Towards Better Support for Machine-Assisted Human Grading of Short-Text Answers

Alexander Meisl
WU Vienna
University of Economics and Business
alexander.meisl@gmx.at

Gustaf Neumann
WU Vienna
University of Economics and Business
gustaf.neumann@wu.ac.at

Abstract

This paper aims at tools to help teachers to grade short text answers submitted by students. While many published approaches for short-text answer grading target on a fully automated process suggesting a grading result, we focus on supporting a teacher. The goal is rather to help a human grader and to improve transparency rather than replacing the human by an oracle. This paper provides a literature overview of the numerous approaches of short text answer grading which were proposed throughout the years. This paper presents two novel approaches (answer completeness and natural variability) and evaluates these based on published exam data and several assessments collected at our university.

1. Introduction

In the last 20 years, automated processing of student's text submissions has gained a lot of attention due to the rise of educational institutions' digital transformation. Most universities and schools have established online e-learning tools where students have access to learning materials and can submit solutions for their training tasks. Consequently, more and more textual information is gathered from students.

In addition to that, through globalization more people have the chance to access and benefit from academic institutions which forced them to broaden their digital environment. As a consequence, a frequent physical appearance at the university is heavily questioned from the professors', the institutions' and the students' point of view. While the latter discussion is highly complex in its nature and widely disputed, educational institutions recognized the opportunity to provide electronic assessments whether as additional training material or as testing the knowledge before the examination [1].

This shift in the test environment did not only have stressful, workload-increasing and demanding consequences inherited, it opened up a wide range of new opportunities. One of these is to pair the manual grading with automatic machine grading [2]. It means that a machine processes, evaluates and grades a person's textual submission while, afterwards, a human supervises the outcome and checks its validity and correctness. We call this machine-assisted human grading. This approach promises a faster throughput of assessed student answers and can possibly lead to higher transparency of the final grade [3, 4, 5].

As a consequence, a huge number of different algorithms have been developed to foster this approach, especially in the area of short text answers. The need of understanding which schemes fit which purposes has increased tremendously. Comparably to the field of statistics, only appropriate tools can lead to right and meaningful assumptions [6]. Thus, the study aims to answer following research questions: *(1) Which factors of a Natural Language Processing tool are crucial for supporting examiners when grading short text answers? (2) How can a software contribute to a transparent grading of short text answers?*

This paper's prime contribution is to highlight and discuss an alternative approach which emphasizes the examiner's knowledge and awareness on submission properties rather than concentrating on an automatic grading.

To accomplish these objectives, this paper reflects on the theoretical background of relevant literature, followed by a thoroughly described research setting. Then, the findings of the study are presented. The paper concludes with a discussion, implications on practitioners, limitations and final remarks.

URI: https://hdl.handle.net/10125/79340
978-0-9981331-5-7

2. Review of relevant literature

Research associated with Natural Language Processing (NLP) has been done from numerous different researchers since many years. Especially in the area of essay grading, investigative studies have been conducted since approximately 50 years [7]. Dating back to the late 90ies, it was reviewed which methodological approaches are necessary or suitable to assess essays in a virtual setting. They highlighted two methodologies which still exist until today, i.e., statistical methods like regression analysis and latent semantic analysis [8].

Followingly, a study was conducted to further assess or more precisely score essays as well as short answer text submissions [9]. But not only different approaches were investigated the predominant ideas to automatically assess text fragments were challenged with ethical and moral viewpoints. They questioned if the design criterion of the tools and the factors, which influenced the scoring, would fully and fairly assess these text submissions and, moreover, lead to transparent grades [10, 11, 12].

Those early challenges led to many studies which focused on one fundamental aspect in regards to NLP, i.e., the complexity of the topic itself. A few studies [4, 5, 13] address rules or internal guidelines human graders follow to assess written texts. Especially, [13] pinpointed that grading text is far more complex and not streamlined throughout human examiners' training courses. This led to differences in text quality aspects as well as grading guidelines [5]. Furthermore, text quality and design elements within a good essay differentiate enormously among human examiners and teachers worldwide [14, 12]. Subjective and human rater-dependent factors affect the final grade more independently on whether good text quality elements are used within an essay or not [5, 4, 13].

Moreover, [4] further emphasizes key advantages of NLP tools and techniques as the following; "it may potentially mitigate [...] inconsistent grading; conscious or unconscious bias; fatigue; working memory overload; discouragement or mood changes while scoring [...]" [4, p. 70].

2.1. Short text answer grading

A large body of research addresses short answer grading related systems and methodologies, many specific approaches have been developed to process and evaluate short text answers. Implementations range from pure statistical metrics to machine learning algorithms like neural networks [15]. Especially, comparisons of pre-trained transfer models are in focus of recent studies while they are also evaluated against more traditional approaches like Bag-of-Words or Support Vector Machines [16, 17].

