Work Smart – Reducing Effort in Short-Answer Grading

Margot Mieskes
Hochschule Darmstadt, h_da
Darmstadt, Germany
margot.mieskes@h_da.de

Ulrike Padó
Hochschule für Technik Stuttgart, HFT
Stuttgart, Germany
ulrike.pado@hft-stuttgart.de

Abstract

In language (and content) instruction, free-text questions are important instruments for gauging student ability. Grading is often done manually, so that frequent testing means high teacher workloads. We propose a new strategy for supporting manual graders: We carefully analyse the performance of automated graders individually and as a grader ensemble and present a procedure to guide manual effort and to estimate the size of the remaining grading error. We evaluate our approach on a range of data sets to demonstrate its robustness.

1 Introduction

Using computers in teaching has opened up new possibilities for learning independent of time or location while receiving individual feedback through frequent testing. For this, automated evaluation of student answers, supported most easily by closed question formats like multiple choice, is key. This means that tests usually do not contain open question types like short answer questions, although these are didactically valuable because they provide insight into students' reasoning.

There is a substantial body of research addressing automated short-answer grading (SAG, see Burrows et al. (2015) for an overview). However, the resulting tools are not widely used to produce completely automated student feedback. Instead, automated methods to reduce manual grading workload have been proposed (which can also be used to reduce annotation workload for training data in general). The use of clustering for label propagation (Basu et al., 2013; Horbach et al., 2014; Zesch et al., 2015; Horbach and Pinkal, 2018) and of Active Learning (Horbach and Palmer, 2016) has been investigated.

In this paper, we describe a new strategy to reduce human graders' workloads. We pre-grade student answers with automated methods that have been carefully analysed to reveal their strengths and weaknesses with regard to the target categories. Combining several automated graders into an ensemble additionally yields insight into the reliability of individual machine grades. Human grading effort can now be focused on reviewing those answers that were most likely not graded correctly.

Effectively, we harness two basic insights of machine learning: Learners perform best on frequently-attested classes (and consequently, under-represented classes require more human attention), and ensembles of learners outperform any given single model (and consequently, automated decisions with high agreement across learners are likely reliable).

Our strategy allows a sizeable reduction of human effort (by at least 40% and up to 93%), while grading accuracy remains at or even improves beyond purely human grading.Since not every student answer is reviewed by a teacher, our approach does not support individual teacher comments on each answer. It is useful in situations where overall performance is being determined by accumulating the grades for individual answers, for example placement tests or recurring text comprehension tests.

Our paper is structured as follows: We first give an overview over related work in manual grader support for SAG (Section 2). We then describe our method and our seven data sets, the machine learning algorithms and features, as well as the evaluation measures in Section 3. In Section 4, we analyse human grading performance in terms of Precision, Recall and Inter-Annotator Agreement to establish a point of comparison. We then investigate the strengths and weaknesses of an automated

Margot Mieskes and Ulrike Padó 2018. Work smart – Reducing effort in short-answer grading. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018).* Linköping Electronic Conference Proceedings 152: 57–68.

grader compared to the human gold standard (Section 5). In Section 6, we assess how much grading effort can be saved and how much grading error remains when we use reliability estimates that are based on the Inter-Annotator Agreement of machine grades only. We summarise our conclusions and point out future work in Section 8.

2 Related Work

Recent work in minimising human annotation effort for short-answer questions has followed two strategies: Clustering similar answers so that each set can be graded together (Basu et al., 2013; Brooks et al., 2014; Horbach et al., 2014; Zesch et al., 2015) or existing grades can be propagated (Horbach and Pinkal, 2018), and selecting the most informative answers for Active Learning (Horbach and Palmer, 2016). Manual workload is reduced either in order to directly benefit teachers (Basu et al., 2013; Horbach et al., 2014; Horbach and Pinkal, 2018) or in order to assist the creation of training data for automatic grading (Zesch et al., 2015; Horbach and Palmer, 2016).

Beyond faster grading, clustering similar answers can also provide interesting insights into common (mis-)perceptions of the subject matter according to Basu et al. (2013), which underscores the didactic usefulness of short-answer questions. In follow-up work, Brooks et al. (2014) demonstrate a speed increase for assigning an initial grade of a factor of three when using clustering support (which corresponds to 66% of time saved). They work on native-speaker content-assessment data, while Horbach et al. (2014) develop a similar approach for language learner data and report a comparable speedup: Using their method could save the correction of 60% of items at 85% grading accuracy. Horbach et al. (2014) acknowledge that perfect scoring accuracy is not necessary in many testing settings; we will investigate human performance levels in Section 4 below. Zesch et al. (2015) aim to reduce the amount of manual annotation required to create training data for automated graders and find that the clustering approach is most useful for very short answer texts.

