling used in the data.

Mean Absolute Error. The Mean Absolute Error *(MAE)* is calculated from the average deviations of the prediction from the target value. It depends on the units and scaling in the evaluated data set.

Root Mean Square Error. Unlike Mean Absolute Error, in the Root Mean Square Error *(RMSE)* the error is squared, and the square root of the mean square deviation is considered. Squaring the error results in strong deviations being weighted more heavily than small ones.

3.2 Corpora

We evaluate our experiments with a German and an English data set which are compared in Tab. 1. To provide insights into both data sets[2], Tab. 2 and 3 indicate for each

[2] Questions and answers were modified for the German data set due to confidentiality.

data set, a typical question, the corresponding model answer and two student answers. One of them was given the full score, the other one is a rather weak answer.

Table 1. Information of the data sets.

	German	English
Subject	Business Administration	Data Structures
#questions with model answer	233	87
#answers (total)	3.560	2.442
#answers per question	15.4	28.1
Ø length of answer (#words)	87.6	18.4
Maximum scores (in points)	6 / 8 / 10	5
Annotated model answer	yes	no

3.3 English Short Answer Grading Data Set

The short answer grading data set of the University of North Texas [2] contains 87 questions with corresponding model answer and on average 28.1 evaluated answers per question about the topic *Data Structures* from the undergraduate studies. The questions are rather short and are not divided into sub-questions. They can usually be answered in only one sentence and no knowledge transfer is required.

Table 2. Original sample question and answers from the English data set.

Question	What is a variable?
Model answer	A location in memory that can store a value.
Example: Answer 1	A variable is a location in memory where a value can be stored.
Grading: Answer 2	5 of 5 points
Example: Answer 2	Variable can be an integer or a string in a program.
Grading: Answer 2	2 of 5 points

3.4 German short answer grading data set

The German data set is taken from an online exam system in the learning management system *Moodle*[3]. It contains 233 questions with corresponding model answer and on average 15.4 evaluated answers per question from the bachelor module *Business Administration*. A special feature of the German data set is that the maximum achievable score varies from question to question. Depending on the question a maximum of 6, 8 or 10 points can be achieved. Another feature is that the model answers include annotations with the criteria for grading performance. Many model answers contain only short hints for the corrector, so that in many cases additional background knowledge is needed for correction in addition to the model answer. A question usually consists of several sub-questions on a common topic and, in addition to the pure

[3] https://moodle.org

reproduction of knowledge. In many cases knowledge transfer is expected from the students.

Table 3. Modified and translated sample question and answers from the German data set.

Question	• Explain: What is the role of models in business administration? • Explain how statements of a model can be distinguished from each other according to the completeness of the information.
Model answer	In business administration, models are used to obtain, formulate, and test knowledge from the operational context (2 points). Statements with complete information are statements with certainty (3 points). Statements with incomplete information are statements under uncertainty or risk (3 points).
Example: Answer 1	a) In business administration, models are used to explain, describe, forecast and design macro- and microeconomic phenomena. b) Complete information represents security. Incomplete information represents uncertainty and risk.
Grading: Answer 2	8 of 8 points
Example: Answer 2	Models are used for information. Explanatory model: Explains reasons in the company, e.g., employee motivation. Descriptive model: Describes business phenomena in the company, e.g., accounting which records the entire flow of money in the company. Decision model: Here different information is combined with each other. For example, the optimal order quantity. This depends on various factors.
Grading: Answer 2	2 of 8 points

4 Techniques

This paper describes and compares the following two architectures for transfer learning: A *feature extraction architecture* and a *fine-tuning architecture*.

4.1 Feature Extraction Architecture

This architecture is based on the *Ans2vec* approach described by [8]: The model answer and the student answer are first converted into the two embedding vectors *MA* and *SA*. Then the dot product and the absolute difference of the two embedding vectors are calculated and concatenated. The result of the concatenation is the input vector for a linear model to predict the score.

Ans2vec-Skip-Logit-Baseline. [8] use combine-skip vectors for the embeddings and logistic regression as a classifier (*Ans2vec-Skip-Logit-Baseline*). Since no combine-skip embeddings are available for German, we evaluated this model only on English.