One facet, which is still missing but very important, is a critical discussion about computer-based student submission grading and/or evaluation. Several publications [18, 19, 20] concentrate on empirical results (surveys, interviews) about automated grading tools. As for other technologies, validating its effects, proximity to desired goals and justification of relevance is very important for supporting grading. The articles [18, 19, 20] elaborate on the complexity of student submission grading when considering that writing a text needs multiple levels of understanding. Especially, [21] demonstrates that integrating complex scoring rubrics has the potential to seize on many levels of understanding but cannot fully replace human raters.

Nevertheless, criticism has not fallen silent throughout the years. Many publications discuss the potential threat such systems could cause when not configured well. Several publications [19, 22, 23, 3, 24, 25] critically assess current methodological approaches and applicable tools. Conclusively, almost all of the mentioned publications emphasize that computer-based evaluation has strengths and weaknesses associated. It heavily depends on the why and how, i.e., "Why do I want to incorporate such tools into my way of teaching?" and "How far do I use them in terms of text grading or preparing students to automatic evaluations?".

3. Research goals and experimental setup

The following section covers the research goals and the experimental setup in which this study was conducted. Then, the newly proposed approaches, answer completeness and natural variability, are described.

As the research trend heads towards automatically assessing and grading student answers based on various syntactic and semantic features, the proposed approaches highlight a different way of addressing this task. Rather than substituting the human grader by a system, the support for the human grader is in focus. How can a system be created to enable a human person to process quicker and more consistent potentially high number of student submissions? In this paper we provide techniques towards this goal, focusing on *Machine-Assisted Human Short Answer Grading*. In particular, the provided techniques focus on automatically extracting grading relevant factors for the human grader.

3.1. Addressed Requirements

The goal is to target real world exams, which are potentially different in every instance. This means that the setup costs of a grader should be little (and error robust) and should not require a reference answer (golden essay) or a potentially large exam-specific corpus.

For one-time exams, grading approaches based on supervised machine learning are very difficult to set up for the majority of teachers since there is usually no corpus or other training data available to train the system upfront.

Many automated grading systems work only with reference answers, and compare the distance of submissions to the reference answer. This approach has several disadvantages. First of all, especially for short text questions, there might be many different textual ways to express the result. Secondly, measuring the distance fosters rather exams on the lower levels of Bloom's taxonomy (memorizing) rather than transfer oriented questions, which are especially important in higher education. Knowledge transfer oriented questions, such as "express in your own words..." have also the advantage to make some forms of cheating more complex. So the goal was to explore methods independent of reference answers.

Further goals are subject independence and little limitations in regards to the natural language in use. The research questions of this paper address the feasibility of these goals as well as the development of techniques to approach these.

3.2. Processing steps

Figure 1 depicts the major processing steps for our experiments. Exams consist of one or many test-items, which are short text questions expecting submissions consisting of single lines up to a few paragraphs of answer text. These assessments are preprocessed using standard tools as provided by natural language processing tools such as NLTK [26]. Important elements of preprocessing are stopword elimination, lemmatization (transform provided words to its canonical form) and part-of-speech analysis (categorize words according to their syntactic function in sentences, e.g., 'noun'). In our implementation, we used the freely available tool TreeTagger [27] for tokenization, part-of-speech tagging and lemmatization. This selection emerged out of two specific reasons: accuracy and multilingualism. It supports a vast amount of different natural languages like English, French, German or Spanish, and achieved a strong score of

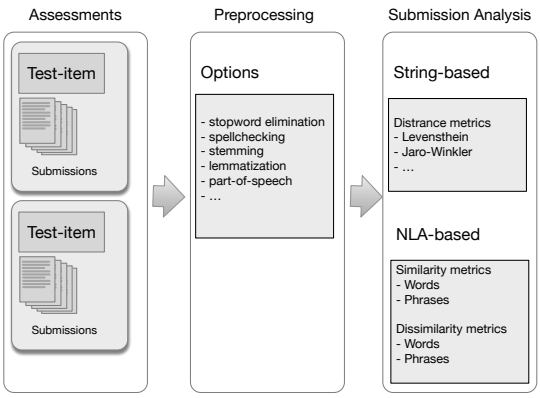

Figure 1. Processing Pipeline.

roughly 97% correctly tagged words (as presented in their study [28]).