Horbach and Palmer (2016) perform Active Learning, where instances to be manually labelled are selected to quickly optimise classifier performance. They find that uncertainty-based sample selection is more efficient in improving the classifier than a random and a cluster-based baseline. However, there is great performance variability across the question corpus.

Zesch and Horbach (2018) introduce a clustering and classification workbench intended to facilitate both first practical applications of human grader support and further research.

3 Experimental Setup

Our experiments target ad-hoc tests such as weekly quizzes or end-of-term exams, where question re-use is limited. This sets us apart from approaches that use a corpus of sample answers to prepare grading models for a standardised question pool. Rather, it restricts us to an unseen-question setting in which no training answers are available for any of the questions in the test set. We use various data sets (Section 3.1) to train machine learners (Section 3.2) and evaluate their performance using Precision/Recall and Inter-Annotator Agreement (IAA) (Section 3.3).

3.1 Data

We test the generality of our findings by using a range of standard corpora that vary in size, language and test setting (see Table 1). Our largest corpus is ASAP, although only five out of ten questions have a reference answer and can be used in the unseen question setting. Five corpora are in English, two in German. Half of our corpora are collections of questions generated for low-volume testing (of tens of students at a time) that are graded by the teachers. SEB, Beetle and ASAP (Higgins et al., 2014) are from high-volume, standardised testing and grading situations. Both corpora of language learner data (CREG and CREE) fall into this category; the other corpora test content mastery. For four data sets (ASAP, CREE, CSSAG and Mohler) we have more than one set of human annotations.[1]

We also show the number of grade categories present in each corpus.[2] We generally observe a strong skew towards the majority category in our multi-class corpora. These characteristics of the data will be relevant in Sections 4 and 5 below.

[1] CREG also has multiple annotations, but was constructed to contain only answers with agreeing human annotation.

[2] We use the unseen question, two-way versions of the SEB and Beetle training data.

Corpus	#Questions/ #Answers	# Classes (% max class)	Lan- guage	Task	Human Annotation	Testing Volume
ASAP (www.kaggle.com/c/asap-sas)	5/8182	5 (46%)	EN	Content	Double	high volume
SEB (Dzikovska et al., 2013)	135/4969	2 (60%)			Single	
Beetle (Dzikovska et al., 2013)	47/3941	2 (58%)				
Mohler (Mohler et al., 2011)	81/2273	11 (49%)			Double	low volume
CREE (Meurers et al., 2011a)	61/566	2 (72%)		Language		
CREG (Meurers et al., 2011b)	85/543	2 (50%)	GER			
CSSAG (Padó and Kiefer, 2015)	31/1926	9 (38%)		Content	Double (subset)	

Table 1: Corpus sizes and characteristics

3.2 Automated Graders and Features

We follow the most common literature conceptualisation and treat the prediction of human short-answer grades as a classification task: The human grades are ordinal in nature, which means the order of the categories is defined, but the distance between individual categories is not. We normalise the categories by using the percentages of the maximum score (e.g., 0% and 100% for the two-category corpora). This is useful because questions can have different maximum scores, which means that the impact of absolute points differs across the corpus (2 points could be partially correct for one question, but fully correct for another). Of course this also means that some of the intermediate percentage-based categories will be rare (e.g., 33% will only occur for the subset of 3-point questions if grading is in one-point steps).

As four out of seven corpora have little data, which reduces the possibility to tune parameters, we follow recommendations by Madnani et al. (2016) and employ Random Forest (RF) and Support Vector Machines (SVM), adding Decision Trees (DT) as a third algorithm for their ease of interpretation (all from the Weka machine learning toolkit[3]).

Individual models are trained by leave-one-question-out cross-validation to make the most of our smaller data sets. We experimented with further parameter tuning on the Beetle and SEB data sets, which provide unseen-question dev and test sets. Tuning did improve performance on the test sets, but rarely affected performance in the leave-one-question-out setting. By Occam's razor, we therefore do not further tune parameters. This also applies to modifications of the training regime such as cost-sensitive learning. As a result, our learner performance underestimates the tuned, optimal case.

We use the feature set described by Padó (2016), who selected representative features explored in the literature: N-Gram features (token and lemma-based), text similarity features (with/without stop words), the overlap between student and reference answer in terms of dependency parse and deep semantic representations, and textual entailment (decision and confidence).