Ans2vec-MUSE-Logit. *Ans2vec-MUSE-Logit* refers to a model that corresponds to *Ans2vec-Skip-Logit-Baseline*, but for sentence embedding the *Multilingual Universal Sentence Encoder* (*MUSE*) [17] is used.

Ans2vec-Skip-SVM. *Ans2vec-Skip-SVM* refers to a model that corresponds to *Ans2vec-Skip-Logit-Baseline*, but for the classification a *Support Vector Machine* (*SVM*) is used [18]. Due to the lack of German combine-skip embeddings, we also evaluated this model only on the English data set.

Ans2vec-MUSE-SVM. *Ans2vec-MUSE-SVM* refers to a model that corresponds to *Ans2vec-Skip-Logit-Baseline*, but for the sentence embedding the *MUSE* is used and for the classification an *SVM.*

4.2 Fine-Tuning Architecture

This architecture is based on *BERT* [1] from the family of transformer models. We supplemented *BERT* with a linear regression layer that provides a prediction of the score given an answer. The model takes the model answer and the student answer without prior embedding as input, separates the model answer and the student answer with a *separator token* and performs a tokenization into *word pieces*. Since in our German and English data sets the scores are not only discrete, integer values, this approach uses regression instead of classification. Our evaluated *BERT* models are characterized in the following sections.

BERT-EN. *BERT-EN* refers to the English BERT[4] published by [1].

BERT-DE-Deepset. *BERT-DE-Deepset* refers to a German *BERT* model provided by Deepset GmbH. The model is trained on Wikipedia and Open Legal Data[5].

BERT-Multilingual. *BERT-Multilingual* refers to a multilingual *BERT* model[6] which is published by Google, supports 104 languages and is trained on Wikipedia.

5 Experiments and Results

Randomization and splitting of the data sets into training, validation and test data using 5-fold cross-validation is performed in all experiments to determine most accu-

[4] https://github.com/google-research/bert
[5] https://deepset.ai/german-bert
[6] https://github.com/google-research/bert/blob/master/multilingual.md

rate models for the German and English data as shown in Tab. 4. After the most accurate models for the German and English data set were determined, we evaluated them on the held-out test set.

Table 4. Preparation of data sets for cross-validation.

Data set	Portion	German	English
Total	100%	3,560	2,442
Cross validation	80%	2,848	1,953
Test (held-out)	20%	712	489
Cross validation	100%	2,848	1,953
Training	80%	2,278	1,953
Validation	20%	570	391

Table 5. Basic experiments with the English data set.

Model	Pearson	RMSE	MAE
[7]	0.63	0.91	-
Ans2vec-Skip-Baseline	0.33 (+0.0%)	1.27 (+0.0%)	0.73 (+0.0%)
Ans2vec-Skip-SVM	0.49 (+48.6%)	1.09 (-14.0%)	0.60 (-16.8%
Ans2vec-MUSE-Logit	0.38 (+13.6%)	1.24 (-2.2%)	0.69 (-4.5%)
Ans2vec-MUSE-SVM	0.56 (+67.5%)	1.02 (-19.1%)	0.56 (-23.7%)
BERT-EN	**0.79** (+138.5%)	**0.69** (-45.3%)	**0.41** (-43.3%)
BERT-Multilingual	0.79 (+137.1%)	0.70 (-44,6%)	0.43 (-44,6%)

5.1 English Automatic Short Answer Grading

The results of the experiments with the English data set and their relative improvements compared to *Ans2vec-Skip-Logit-Baseline* are shown in Tab. 5. For comparison, the first line also contains the values published by [8]. Since they do not provide further details on the implementation, parameters, and the procedure for evaluating the model, the reasons for deviation cannot be further analyzed. If instead of the combine-skip embeddings the *MUSE* embeddings are used, the results improve, and the training effort is reduced considerably. With only 512 dimensions, the *MUSE* embeddings are significantly more compact than the combine-skip vectors with 4,800 dimensions. The *Ans2vec* model also provides better predictions if the linear regression is replaced by an *SVM* classifier. However, the *BERT* models provide the best results. *BERT-EN* is the best of all models, but *BERT-Multilingual* provides only slightly worse numbers. With a *Pearson* of 0.79, we see that the scores predicted by the *BERT-EN* model have a strong linear relationship with the scores decided by the human corrector. On average, the evaluation of an answer by the model deviates by 0.41 points (see *MAE*). The *RMSE*, which weighs more strong deviations, also has the lowest number in this model. Compared to the numbers published by [8], the *BERT-EN* model achieves a relative improvement of more than 25% in *Pearson* and 23% in *RMSE*.