We are comparing in this paper several algorithmic approaches that have different requirements concerning the processing pipeline. While the string-based methods are based on the plain bytestream, other approaches require part-of-speech (POS) tagging or lemmatization. The computation of the term frequencies based on the tf-idf (frequency–inverse document frequency) method is the only one requiring stop word elimination. Therefore, we create in our implementation[1] from the assessment and submission data a data structure containing the input in various stages such that these formats are available with only little overhead for further processing.

3.3. Answer Distance

One promising approach is to compute the distance matrix of all submissions to figure out how similar submissions are (see e.g., the power-grading approach [2]). Such a distance matrix could be used in a grading process to suggest to a human grader who has graded a submission, some other submissions as next which are the most similar ones to this. For this distance matrix, either string-based approaches might be used or also measures based on the lemmatized submissions (e.g., using a cosine distance). The string-based approaches have the advantage to cope quite well with misspelled submissions, since they work on the full character (byte) sequence of the submission. Natural language processing software has usually problems to work with misspelled inputs, since the attempted word lookup fails. In such cases automated spellchecker might be used; one can use e.g., Levenshtein distance [29] to

[1]The prototypical implementation containing pre-processing and analysis is available from `https://github.com/nm-wu/haSAAS`.

find the most appropriate word for a misspelled one, but certainly, this might change words into terms with different meanings, which is problematic to support the grading. In this paper, we just look at distance matrix calculations based on Jaro-Winkler [30], which takes into account only matching character sequences and character swappings (transpositions) which is well suited for misspelled texts.

3.4. Answer Completeness

When a grader receives multiple answers for a test item, one important aspect is, how complete the answer is, i.e., up to which degree an answer covers all relevant aspects of the question. To compute this task, the lemmatized submissions are a good starting point, since the singular/plural etc. differences can be eliminated. In absence of a golden essay (covering all aspects perfectly), we compute for every submission a text consisting all other submissions and compute the similarity of the submission to this text conglomeration.

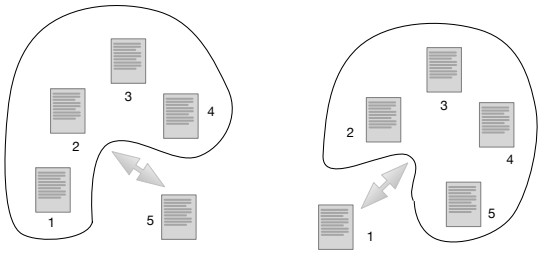

Figure 2. Exemplary visualization of the Completeness Similarity computation.

Figure 2 visualizes the completeness similarity computation: e.g., submission 1 is compared with $2 \cup 3 \cup 4 \cup 5$ whereas e.g., submission 5 is compared against $1 \cup 2 \cup 3 \cup 4$. Our hypothesis is that good submissions will have a large similarity with this text base.

More generally, for a submission S_i the similarity value with the text base can be defined as in Equation 1:

$$sim_i = \cos(\bigcup_{j=1}^{n} S_j - S_i, S_i) \tag{1}$$

Since the text quantity of the text base is much larger than a single submission, the term frequencies are weighted via the *tf-idf* (term frequency–inverse document frequency) method, which computes the term frequency relative to the document size. The weighted factors W are used then to compute the cosine distance between a weighted submission W_i and the weighted text base W_{base} (see Equation 2).

$$\cos(W_i, W_{base}) = \frac{W_i W_{base}}{\|W_i\| \|W_{base}\|} \tag{2}$$

3.5. Natural Variability

The basic idea of this metric is that when several students formulate answers to questions in their own words, they should differ up to a certain degree from each other. While grading student submissions we observed frequently that several students submit the same wrong arguments using very similar words, where it is not clear where they got this from. Or sometimes, some student submissions contain exactly the same passages. The metric of the natural variability should detect such cases, which could also be characterized as ”answer clones”. When students use pre-assembled summaries or copy the solution during an exam, the natural variability decreases because the same argumentation and syntactic elements are expressed respectively.

Our hypothesis is that good submissions will have a high natural variability. Alternatively, one can state that a variability below a certain threshold should be checked by the human grader. In general, a low variability can be the consequence of fully correct answers or as well totally wrong ones, depending on the type of question and the expected text amount.