3.3 Evaluation Measures

We report weighted **Precision** (P) and **Recall** (R)[4] – on the whole corpus in Section 4, for comparison to human performance; and per predicted category (see Section 5), for a more detailed performance analysis. P and R indicate how reliable a learner's category predictions are and how well they overlap with the actual incidence of that category. Note that weighted overall Recall corresponds to overall **Accuracy**. Overall weighted Recall Rec is computed as in Equation 1.

$$Rec = \frac{\sum_c \frac{TP_c}{TP_c+FN_c} * N_c}{N_{total}} \tag{1}$$

Since $TP_c + FN_c = N_c$,

$$Rec = Acc = \frac{\sum_c TP_c}{N_{total}}. \tag{2}$$

The advantage of Inter-annotator Agreement (IAA) measures such as **Fleiss' κ** ((1971), which is more general than Cohen's κ (1960)) is that they take into account chance agreement by considering the study-specific distribution of annotation categories. Fleiss' κ allows us to compute agreement for individual answers as well as on the question level. κ estimates the annotation reliability in cases where two or more annotators (human or machine) are present. It is computed as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{3}$$

[3] https://www.cs.waikato.ac.nz/ml/weka

[4] We do not report F_1 scores, as they are most useful to compactly compare overall classifier performance, while we are most interested in individual, class-based performance.

where $1 - \bar{P}_e$ denotes the agreement predicted by chance and $\bar{P} - \bar{P}_e$ denotes the agreement actually attained. $\bar{P}$ and $\bar{P}_e$ are calculated as:

$$\bar{P} = \sum_{i=1}^{N} \frac{P_i}{N} \quad (4)$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \quad (5)$$

For each answer that receives a grade, we can calculate the individual agreement P_i as

$$P_i = \frac{\sum_{j=1}^{k} n(n_{ij} - 1)}{m(m-1)} = \frac{\sum_{j=1}^{k} n_{ij}^2 - n_{ij}}{m^2 - m} \quad (6)$$

where m is the number of annotators, n_{ij} is the number of annotators that chose a category j for token i, k the number of categories and N the number of tokens.

We follow Yannakoudakis and Cummins (2015) and do not report correlation measures like Pearson's r and Spearman's ρ, as they are not appropriate for data with many ties (such as grading data sets with their fixed range of categories). Furthermore, r is sensitive to outliers, while ρ inherently measures the ability of a system to rank answers appropriately, as opposed to predicting the correct category. As such, correlation measures do not support our goal of determining how reliable item-wise machine predictions are.

4 Experiment 1: Comparing Human and Machine Performance

Our first experiment investigates human-human performance and compares it to the reliability of automated grading. We compute human P, R and κ for the data sets where informative double manual annotations are available (ASAP, CREE, CSSAG and Mohler). For the human data, we report the performance of the best single annotator against the gold labels and show machine P and R for comparison. We begin with the easiest setting: binary *correct-incorrect* classification for all corpora (where *correct* means $> 50\%$ of the max score). Results are shown in Table 2.

Human P/R results (H P and H R in Table 2) are in the eighties (up to 94 for ASAP) throughout. For the Mohler data, we show two performance numbers: For this data set, the gold standard is created by averaging the two single annotations; therefore, every annotator's grades are highly correlated with the gold. This leads to artificially inflated P and R values (shown in brackets in Table 2). For the other grades, the gold standard was created independently (CSSAG) or one annotator's grades are marked as the gold standard (CREE and ASAP), so that the other's grades are independent of gold. Treating the Mohler data in this way (using annotator "me" as gold annotation) results in performance in the low eighties. We refer to this evaluation method als *Mohler strict evaluation* below.

Our results show that human annotation with up to 16.5% of error (Mohler strict evaluation, 14.2% for CREE) has been accepted in the past for low-volume testing. (Assuming error to be 1-Accuracy, that is 1-R, since weighted R equals Accuracy). For high-volume testing (ASAP), we see much lower rates at 6% error.

Human κ values vary widely across corpora and range from 0.41 (Mohler) to 0.82 (ASAP). The higher κ, the better the human annotators agreed on the grades, producing clearly defined categories and clean training data.

We find low κs for corpora collected as a by-product of low-volume, ad hoc testing with many different questions and different grade categories (CSSAG, CREE and Mohler). ASAP, collected in a high-volume testing setting, is the opposite, since the reliability of multi-annotator grading is a priority when single-annotator grading is impossible due to testing volume. Consequently, there is a small number of grade categories and clear scoring rubrics exist for each question (and graders were likely carefully trained to apply them consistently).

