Methods and perspectives for the automated analytic assessment of free-text responses in formative scenarios

Sebastian Gombert*

DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany

Abstract

Assessment is the process of testing learners' skills and knowledge. Free-text response items are well suited for the assessment of learners' active knowledge and writing skills. However, the automatic assessment of respective responses is not trivial and requires the application of natural language processing. Accordingly, the automatic assessment of free-text responses is a widely researched topic in educational natural language processing. Most past work targets holistic scoring, the process of assigning overall scores or grades to responses. This is problematic in formative scenarios because learners require feedback rather than summative scores in such scenarios. Such feedback ideally targets specific aspects of responses, and, accordingly, automated systems which only predict holistic scores cannot be used as a basis for providing the same. What is instead needed are systems which implement analytic scoring approaches. Analytic scoring targets specific aspects of responses and scores them according to corresponding criteria. This requires different systems than addressed by the broad research on automated holistic scoring. In my PhD work which is outlined by this paper, I want to explore approaches for implementing analytic scoring systems by means of state-of-the-art natural language processing. These systems are targeted at providing a basis for feedback generation.

Keywords

Assessment, Automated Assessment, Analytic Assessment, Short Answer Grading, Essay Grading

1. Introduction

Educational assessment is the process of empirically measuring and documenting learners' skills and knowledge [1]. This is conducted through tests composed of various kinds of test items. Assessing learners' knowledge and skills is also the basis for providing them with appropriate content-related feedback in formative scenarios [2]. In the context of technology-based assessment, multiple-choice items have grown to be a popular choice to implement tests [3, 4]. This is mostly the case since evaluating multiple-choice items is rather trivial. Test creators simply need to define a set of responses out of whom they define one or more as the correct ones. When test-takers select respective responses during testing, the computer only needs to determine which of them were among the correct ones. Moreover, multiple-choice items take only a short time to answer which makes it possible to include many different of them within tests and test for a broad range of knowledge [4].

However, not every skill and every kind of knowledge can be assessed through multiple-choice items. "A multiple-choice test for history students can test their factual knowledge. It can also determine whether they can discriminate between correct and incorrect statements of the relationships between facts — but it cannot determine whether the students can write a well-reasoned essay on a historical question. [...] A multiple-choice test of writing ability can determine whether the test takers can discriminate between well written and badly written versions of a sentence — but it cannot determine whether they can organize their own thoughts into a logically structured communication in clear and appropriate language" [4]. Moreover, multiple-choice cannot test for active knowledge. A test-taker might simply conduct (informed) guessing and there is no guarantee that they would have been able to actively reproduce this knowledge.

2. Constructed Responses and their Automatic Assessment

To test skills such as the ones described by [4], constructed response items are needed instead multiple choice items. In their most common form, they require students to enter a free text as response into a text field. However, this drastically increases the complexity of assessing learners' responses in an automated fashion, as the computer-based analysis of human language is far from trivial. With natural language processing respectively computational linguistics, a whole interdisciplinary field of research building upon various methods and theories from linguistics, artificial intelligence, statis-

Proceedings of the Doctoral Consortium of Seventeenth European Conference on Technology Enhanced Learning, September 12–16, 2022, Toulouse, France.

*Corresponding author.

gombert@dipf.de (S. Gombert)

https://edutec.science/team/sebastian-gombert/ (S. Gombert)

0000-0001-5598-9547 (S. Gombert)

tics, logic, psychology, cognitive science, software engineering and philosophy is dedicated to this issue, and the automatic processing of many aspects of language remains open research. What makes the automatic analysis of free text difficult are the properties of language itself. Humans can generate an unlimited set of different linguistic utterances, and often, there are many ways to express the same or similar semantics, i.e., through different synonyms, the usage of passive vs. active constructions, or ways of paraphrasing. In past research, many different methods were applied to the automatic assessment of free-text responses. These range from simpler keyword, pattern and regular expression searches, and methods building upon distributional vector space semantics, to fully-fledged machine learning systems [5, 6].