The variability metric between two texts $T1$ and $T2$ is computed in three stages: First, all possible tri-grams are extracted from the lemmatized texts. Then the extracted tri-grams of $T1$ are checked against the ones of $T2$ and counted if they match. Ultimately, the actual number of occurrences is divided through the maximum possible ones to get a percentage of matched tri-grams. The same procedure is applied for bi-grams as well as for uni-grams. In total, three percentages are calculated which express the ratio of tri-/bi-/uni-grams occurring in both, $T1$ and $T2$. Second, these three ratios are weighted dependent on its severity, i.e., the presence of many tri-grams in both answers reflects a strong relation while the same uni-grams (single words) are mandatory in certain contexts. Therefore, the ratio of tri-/bi-/uni-grams is weighted as 0.6, 0.3 and 0.1, respectively (see Equation 3). As a consequence, the resulting number represents the percentage of overall matched occurrences in the reference answer.

$$\begin{aligned} \textit{total\_occurrence}(\%) = &(N_{\textit{tri-grams}} * 0.6) \\ &+ (N_{\textit{bi-grams}} * 0.3) \\ &+ (N_{\textit{uni-grams}} * 0.1) \end{aligned} \tag{3}$$

$$variability(\%) = 1 - total\_occurrence \tag{4}$$

As the *total_occurrence* stands for the actual presence in the referenced answer, the natural variability is semantically the contrary of it. To transform this score into the intended form of meaning, the complement of the actual presence is calculated as seen in Equation 4.

The computed variability can also be stipulated as the percentage of uniquely used collocations. Simply spoken, a weighted ratio of tri-/bi-/uni-grams used in student answer A existing in student answer B.

As a result, a complete matrix is calculated which signifies a quadratic effort. In total, $n^2 - n$ scores have to be computed where n stands for the number of submitted student answers. n is subtracted as the diagonal of the matrix represents the variability score of a student's answer to itself which is not used. This complete matrix is not symmetrical as it can occur that student answer A and B differ in their length. For a better understanding a small example is given below in (5-6):

$$\begin{aligned} A &= \textit{"aa bb cc dd ee"} \\ B &= \textit{"aa bb cc"} \\ C &= \textit{"aa bb cc dd ee ff gg"} \end{aligned} \tag{5}$$

$$\begin{array}{c|ccc} & A & B & C \\ A & . & 0.5900 & 0.0000 \\ B & 0.0000 & . & 0.0000 \\ C & 0.3686 & 0.7372 & . \end{array} \tag{6}$$

In this small example, the asymmetry can be clearly identified. The variability of A against B is 0.59, while the opposite results in 0.0. If we compare these scores against the ones between B and C, the variability of C against B is 0.73. Hence, the percentage of possible similar words in C is higher than in A if both are compared to B, it can be seen that the variability score increases the longer the student answer is. Furthermore, this asymmetrical complete matrix allows for detection of clones, i.e., if student answer B is a subset of student answer A, B as a clone of A can be therefore detected.

4. Results

For evaluating the presented metrics, we used previously published test cases of [17] and in addition, we tested with 73 assessments with a total of 995 submissions (738 in German and 257 in English) collected from several Information Systems courses at the Vienna University of Economics and Business (WU). Since all these assessments were manually graded, we used the manual grades for determining the usefulness and precision of the newly introduced metrics.

The upcoming section follow the order in which these metrics were introduced in Section 3. First we show how to use a hierarchical cluster analysis based on Jaro-Winkler distance to visualize distance-based (dis)similarity. It allows for grouping of student answers to quickly detect ones which could be graded in a sequence by a human grader. In Section 4.2 we look into use cases of the novel measure of the answer completeness. The main goal is to show, how well this metrics can be used to support grading by comparing it with the manual grading of results. This can be used to visualize the correlation between answer completeness and the grading result, where one can also see outliers or detect exams which were hard to solve by students. Although, it is not the main purpose of this paper, we compare how well the answer completeness compares with approaches based on golden essays. Finally in Section 4.3 we show how the metrics of natural variability can detect interesting facts about student submissions, which cannot be detected by string-based methods like the Jaro-Winkler distance metrics.

In the examples targeted on supporting the grader with visual tools, we focused on certain assessments. Section 4.2 contains the most complete analysis, addressing an comparison with other approaches and languages independence.

4.1. Answer Distance

In our first example we provide a graphical representation based on a hierarchical cluster analysis to visualize the textual similarity of the submissions of the exam. Figure 3 shows a single exam in English with 14 submissions which are displayed in form of a dendrogram that groups submissions by similarity.

The X axis is labeled with the submissions and the achieved points. The textually most similar submissions were submissions 5 and 12, one is graded with 2.5 and the other one with 2 points. There is a textually very similar orange colored cluster on the left resulting in gradings between 1.5 and 2.5 points.

Since the clustering is performed based on the text similarity, a grader can e.g., grade one exam and continue with a very similar submission from the same cluster which is following probably a similar argument line and will get probably similar grades too.