We present machine P and R for the RF learner, our best individual machine grader. It outperforms literature results from Padó (2016) (who used the same features). The SVM and DT learners show similar result patterns as RF, but perform an average of 3 (SVM) and 4 (DT) percentage points worse. Machine results are worse than human results except for the CREE corpus and also outperform the strict Mohler human-human P and R values (see above). Note that both these corpora are strongly skewed towards one class (87% and 72% of the items, respectively). In CSSAG (as well as in SEB and Beetle), the class balance moves to 60-40 and learner performance is noticeably worse. In fact, human P/R results for CSSAG are the strongest among the low-volume corpora, but

Measure		ASAP	CREE	CREG	CSSAG	Mohler	Beetle	SEB
H	P	93.7	86.0	n.a.	89.2	82.7 (95.4)	n.a.	n.a.
	R	93.7	85.8	n.a.	89.9	83.5	n.a.	n.a.
RF	P	86.0	85.4	84.6	71.0	87.5 (93.6)	78.4	70.6
	R	86.2	86.0	84.5	70.4	89.0	78.0	70.7
H	κ	0.82	0.64	n.a.	0.54	0.41	n.a	n.a.

Table 2: Weighted Precision (P) and Recall (R): Human-gold (H) and machine-gold (Random Forest, RF) performance for binary classification. Human-human Fleiss' κ. n.a.: Single human annotation only.

the machine results are the lowest of all four corpora. This may be caused by the low reliability of the human annotations (evidenced by low κ). The strong skew of the Mohler data (49% of data points are annotated with the highest of 11 categories) probably masks a similar effect for that corpus.

5 Experiment 2: Strengths and Weaknesses of Single-Model Grading

Experiment 1 has presented human annotation standards and the performance of a vanilla automated grading model. While the automated grader clearly has room for improvement, our next analyses show that even unreliable machine predictions can considerably reduce human grading effort.

Our goal is to focus the human grading effort on those answers where it is most needed. We accept the consequence that not every student answer will be reviewed by a human grader and that some errors will remain in the final grades. Therefore, the approach is most suitable for testing situations where the grades for individual answers are combined into an overall grade. This accumulated grade is more robust towards some remaining error.

Note that the notion of "most needed human attention" depends on the testing context. In formative feedback situations, it is more acceptable to receive approximate grades than in high-stakes testing, since no decisive consequences depend on formative feedback. We will further discuss these issues below, where we strive to present the trade-off between grading accuracy and grading effort in order to allow users to find the ideal balance for their situation.

In this Section, we take a first step and discuss how to identify reliable machine grades based only on the RF grader's strengths and weaknesses. In Section 6, we will move on to comparing automated predictions from several learners for improved reliability estimates.

	Correct		Incorrect		Majority
Corpus	P	R	P	R	class
ASAP	69.8	66.3	90.6	91.9	I
CREE	89.8	93.1	68.0	57.9	C
CREG	83.3	85.2	85.8	84.1	–
CSSAG	54.3	58.3	79.1	76.3	I
Mohler	89.4	99.2	74.2	16.4	C
Beetle	70.8	76.5	83.4	78.9	I
SEB	70.0	52.7	70.9	83.7	I

Table 3: Weighted P and R per category (binary classification) and majority class (CREG is balanced by design). RF classifier.

5.1 Case 1: Binary classification

Table 3 shows category-wise P and R for binary classification. As can be expected, the majority class is predicted more reliably and with fewer errors in all cases. As CREG is balanced by design, there is no such frequency effect. For the highly imbalanced data sets, R drops steeply in the minority category (between 25%-points for ASAP – 71% incorrect – and 83%-points for Mohler – 87% correct) as the machine grader over-generalises to the majority category.

These results indicate that in a binary setting, manual effort should focus on reviewing the predicted minority class results, as the majority class is fairly reliably marked. For a strongly skewed corpus like ASAP, 1717 instances out of 8182 (21%) need to be reviewed, while for a less skewed corpus such as Beetle, 1708 out of 3942 instances (43%) need to be checked.

Since most corpora are imbalanced, checking only the minority class predictions would save 60-80% of labor while eliminating the largest error source. However, when relying on automatic graders, the information about which answers may have wrongly received the majority class is not available. Additionally, due to the binary setting, no additional information is available to reduce the error further.