5.2 German Automatic Short Answer Grading

The results of the experiments with the German data set are shown in Tab. 6. Comparing these results with those of the evaluation of the English data set, only *Pearson* may be used. *MAE* and *RMSE* depend on the scaling of the score, which is different for both data sets. Looking at the linear correlation, one will notice that *Ans2vec-MUSE-Logit* performs slightly better on the German data set, while *Ans2vec-MUSE-SVM* performs slightly better on the English data set. The results of *BERT* on the German data set are only slightly worse than on the English data, even if the German exam questions are considerably more complex and extensive.

Table 6. Basic experiments with the German data set.

Model	Pearson	RMSE	MAE
Ans2vec-MUSE-Logit	0.39	2.52	1.87
Ans2vec-MUSE-SVM	0.44 (+11,2%)	2.68 (+6,6%)	1.90 (+1,9%)
BERT-DE-Deepset	0.74 (+89.2%)	1.76 (-30.3%)	1.31 (-29.7%)
BERT-DE-DBMDZ	**0.75** (+90.2%)	**1.75** (-30.6%)	**1.30** (-30.6%)
BERT-Multilingual	0.71 (+81.4%)	1.83 (-27.4%)	1.38 (-26.3%)

The German data set also demonstrates that the *BERT fine-tuning architecture* produces significantly better results. Again, the monolingual *BERT* model–in this case the *BERT-DE-DBMDZ*–is slightly better than the multilingual model. The best model is *BERT-DE-DBMDZ* which–with a *Pearson* of 0.75–shows a strong linear relationship between prediction and actual scores. The model's prediction deviates by 1.30 points from the human corrector's grading (see *MAE*). Compared to *Ans2vec-MUSE-Logit*, *Pearson* could be improved by 90% relative. The *RMSE* and *MAE* are almost 31% lower than the numbers of the *Ans2vec-MUSE-Logit* model.

5.3 Experiments with Removed and Added Annotations on the German Data Set

We removed the annotations with the criteria for grading (e.g., "(2 points)", see Tab. 3) and re-evaluated the best model *BERT-DE-DBMDZ*. Additionally, we added annotations for the maximum achievable score of each question to the model answers (e.g., "maximum 8 points"). The results of the further experiments compared to *BERT-DE-DBMDZ* can be found in Tab. 7. Removing the annotations reduces the quality of the model's predictions, adding the annotations slightly improves them.

Table 7. Further experiments with the German data set.

Model	Pearson	RMSE	MAE
BERT-DE-DBMDZ	0.75	1.75	1.30
Annotations removed	0.74 (-1.4%)	1.78 (+2.0%)	1.31 (+1.3%)
Max. score annotated	**0.75** (+0.8%)	**1.73** (-0.9%)	**1.28** (-1.7%)

5.4 Final Results

After the most accurate models for the German and English data set were determined, we evaluated them on the unseen *test set*. As shown in Tab. 5 and 7, *BERT-EN* is the best model on the English data set and *BERT-DE-DBMDZ (Maximum score annotated)* on the German data set. The results of the final evaluation are illustrated in Tab. 8. For English, the relative improvement compared to [8] is also shown.

Table 8. Final results on the unseen text sets.

Language	Pearson	RMSE	MAE
English	**0.73** (+15.5%)	**0.72** (-20.9%)	**0.42**
German	**0.78**	**1.62**	**1.19**

6 Experiments and Results

We investigated and compared state-of-the-art deep learning techniques for *Automatic Short Answer Grading*. With our *BERT* models we achieved a significant performance improvement compared to our baseline system and related work. Our system achieves a Pearson correlation coefficient of 0.73 and a Mean Absolute Error of 0.4 points on the Short Answer Grading data set of the University of North Texas [2]. On our German data set we report a Pearson correlation coefficient of 0.78 and a Mean Absolute Error of 1.2 points. The result on our English and German data sets were comparable even though the German data set contains more complex questions and has variable maximum scores.