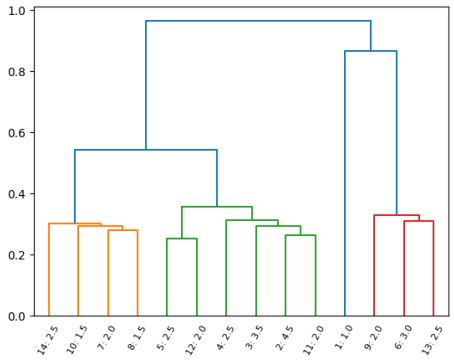

Figure 3. Hierarchical Cluster Analysis based on Jaro-Winkler Distance

4.2. Answer Completeness

In our experiment we observed differences depending on the type of the exams, whether these exams are exams during the semester, or end-of-term exams. In this section we compare eight different assessments in English which were all during the semester, taking 5 minutes per test item. For these eight assessments, a total of 102 submissions were graded. The grading result is normalized to a percentage (i.e., between 0 and 1).

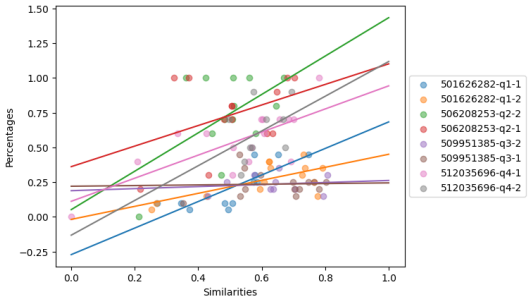

Figure 4. Correlation of Grading (in %) vs. Answer Completeness

Figure 4 shows on the X axis the computed answer completeness and on the Y axis (ordinate) the achieved percentage of points. The graphic shows only positive correlations between answer completeness and the manual grading. The figure gives also an impression of the difficulty of the questions to the student: assessments with higher difficulty show regression lines with higher ordinate values. Two of the assessments (509951385) have only a flat slope, indicating only a weak positive correlation. Interestingly, the students scored very bad on these assessments, since they got only 25% of the achievable points. Exactly obtaining the same results of a grader is not realistic, since also two humans are not grading the same submission identically. To get results from auto-grading comparable to human variability, the correlation should be above 0.6 [31]. The Pearson correlation factors behind Figure 4 can be seen as in the tabular representation. The RMSE (Root Mean Square Error) shows the error between the true and predicted scores (lower value is better).

Assessment	Pearson	RMSE	n
501626282-q1-1	0.8021	0.3089	12
501626282-q1-2	0.6743	0.3652	14
506208253-q2-2	0.5378	0.3478	13
506208253-q2-1	0.4128	0.3268	13
509951385-q3-2	0.1305	0.4193	10
509951385-q3-1	0.0350	0.4353	14
512035696-q4-1	0.7441	0.1509	16
512035696-q4-2	0.4086	0.1768	10
Weighted average	0.4824	0.3143	102

We see from this data that in some cases, the Pearson correlation and RMSE values on this assessment data from WU are even better than the results reported in [16] (best result Pearson correlation 0.592, RSME 0.8887). As indicated, the results of 509951385 are especially bad - probably for reasons out of the scope of an algorithm. As weights of the weighted average, the number of submissions are used such that results of larger assessments have a stronger influence on the results. Also, the weighted average of these assessment results would be in the top third of the 14 compared approaches of [16]. When performing this comparison on the Data of Mohler [17, 16] we see results very similar to the local tests - showing as well sometimes big differences.

Assessment	Pearson	RMSE	n
6.3	0.8586	0.3761	26
6.2	0.8536	0.3652	26
...			
11.7	-0.1698	0.2797	30
8.1	-0.3207	0.6144	27
Weighted average	0.4374	0.4438	4492

Next, we look into language independence, based on the English and German assessments from our collected data. We filtered from the assessments the cases, where we have just aggregated scores.

Assessment	Pearson	RMSE	n
Best (en)	0.8022	0.3089	12
Best (de)	0.8030	2.9689	16
Worst (en)	-0.3916	0.2334	13
Worst (de)	-0.5360	0.3141	11
Weighted average (de)	0.2156	0.1775	194
Weighted average (en)	0.4825	0.3143	102
Weighted average (all)	0.3075	0.2246	296

We can see that there are quite high differences in both English and German assessments, that overall the results are worse for our German assessments, but that some of the German assessment scores are comparable with the English ones. Future work will address a better understanding of the origin of these differences, and to find factors indicating the reliability of the results. One should also note that the distribution of scores from the Mohler testset (as shown in [16]) is strongly right-skewed (median 90% correct), whereas the distribution of our assessment data is more balanced (median 35% correct).

4.3. Natural Variability

The example we are presenting here is the example 1.1 from the Mohler testset [17]. We use here dendrograms to visualize the differences between the natural variability and the Jaro-Winkler distance. The colors are just used to ease the visual separation. The top dendrogram in Figure 5 shows the Jaro-Winkler string distances as presented in section 4.1. We see in the string-distance based version submissions with 5 points (maximal grades) distributed over all clusters.