In the case of low minority class recall in a high stakes situation, the risk of wrongly-assigned "pass" or "fail" grades is high. This means that all majority class predictions or at least a sample should additionally be checked to catch misassigned minority class answers.

5.2 Case 2: Multi-class classification

We now move on to the more complex multi-class case (where a spectrum of grades is assigned instead of just pass/fail). We have three data sets with more than two target categories: ASAP (five categories), CSSAG (nine) and Mohler (11). We again evaluate RF classification using P/R. We also report category-wise human-human performance for comparison.

Table 4 shows results for the CSSAG data. Clearly, in the harder multi-class case, both human and machine grader performance degrade compared to the results in Table 2. Recall that the human-human data is for a subset of CSSAG – there are additional categories in the whole data set that are not covered in the subset. Human performance on all metrics is best for the categories 0 and 1, with similar κ for 0.5. For categories 0.25 and 0.75, human agreement becomes erratic, with low P/R and κs, which indicates that these categories are not assigned consistently. This implies that these intermediary categories are not well-defined in the annotators' minds, which in turn causes data quality to suffer. Not surprisingly, therefore, the RF P and R show patterns of frequency (rare categories not attested in the human-human subset are predicted badly) and of annotation cleanness. Therefore, predictions of 0, 0.75 (high P) and 1 (which make up 74% of the training data) can be trusted, while the other 25% of predictions should be checked. Additional spot checks of 1 predictions are also advisable due to the lower P in high-stakes settings, while in formative settings, it may not be as important to differentiate between the fine-grained grade steps and over-generalisations to 1 may be acceptable.

Table 5 shows the results for ASAP. Again, both human and machine overall performance drop for the harder task, but with just five categories, the drop is not as steep. Also, human and machine performance is much more robust across all categories. The automated grader performs worst on categories 0.33 and 0.66. Since human performance is stable for these categories, this is probably a frequency issue as the categories are well-defined and clear to the annotators, and the automatic grader is generally reaping the benefits of clean data for the majority class. The machine predictions for category 0, which makes up about half of the gold annotations, can generally be trusted, while predictions of 0.33 and 0.66 (23% of the data) should always be checked.

For the Mohler data set with 11 categories, the drop in performance from the binary classification case is clearest (Table 6, "overall" column; we present the stricter evaluation of one human annotator against the other). Looking at the category-wise results, the Mohler data set, like CSSAG, suffers from both ill-defined and sparsely attested categories. We see low human P/R except for 0 and 1 and low human κ except for 0.4, 0.6 and 1. Additionally, in case of human disagreement the difference between the human grades is often in the range of just one grade; this begs the question whether the difference between the 11 categories can in fact be reliably annotated. The data sparseness stems from the fact that the majority of questions uses only six categories. This results in no machine predictions or very low P/R except for categories 0 and 1.

Together, these two categories make up 50% of the training data. Any other category predictions are likely to be incorrect and should be checked; in a high stakes setting, even predictions for 1 could be additionally reviewed because the relatively low P at high R indicates over-generalisation towards this category.

In sum, when using a single, imperfect machine grader, we can already identify a relatively large set of student answers that is likely graded correctly and probably does not need further human attention. The more target categories there are in the data, the more fine-grained the analysis becomes, but also the reliability of both human and machine grades suffers. Therefore, in a high stakes situation, human graders can be most reliably supported by automated grades for a binary pass/fail decision if the machine grader shows high recall for the minority class. If this is not the case or if the distinction between more grade steps matters, the setup presented here may still be useful for formative feedback since repeated, formative feedback is a large drain of human grader resources and human time saved may outweigh the approximate nature of the grades.

As we use various data sets from a range of scenarios, our conclusions should be generalizable.

		Overall	0	0.17	0.25	0.33	0.5	0.66	0.75	0.83	1
H	P	77.2	79.2	–	62.5	–	57.4	–	39.1	–	93.5
	R	76.2	98.3	–	57.7	–	50.0	–	45.0	–	62.3
RF	P	49.1	63.7	0	18.2	0	40.7	15.8	75.0	0	38.4
	R	67.4	73.6	0	2.0	0	10.0	10.3	2.6	0	71.6
H	κ	0.54	0.65	–	0.27	–	0.59	–	0.36	–	0.68

Table 4: CSSAG: Human-gold (H) and machine-gold (RF) P and R, human-human κ values. – : No prediction made.