Most recently, transformer language models such as BERT [7] were successfully applied to the problem of free-text assessment [8, 9, 10, 11]. The application of transformers to the assessment of constructed responses promises major advancements in the field, but nonetheless, most of the systems available are built to predict only holistic scores [5, 6], ergo scores aimed at denoting the overall quality of a response [4]. Most of the established datasets, especially the ones focused on short answers, also cater towards this approach [5, 9, 6]. While holistic scores reflect how well learners were able to overall solve a given task, they do not necessarily denote which aspects of their response were of good quality and in which regards they could improve. However, especially in formative scenarios, providing students with feedback is crucial, which puts the application of holistic scoring systems in formative scenarios into question.

There is a second scoring approach in constructed response assessment which can be seen as a better basis for providing detailed, personalized feedback: analytic scoring. In analytic scoring, rather than judging responses as a whole, they are assessed for multiple different aspects which need to be specifically defined in a coding rubric [4]. I.e., "[o]n a science question, the scorer may award two points for providing a correct explanation of a phenomenon, one point for correctly stating the general principle that it illustrates, and one point for providing another valid example of that principle in action" [4]. Drawing such distinctions and coding responses for multiple different aspects allows to provide more detailed and concise feedback as the same can specifically address these aspects.

3. Research Questions

The two most common types of free-text responses are short answers and essays. While short answers are used to test students' ability to explain phenomena or demonstrate their active knowledge, essays are used for analysing their writing skills of students, e.g., their skill to clearly and coherently discuss or communicate a given issue or argue against or in favour of an opinion. Accordingly, approaches for the analytic assessment of both text forms must inevitably differ. For short answers, it presumably should be sufficient to simply assess whole responses for the different aspects, as short answers are rather condensed texts. From a formal point of view, this can be interpreted as a (multi-label) text classification task [5].

On the other hand, for essays, the respective coding can require more varied approaches. Are the aspects coded related to content or writing style? Does a content-specific code apply to the whole text or to specific sections? These questions need to be addressed in order to come to appropriate operationalisations. E.g., if it is likely that each code corresponds to a specific part of an essay, one needs to first semantically segment it into the respective parts. One could then in a second step separately classify these parts for the actual codes. On the other hand, if a code corresponds to a whole essay, such separation is not needed.

I plan my PhD to be paper-based where the single papers are connected by the overarching topic of analytic constructed response coding. First and foremost, I want to explore what has been already done in past work and how my own work can benefit from these insights. The acquired knowledge is then to be used for the practical implementation of constructed response scoring systems in a range of case studies. For these case studies, I plan to leverage data sets from several research projects I am involved in. In the projects *AFLEK* and *ALICE*, I have access to a set of short answers to different science-related tasks with detailed coding rubrics focusing on scientific knowledge and argumentative skills. On the other hand, the project *HIKOF* provides a data set of essays in which students discuss learning tips from a YouTube video with respect to their grounding in educational psychology. Both data sets are coded in a way which allows for the implementation of automated analytic assessment systems.

Another important aspect of my work is the question how codes from response scoring systems can be transformed into concrete learner feedback. Feedback can be given on an item-specific level as well as on a more global one. It can focus the content or the form of concrete responses, and it can also target the overall domain knowledge of a student across multiple items. For the prior case, generative language models could be promising [9, 12]. For the latter case, a way of modeling learners' domain knowledge is required. A conceptual framework which goes into this direction was provided by [13] with their *expanded evidence-centred design* model, which adds multiple feedback-related aspects to the well-known *evidence-centred design* [14]. However, to my best

knowledge, this conceptual framework was not operationalised into a concrete feedback-driven assessment system so far.

The last aspect I want to address is the one of explainability. Ethical frameworks in learning analytics and educational technology such as [15] often call for the application of transparent and explainable models where possible. It is likely that providing learners with simple explanations on why models made a given prediction, which, in turn, led to a particular feedback outcome, can increase their acceptance for respective systems. For natural language processing models, a wide range of methods for providing such explanations has been developed [16]. Research for making state-of-the-art methodology explainable also shows promising results, e.g. [17]. For this reason, I want to leverage this potential and explore, if providing learners with explanations for their feedback can increase trust.