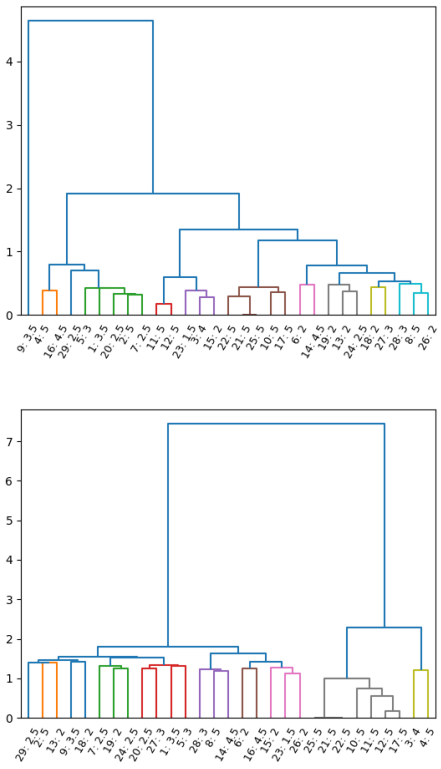

Figure 5. Hierarchical Cluster Analysis of Similarities based on Jaro Winkler Distance vs. Natural Variability

The lower part of Figure 5 displays the dendrogram on the same data based on the natural variability measure. We see a clear cluster on the right containing only top grades.

The main reason for this difference is that these answers are very similar in its formulation but some answers use a slightly different wording. Even more important, these wordings are quite different from other submissions. Below are the answers 10, 11,12, 17, 21, 22, and 25 that form the right cluster for illustration. Aside of the terminating dot, answers 21, 22, and 25 are identical.

10: Simulating the behavior of only a portion of the desired software product.
11: A program that stimulates the behavior of portions of the desired software product.
12: A program that simulates the behavior of portions of the desired software product.
17: Program that simulates the behavior of portions of the desired software product
21: it simulates the behavior of portions of the desired software product
22: It simulates the behavior of portions of the desired software product.
25: it simulates the behavior of portions of the desired software product

Figure 6. Examples of sentences with similar meanings but partly different formulations

The world-level similarities are much better covered by the variability measure. Typically, very similar submissions make a grader suspicious, how this happened. It might be the desired answer (the only correct formulation), or something purely memorized, or a consequence of cheating. While a pure automated grading system would probably just assign a grade to the submission, our approach aims to detect such anomalies - which might have a perfect valid reason, but that is for the grader to decide.

5. Discussion and conclusion

The main contributions of this work are the in-depth analysis of various Short Answer Grading tools and their approaches. All the analyzed grading tools require a desired solution and try to calculate a score based on the distance between the submission and the ”golden essay”. We tried here a radical different approach by questioning the usefulness of the ”golden essay”, especially when the goal is to assess the student’s ability to transfer knowledge. Therefore, we developed an approach which does not require such a ”golden essay”, but where we could obtain comparable quality measures in a language independent approach.

In regards to the first research question factors like answer completeness and natural variability express crucial and important aspects when assessing student answers. Compared to completely manual grading, these metrics assist human graders to highlight basic quality characteristics for easier further evaluation.

Furthermore, our work targets on using assessment tools to improve the transparency of the grading for

human graders. This is very different from the approaches, where machine learning algorithms are used as a black-box producing a grading without explanation. In the last few years various work was published where neural networks predict scores in short text examination settings [32, 33, 34, 35].

Such models have to be trained and have to be fine-tuned [36] to adjust the underlying model weights to each specific context. Machine learning models have biases depending on their training data which are invisible and hard to predict in their consequence, especially when new test items are formulated. The creation of hundreds or even thousands of exemplary reference answers is really strenuous and hard to fulfil by teachers but necessary for reliability and accuracy.

Overall, the trend towards a fully automated grading process with a software should be seen as highly conflicting and morally as well as ethically controversial. At least as long as computer software are not able to grasp, extract and understand rich semantics in all possible variants, a complete replacement of human raters is nearly impossible. Logically, the improvements in newer technology drive the upcoming attempts to fulfil this huge task but until these tools cannot reach up to a sophisticate level of accuracy and reliability, a full replacement of a human should not be considered.