	Overall	0	0.33	0.5	0.66	1
H P	87.2	92.4	87.6	80.4	86.0	86.0
H R	87.2	92.4	86.6	81.3	86.0	85.1
RF P	64.7	81.8	39.3	61.8	34.2	47.6
RF R	67.4	89.0	23.9	64.4	22.8	64.5
H κ	0.82	0.88	0.86	0.74	0.85	0.82

Table 5: ASAP: Human-gold (H) and machine-gold (RF) P and R. Human-human κ values.

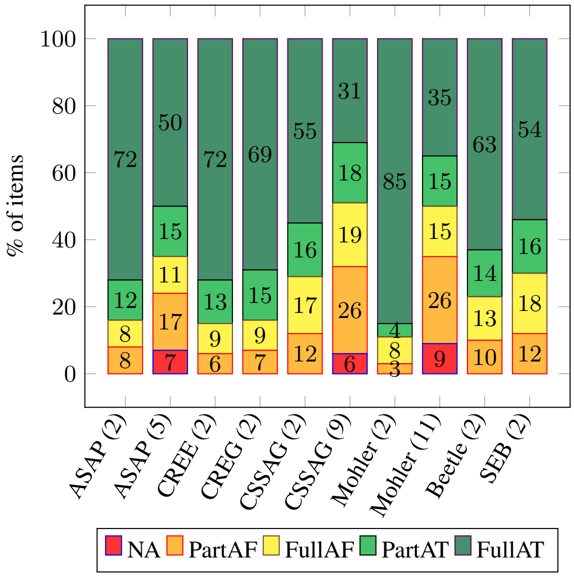

Figure 1: Proportion of items with no agreement (NA), partial agreement on false prediction (PartAF), full agreement on false prediction (FullAF), partial agreement on true prediction (PartAT) and full agreement on true prediction (FullAT), in brackets: # of classes.

6 Experiment 3: Item-wise Reliability of Ensemble Grading

We now switch from using a single automated grader to combining three automated graders (RF, DT and SVM). This approach allows us to generate multiple machine annotations and use them for reliability estimates. We use κ to analyse the automated graders' reliability down to the single-item level and to generate fine-grained reviewing recommendations for manual graders.

We assume that a machine grade is more reliable if more of the graders in our ensemble predict it (and therefore agree better, such that κ is high). With three learners, the item-wise predictions can be in full agreement (FullA, $\kappa = 1$), partial agreement (PartA, $\kappa = 0.83$) or no agreement (NA, $\kappa = 0$).

Figure 1 shows automated grader agreement and disagreement for the binary (and, where applicable) multi-class case. The number of target categories is given in brackets. This figure demonstrates that our assumption of *greater agreement = greater reliability* is generally justified: Compare the proportion of true and false predictions for full agreement and partial agreement. There are vastly fewer cases of FullAF (full agreement, false; yellow) than FullAT (full agreement, true; dark green), while cases of PartAF (partial agreement, false; orange) and PartAT (light green) are closer to balance.

Cases of FullAF (full agreement, false prediction; yellow) are generally around 10%, with a maximum of 19% for CSSAG (9). This is similar to human standards: Human annotators may also agree on a category that does not match the gold standard. In our CSSAG subset, this occurred for 10% of annotations, as well.

Given this picture, our recommendation is to manually review answers in the order of the amount of learner disagreement on the category, beginning with NA. The first lines of Table 7 show that in the binary case, reviewing only NA cases of course means no manual work (three graders have to agree at least partially on two labels, so there is always a clear majority for one label), at error levels between 11% (Mohler (2)) and 30% (SEB). For four out of the seven corpora, error levels would already be close to human agreement error (at $1 - Accuracy$ =15% – recall Section 4). This picture is close to single-grader performance in Section 5.1.

In the more complex multi-class case, this strategy reduces the manual grading effort to between 6% (CSSAG (9)) and 9% (Mohler (11)) of all items in the multi-class case. Assuming that the

		Overall	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
H	P	58.5	96.0	0	37.5	0	30.6	13.6	21.8	3.8	20.8	4.4	78.0
	R	57.3	19.0	4.4	0	0	1.0	13.6	23.8	25.0	27.4	8.7	86.3
RF	P	38.1	67.9	0	0	33.3	13.3	20.6	12.5	37.5	8.8	18.6	56.7
	R	50.9	95.0	0	0	2.9	2.2	5.7	5.8	13.8	2.4	6.7	94.9
H	κ	0.41	0.30	0.01	0.10	0.15	0.62	0.15	0.62	$\kappa < 0$	0.13	0.02	0.82
RF	κ	0.10	0.79	$\kappa < 0$	$\kappa < 0$	0.05	0.02	0.06	0.03	0.15	$\kappa < 0$	0.01	0.15

Table 6: Mohler: Human-gold (H) and machine-gold (RF) P and R. Human-human κ values. n.a.: No prediction made.