To summarzie, I want to address the following research questions:

1. What were the main methods, characteristics and results of past work in constructed response scoring?
2. What techniques were applied for coding constructed responses in an analytic fashion in past work?
3. What machine learning-based pipelines and approaches are effective for the automated analytic assessment of constructed responses and to what extent can they be generalized?
4. How can the predictions of automated analytic assessment systems be transformed into useful learner feedback?
5. To what extent can explaining model outputs make learners trust in the provided feedback?

4. Design

From a technical perspective, the intention behind my PhD work is to implement and evaluate respective methods for the analytic assessment of free-text responses for exemplary use cases drawing from state-of-the-art NLP research. I plan to study and summarize what methods were applied to the assessment of free-text responses in past work via a literature review to address RQ1 and RQ2. For this literature review, I plan to draw from past reviews on the topic, in particular [5] for the text type of short answers and [6] for the text type of essays, but primarily with a focus on work which was not covered by them. The main goal behind the literature review is to provide a concise overview over the methods and features which can be successfully applied to the task.

The review by [5] is, thanks to its publication date, fairly outdated. Moreover, in my opinion, it fails to function as a lookup guide for possible techniques to use, and rather focuses on summarizing papers from past work. The review by [6], on the other hand, is well structured but also fairly short thanks to it being published in conference proceedings. The plan for my literature review is to primarily act as a guide for practitioners which they can refer to when they plan to build their own free-text assessment systems rather than as a pure overview over past work. It shall equip interested researchers with a clear plan on how they can approach their own free-text response assessment system in a structured manner.

The next papers deal with the implementation of respective systems themselves to address RQ3. The most recent achievements in holistic free-text response assessment, in line with the general developments in natural language processing, were achieved using transformer language models [8, 9, 10, 11]. For this reason, my plan is to also apply transformer language models to the task of analytic assessment. However, [5] and [6] document a wide range of methods from the pre-transformers era. It is an interesting question In this context, my plan is to implement and evaluate exemplary systems for assessing both short answers and essays in an analytic fashion.

In a first research paper, which is currently under review, I implemented and evaluated multiple systems aimed at assessing German middle school students' knowledge about energy physics. In particular, the systems classify if students mentioned certain concepts related to energy transformation, i.e., different manifestations of energy, indicators for the same, and if energy is transformed, in a meaningful manner. For this purpose, first data was collected and coded using a coding rubric which targeted the different categories of knowledge. I then implemented and evaluated multiple text classification systems trained to replicate the coding for the respective purpose, transformer- and feature-based. The systems are given the response, a provided sample solution and the item prompt. Moreover, using different methods for generating model explanations, I evaluated the descriptive accuracy of the implemented models. Overall, a transformer-based model based upon GBERT could achieve superior results. In subsequent research, I want to explore how well the predictions of such systems can be concretely translated into feedback.

In another research paper, I want to implement systems targeting essays. In particular, I aim to use a data set of essays collected throughout the *HIKOF* project. These essays discuss ten different learning tips presented in a YouTube video with respect to their grounding in educational and psychological research. For each tip, ten different codes were assigned. Moreover, it was coded which sentences within an essay correspond to which tips. This results in two problems which need to be solved.

First, unseen essays must be segmented into sections corresponding to the different tips. This can be approached as a sentence classification task. In a second step, the resulting sections must then be given to a second text classification system which classifies the sections with respect to the analytic codes corresponding to each tip.

In the next step, feedback needs to be generated from the predicted codes. For this purpose, I use content-related feedback templates which are assembled dynamically depending on the predicted codes. In particular, the predicted codes are matched with ground truth codes, and discrepancies between the two lead to The generated feedback will be tested within a university lecture in an AB setup. In a followup study, I plan to add aspects of explainability to this feedback. In particular, I plan to present learners with highlighted text of what exactly in their response led to a concrete feedback in an AB setup. This shall then be combined with questionnaires evaluating if showing these explanations to learners increases acceptance. For educational recommender systems, findings from [18] suggest that showing explanations to learners can increase the acceptance for respective systems. I want to find out if this is also the case for assessment-driven feedback systems.

5. Conclusion

In this document, I presented my PhD project which deals with systems for the automatic assessment of constructed responses in formative scenarios implemented through machine learning-based natural language processing. In particular, I explore the implementation and evaluation of respective systems for multiple use cases. Moreover, I plan to write a literature review on constructed response scoring in the form of a practitioner lookup guide. Finally, I then want to explore how codes predicted by automatic assessment systems can be translated into automatic actionable feedback, and if explaining the model predictions behind this feedback can contribute to the acceptance of these systems.