The second main contribution addresses the research question if a software can contribute to a transparent grading of short text answers. The prime intention of the presented approach is to focus on support for human graders. The provided metrics can pinpoint to certain characteristics of the submissions and can guide the grading process of a human grader. These metrics can be used to categorize certain variants of student answers into groups like ”good answer”, ”off-topic answer” or ”copied”. This pre-filtering of answers can be used by the human grader to select those answers which are either completely off-topic or copied from another answer, and enables the human grader to quickly jump to those answers and further investigate them, or to use the metrics to provide certain hints for the teacher to provide more consistent and transparent grading.

5.1. Implications on practitioners in the area of educational grading

According to [15] the latest era is called ”Era of Evaluation”, this is manifested by many evaluation challenges and tasks in various years. Beginning in 2012, the Automated Student Assessment Prize (ASAP) organized by Kaggle has led to competitions where different approaches were tested on the same dataset and evaluation metrics. Another extensively reviewed competition was held during the semantic evaluation workshop in 2013 (SemEval) and was called Joint Student Response Analysis and Eight Recognizing Textual Entailment Challenge [15, 37]. From there on it has become more acknowledged that a publicly available dataset is used for evaluating a possible new approach. Referring to previously presented approaches and tools, these publications were evaluated and tested based on publication-specific data as well as context.

Even with the addition of these two competitions, the whole research field lacks of a world-wide standardized evaluation environment in combination with a dataset. The result is that numerous newly published articles have one aspect in common, the differentiating evaluation mechanisms. While some already use publicly available datasets, others still collect their own data and test their approach on this data. Consequently, the possibility to inter-evaluate different papers is decreasing and highly dependent on the data summarization of the presented paper. On top of that the evaluation metrics differ as well. The Pearson correlation coefficient, the root-mean-square error (RMSE) and the Cohen’s kappa coefficient are just three of common metrics applied for evaluations.

5.2. Limitations and further research

As indicated in Section 4.2, the prediction quality of the answer completeness score concerning grading varies depending on the kind of question. Although the main focus of this paper is not autograding, it would be interesting to investigate deeper into these cases in order to improve on the bad cases, and maybe to develop reliability metrics of these scoring results. In addition, automatically detecting when the introduced metrics should not be used, or maybe some other metrics should be used instead, is another crucial point which needs further research. It is straightforward to combine our system with golden essays, which might provide better results in some cases. But in this paper we wanted to explore first, how far we could get in absence of these.

So far, the acceptance of the supportive charts and graphics to common teachers is not evaluated. One path to tackle this issue is to integrate the grading support into our university assessment system. This would enable us to provide empirical evidence and could help to build a growing set of test data from up to 4000 classes per year. This would allow one to further evaluate it and possibly enhance this system for more diversified types of assessment.

5.3. Main lessons and conclusions

We provided an approach based on new metrics to support human graders in their grading task to increase productivity and transparency. It shows a path to direct the machine's computing power to aspects where a human can work with the newly generated information to assess a student's answer more thoroughly. In the end, the examination process is not made for robots or machines, humans develop those examinations, humans carry out those exams and humans should evaluate if the question was correctly answered or not. The many contextual subtleties, semantically varying perspectives and fine-grained personal opinions make grading such a tedious task for the teachers and professors. Only if students are encouraged to express their knowledge in naturally variable ways, the answers achieve the intended goal, i.e., to consolidate their knowledge for the future.

References

[1] J. Lewis, "Ethical Implementation of an Automated Essay Scoring (AES) System: A Case Study of Student and Instructor Use, Satisfaction, and Perceptions of AES in a Business Law Course," *Satisfaction, and Perceptions of AES In a Business Law Course*, 2013.

[2] S. Basu, C. Jacobs, and L. Vanderwende, "Powergrading: a clustering approach to amplify human effort for short answer grading," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 391–402, 2013.

[3] C. Vojak, S. Kline, B. Cope, S. McCarthey, and M. Kalantzis, "New spaces and old places: An analysis of writing assessment software," *Computers and Composition*, vol. 28, no. 2, pp. 97–111, 2011.

[4] R. H. Nehm and H. Haertig, "Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software," *Journal of Science Education and Technology*, vol. 21, no. 1, pp. 56–73, 2012.

[5] S. Lane, "Performance assessment: The state of the art," *Beyond the bubble test: How performance assessments support 21st century learning*, pp. 131–184, 2015.

[6] R. V. Hogg, J. McKean, and A. T. Craig, *Introduction to mathematical statistics*. Pearson Education, 2005.

[7] E. B. Page, "The imminence of grading essays by computer," *The Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966.

[8] G. K. Chung and H. F. O'Neil Jr, *Methodological Approaches to Online Scoring of Essays.* ERIC, 1997.

[9] L. S. Larkey, "Automatic essay grading using text categorization techniques," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 90–95, 1998.