		Binary Classification							**Multiclass**		
Strategy	**Measures**	**ASAP**	**CREE**	**CREG**	**CSSAG**	**Mohler**	**Beetle**	**SEB**	**ASAP**	**CSSAG**	**Mohler**
NA only	Effort (%)	0	0	0	0	0	0	0	7	4	9
	Error (%)	16	**15**	16	29	**11**	23	30	28	44	41
PartA/ weak	Effort (%)	12	7	–	12	5	13	12	23	35	50
	Error (%)	**12**	**13**	–	23	**9**	17	25	18	25	**15**
PartA/ all	Effort (%)	20	19	12	27	7	24	28	39	50	59
	Error (%)	**8**	**9**	**9**	17	**8**	**13**	18	**11**	19	**15**

Table 7: Remaining effort (in % of items) and remaining error for all corpora following different review strategies. NA only: Review no-agreement answers; PartA/weak: Revise PartA predictions of classes with weak classifier performance; PartA/all: Revise all PartA predictions. **Bold**: Error **at** or **below** observed human agreement. –: CREG has no majority class.

hand-assigned categories are always correct, remaining error would then range between 28% (ASAP (5)) and 45% (CSSAG (9)). More human effort is clearly needed to further reduce error in most grading situations, even though grader workload has been greatly reduced over the single-grader case (where 25-50% of predictions had to be reviewed) and remaining error also drops for ASAP and Mohler compared to using a single grader.

Finding Errors Figure 1 implies that grades predicted in partial agreement are unreliable between 32% (CREE, CREG) and 43% (CSSAG (2), Mohler (2)) for the binary case and (at best) half of the time for the multiclass case. For comparison, grades predicted in full agreement are unreliable between 4 and 25% in the binary case and between 14 and 38% in the multiclass case. Focusing on PartA predictions is therefore an efficient use of human effort.

We can zoom in further on likely errors by concentrating on the categories that are most likely affected because the machine graders perform weakly on them. For ASAP (5), machine grading performance is known to be worst for classes 0.33, 0.66 and 1 (see RF performance in Table 5). 60% of the erroneous PA predictions are in fact for those classes. Reviewing all PartA cases for these categories, which make up 16% of the total data set, and additionally checking all items where the machine graders disagree (7%) results in a reduction of manual grading effort of 77% of the items, while holding remaining error at 18%. Remaining error can be further reduced to 11% by revising the PartA predictions for *all* classes instead of just the weakest classes. Humans still review only 39% of all answers in this case (corresponding to 61% of effort saved).

Table 7 shows the remaining manual effort and error for all data sets for the PartA/weak strategy (revise cases of NA and those PartA categories that the RF grader is known to perform weakly on) as well as for the PartA/all strategy. For the binary corpora, the predictions for the minority class are reviewed for the PartA/weak strategy.

Clearly, the same patterns hold across all data sets: For binary classification, just using the ensemble predictions reaches error at human levels for CREE and Mohler (recall, however, that these corpora are strongly biased towards the majority class). When reviewing all PartA predictions, five out of seven corpora show remaining error levels below the observed human error level of 15%, and for four of these five, the error is even below 10% at a maximum of 28% of items reviewed. In the harder multiclass case, two of the three corpora show human-level remaining error, but some more reviewing effort is needed (up to 60% of items, or 50% for CSSAG at 20% remaining error). This mirrors the complexity of the task, but is still a sizeable reduction.

Relaxing the Evaluation A second measure that helps save human effort is reconsidering the gravity of machine errors in the multiclass case: In repeated formative testing, a difference between actual and predicted grade of one grade step out of five (or even eleven) may not be of much consequence for the student. To model this relaxed evaluation, we use the definitions from above for FullA, PartA and NA predictions. However, we now count a prediction as correct if it is within one grade step of the gold category. We also apply the relaxed prediction matching to the reviewing recommendations: We now only review NA cases and those PartA cases where the predictions differ by more than one grade step (the majority prediction is accepted for the other PartA cases).