Future work will include an analysis of what types of questions and answers the system still has issues and how we can tackle them. For example, examination questions often contain sub-questions. It could be evaluated whether their separation into individual questions leads to better predictions. We also plan to analyze to what extent the quality of the model improves by training the model on further subject-specific corpora such as lecture notes or textbooks as suggested by [19]. The developed model could also be used as a warning system: The system detects when a human corrector's grading significantly deviates from the model (e.g., by a defined threshold) and initiates further steps, e.g., the transfer of the relevant student answer and correction to a further review process. Furthermore, we plan to investigate the time savings by our automation. For example, our first more detailed analyses of the results of the German data set indicate that in 31% of all cases the score can be just accepted. In 39.6% of the cases the suggested score only needs to be corrected 1 point up or down out of 6, 8 and 10 points. This means that in 70.6% the total score does not have to be corrected at all or only by 1 point, which could lead to significant time savings in the correction process and to be used for learning systems that prepare students for exams [20].

References

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
2. 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, Portland, Oregon, USA. Association for Computational Linguistics.
3. Paul Libbrecht, Thierry Declerck, Tim Schlippe, Thomas Mandl, and Daniel Schiffner. 2020. NLP for Student and Teacher: Concept for an AI based Information Literacy Tutoring System. In The 29th ACM International Conference on Information and Knowledge Management (CIKM2020), Galway, Ireland.
4. Neslihan Süzen, Alexander N. Gorban, Jeremy Levesley, and Evgeny M. Mirkes. 2020. Automatic short answer grading and feedback using text mining methods. Proedia Computer Science, 169:726–743.
5. Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education, 25(1), 60-117.
6. Fabian Zehner. 2016. Automatic Processing of Text Responses in Large-Scale Assessments. Ph.D. thesis, TU München.
7. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.
8. Wael Hassan Gomaa and Aly Aly Fahmy. 2019. Ans2vec: A scoring system for short answers. In The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019), pages 586–595, Cham. Springer International Publishing.
9. Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15, page 3294–3302, Cambridge, MA, USA. MIT Press.
10. Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.
11. Surya Krishnamurthy, Ekansh Gayakwad, and Nallakaruppan Kailasanathan. 2019. Deep learning for short answer scoring. International Journal of Recent Technology and Engineering, 7:1712–1715.

12. Chul Sung, Tejas Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. Artificial Intelligence in Education, pages 469–481.
13. Leon Camus and Anna Filighera. 2020. Investigating Transformers for Automatic Short Answer Grading. Artificial Intelligence in Education, 12164:43–48.
14. Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In Proceedings of the TextInfer 2011 Workshop on Textual Entailment, pages 1–9, Edinburgh, Scottland, UK. Association for Computational Linguistics.
15. Ulrike Pado and Cornelia Kiefer. 2015. Short Answer Grading: When Sorting Helps and When it Doesn't. In Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning, NODALIDA 2015, Linköping Electronic Conference Proceedings, pages 42–50, Wilna. LiU Electronic Press and ACL Anthology.
16. Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58:240–242.
17. Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval. arXiv:1907.04307.
18. Theodoros Evgeniou and Massimiliano Pontil. 2001. Support vector machines: Theory and applications. Machine Learning and Its Applications: Advanced Lectures, 2049:249–257.
19. Matthias Wölfel. 2021. Towards the Automatic Generation of Pedagogical Conversational Agents from Lecture Slides. 3rd EAI International Conference on Multimedia Technology and Enhanced Learning (EAI ICMTEL 2021). Cyberspace.
20. Tim Schlippe and Jörg Sawatzki. 2021. AI-based Multilingual Interactive Exam Preparation. The Learning Ideas Conference 2021 (14th annual conference). ALICE - Special Conference Track on Adaptive Learning via Interactive, Collaborative and Emotional Approaches. New York, New York, USA.