References

[1] R. Nelson, P. Dawson, A contribution to the history of assessment: how a conversation simulator redeems socratic method, Assessment & Evaluation in Higher Education 39 (2013) 195–204. URL: https://doi.org/10.1080%2F02602938.2013.798394. doi:10.1080/02602938.2013.798394.

[2] P. Black, D. Wiliam, Assessment and classroom learning, Assessment in Education: Principles, Policy & Practice 5 (1998) 7–74. URL: https://doi.org/10.1080/0969595980050102. doi:10.1080/0969595980050102. arXiv:https://doi.org/10.1080/0969595980050102.

[3] S. K. Mangal, S. Mangal, Assessment for Learning, PHI Learning, Delhi, India, 2019.

[4] S. A. Livingston, Constructed-response test questions: Why we use them; how we score them. r&d connections. number 11., Educational Testing Service (2009).

[5] S. Burrows, I. Gurevych, B. Stein, The eras and trends of automatic short answer grading, International Journal of Artificial Intelligence in Education 25 (2014) 60–117. URL: https://doi.org/10.1007%2Fs40593-014-0026-8. doi:10.1007/s40593-014-0026-8.

[6] Z. Ke, V. Ng, Automated essay scoring: A survey of the state of the art, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2019. URL: https://doi.org/10.24963%2Fijcai.2019%2F879. doi:10.24963/ijcai.2019/879.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[8] L. Camus, A. Filighera, Investigating transformers for automatic short answer grading, in: Lecture Notes in Computer Science, Springer International Publishing, 2020, pp. 43–48. URL: https://doi.org/10.1007%2F978-3-030-52240-7_8. doi:10.1007/978-3-030-52240-7_8.

[9] A. Filighera, S. Parihar, T. Steuer, T. Meuser, S. Ochs, 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), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8577–8591. URL: https://aclanthology.org/2022.acl-long.587. doi:10.18653/v1/2022.acl-long.587.

[10] A. Poulton, S. Eliens, Explaining transformer-based models for automatic short answer grading, in: 2021 5th International Conference on Digital Technology in Education, ACM, 2021. URL: https://doi.org/10.1145%2F3488466.3488479. doi:10.1145/3488466.3488479.

[11] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, R. Arora, Pre-training BERT on domain resources for short answer grading, in: Proceed-

ings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6071–6075. URL: https://aclanthology.org/D19-1628. doi:10.18653/v1/D19-1628.

[12] A. Filighera, J. Tschesche, T. Steuer, T. Tregel, L. Wernet, Towards generating counterfactual examples as automatic short answer feedback, in: Lecture Notes in Computer Science, Springer International Publishing, 2022, pp. 206–217. URL: https://doi.org/10.1007%2F978-3-031-11644-5_17. doi:10.1007/978-3-031-11644-5_17.

[13] M. Arieli-Attali, S. Ward, J. Thomas, B. Deonovic, A. A. von Davier, The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design, Frontiers in Psychology 10 (2019). URL: https://doi.org/10.3389%2Ffpsyg.2019.00853. doi:10.3389/fpsyg.2019.00853.

[14] R. J. Mislevy, R. G. Almond, J. F. Lukas, A brief introduction to evidence-centred design, ETS Research Report Series 2003 (2003) i–29. URL: https://doi.org/10.1002%2Fj.2333-8504.2003.tb01908.x. doi:10.1002/j.2333-8504.2003.tb01908.x.

[15] S. Slade, A. Tait, Global guidelines: Ethics in learning analytics (2019).

[16] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459. URL: https://aclanthology.org/2020.aacl-main.46.

[17] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 782–791.

[18] K. Takami, Y. Dai, B. Flanagan, H. Ogata, Educational explainable recommender usage and its effectiveness in high school summer vacation assignment, in: LAK22: 12th International Learning Analytics and Knowledge Conference, LAK22, Association for Computing Machinery, New York, NY, USA, 2022, p. 458–464. URL: https://doi.org/10.1145/3506860.3506882. doi:10.1145/3506860.3506882.