[10] J. Burstein, C. Leacock, and R. Swartz, *Automated evaluation of essays and short answers*. Loughborough University, 2001.

[11] M. A. Hearst, "The debate on automated essay grading," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 5, pp. 22–37, 2000.

[12] F. White, "Grading the readability of articles," *Canadian Journal of Public Health*, vol. 91, no. 1, p. 73, 2000.

[13] M. Lesterhuis, "When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality," *L1 Educational Studies in Language and Literature*, vol. 18, no. Running Issue, 2018.

[14] C. Beauvais, T. Olive, and J.-M. Passerault, "Why are some texts good and others not? relationship between text quality and management of the writing processes.," *Journal of Educational Psychology*, vol. 103, no. 2, p. 415, 2011.

[15] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60–117, 2015.

[16] S. K. Gaddipati, D. Nair, and P. G. Plöger, "Comparative evaluation of pretrained transfer learning models on automatic short answer grading," 2020.

[17] M. Mohler, R. Bunescu, and R. Mihalcea, "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*, pp. 752–762, 2011.

[18] S. Darus, S. H. Stapa, S. Hussin, and Y. L. Koo, "A survey of computer-based essay marking (CBEM) systems," in *International Conference 'Education & ICT in the New Millennium' at Parkroyal Kuala Lumpur*, vol. 27, p. 28th, 2000.

[19] Y. Yang, C. W. Buckendahl, P. J. Juszkiewicz, and D. S. Bhola, "A review of strategies for validating computer-automated scoring," *Applied Measurement in Education*, vol. 15, no. 4, pp. 391–412, 2002.

[20] Y. Cheung, "Feedback from automated essay evaluation systems: A review of selected research," *TESL Reporter*, vol. 48, no. 2, pp. 1–15, 2016.

[21] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. C. Linn, "Automated scoring of constructed-response science items: Prospects and obstacles," *Educational Measurement: Issues and Practice*, vol. 33, no. 2, pp. 19–28, 2014.

[22] P. F. Ericsson and R. H. Haswell, *Machine scoring of student essays: Truth and consequences*. Utah State University Press, 2006.

[23] C. Scharber, S. Dexter, and E. Riedel, "Students' Experiences with an Automated Essay Scorer.," *Journal of Technology, Learning, and Assessment*, vol. 7, no. 1, p. n1, 2008.

[24] K. Uzun, "Home-Grown Automated Essay Scoring in the Literature Classroom: A Solution for Managing the Crowd?.," *Contemporary Educational Technology*, vol. 9, no. 4, pp. 423–436, 2018.

[25] B. Lewis Sevcikova, "Human versus automated essay scoring: A critical review," *Arab World English Journal (AWEJ) Volume*, vol. 9, 2018.

[26] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[27] H. Schmid, "Probabilistic part-ofispeech tagging using decision trees," in *New methods in language processing*, p. 154, 2013.

[28] H. Schmid, "Improvements in part-of-speech tagging with an application to German," in *Natural language processing using very large corpora*, pp. 13–25, Springer, 1999.

[29] G. Navarro, "A guided tour to approximate string matching," *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.

[30] W. E. Winkler, *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.* ERIC, 1990.

[31] T. K. Landauer and J. Psotka, "Simulating text understanding for educational applications with latent semantic analysis: Introduction to lsa," *Interactive Learning Environments*, vol. 8, no. 2, pp. 73–86, 2000.

[32] S. Kumar, S. Chakrabarti, and S. Roy, "Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading.," in *IJCAI*, pp. 2046–2052, 2017.

[33] T. Liu, W. Ding, Z. Wang, J. Tang, G. Y. Huang, and Z. Liu, "Automatic short answer grading via multiway attention networks," in *International Conference on Artificial Intelligence in Education*, pp. 169–173, Springer, 2019.

[34] H. Qi, Y. Wang, J. Dai, J. Li, and X. Di, "Attention-Based Hybrid Model for Automatic Short Answer Scoring," in *International Conference on Simulation Tools and Techniques*, pp. 385–394, Springer, 2019.

[35] Y. Zhang, R. Shah, and M. Chi, "Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading.," *International Educational Data Mining Society*, 2016.

[36] F. Wild, C. Stahl, G. Stermsek, Y. Penya, and G. Neumann, "Factors influencing effectiveness in automated essay scoring with lsa," in *Proc. of the 12th International Conference on Artificial Intelligence in Education (AIED)*, (Amsterdam, The Netherlands), IOS Press, July 2005.

[37] M. O. Dzikovska, R. D. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. T. Dang, "SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Embodiment Challenge," in *Second Joint Conference on Lexical and Computational Semantics (* SEM): Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, vol. 2, Association for Computational Linguistics, 2013.