This relaxation is very relevant: 75% of PartA predictions for the ASAP (5) data and 72% for the Mohler (11) data differ within one grade step. For CSSAG (9), however, only four out of more than 800 PartA predictions are within one grade step of another. This pattern of results can be explained by a tendency of the Decision Tree (DT) learner to predict the extreme categories. Since the Mohler and ASAP data are biased towards those categories, all the learners show this pattern and predictions match closely. CSSAG has a bias towards the middle category as well as the top that SVM and RF reflect better than DT. Therefore, they may cast votes that differ strongly from the DT vote.

The results of relaxed evaluation and review are very encouraging for practical application: For ASAP, error drops to 7% when reviewing just 25% of the data. This is at the human level observed for ASAP. Previously, reviewing all PartA predictions in strict evaluation, 11% of error remained and 38% of data were reviewed. For Mohler, 5% error remain after reviewing 38% of data (from 15% of error while reviewing 59% of data). For CSSAG, there is of course no change. This pattern of results makes the approach very promising for formative assessment, where testing is frequent (causing high grader workload) and the individual test result can still be informative even if it is approximate.

7 Implications for Real-World Users

Our motivation for this work was to help human graders in language instruction (and elsewhere) save time and effort on manual grading of free text answers. Our analysis shows that human effort can be reduced drastically by following our reliability-guided review strategy, while grading error stays at or even drops below the human level. However, there are a few points to consider for real-world graders as they choose the correct level of revision strictness for their testing context.

Human revision error We make the assumption that human review always determines the correct category. This may seem optimistic given the human error rate of up to 15% in Section 4, but it is hard to predict grader error more precisely for the general case. Our results in Section 4 show that grading error is lowest in a situation where graders have clear scoring rubrics and are (presumably) carefully trained. Ad hoc grading by teachers shows the highest error rates. For this scenario, it can be hoped that, as the number of answers to review drops, grader alertness and motivation will rise, leading to cleaner annotation. We therefore report the assumed-to-be-perfect numbers and leave it to each grader to discount them by the likely error rates incurred in their process.

Distortion of the grade distribution How will using the proposed method alter the grade distribution? The most conservative case means reviewing all PartA judgements, as this issue is likely to matter most when stakes are high. The only remaining system error (following our assumption about perfect human revisions) are the cases where the ensemble agrees on the wrong category. We analysed CSSAG multiclass, because it has multiple human annotations that are independent of gold (and therefore shows the phenomenon on humans agreeing on non-gold categories, like the machine ensembles). First of all, in both data sets (multiple human-annotated subset of CSSAG and multiple machine-annotated complete CSSAG), the most frequent categories are 0, 1 and 0.5, in this order. Another similarity is that in both cases, mis-assignments end up mostly in the more frequent classes. For the human annotations, mis-assignments are most often labelled 0 and 0.5, for the machine annotation, 0 and 1 (the two most frequent classes). There are, however, some differences: The humans showed a clear tendency to assign the next lower frequent class (mis-labelling 0.5 as 0 or 0.75 as 0.5). They rarely wrongly agreed on the label 1, and true 1s were labelled as 0.75 or even 0.5. Conversely, the machines are overly generous: They tend to mis-label as 1, even though 0 is the most frequent class in the data (almost twice as frequent as 1).The reason may be that many true 0 answers are simply empty or contain just a few (non-informative) words. The machines therefore tend to over-generalise to the next more frequent category, 1, if an answer does not fit that pattern, even if it is still incorrect. There is an indication of this behaviour in the learner performance in Table 4: Precision for class 1 is much lower than for class 0, at similar recall. This indicates that class 1 predictions are more often unreliable, and the machine is therefore being overly generous. Note, however, that this distortion affects only 11% of the data.

This error analysis suggests that we might improve the automated grader by training it only on non-empty incorrect answers, thereby removing the bias towards distinguishing between empty and non-empty rather than correct and incorrect. The empty answers would then be trivially labelled as incorrect after filtering.

In general, since items to review are chosen on the basis of classifier grades, if the ensemble shares a bias towards a specific category, the grades of that category will be reviewed disproportionally rarely. Fortunately, learner bias follows the frequency biases in the data, so that the bias categories are generally graded reliably, as mirrored in the remaining error levels for our strategy. However, classifier bias and it's tendency may still be relevant depending on the testing context, as we have seen.

Language lerner corpora In our experiments, language learner corpora generally fare better than content assessment corpora. The CREE and CREG corpora feature binary categories which were annotated reliably ($\kappa = 0.64$ for CREE, only items with annotator agreement for CREG). While a κ of 0.64 might not sound impressive, it gives an idea how hard the task of awarding the equivalent of "pass" or "fail" in these contexts are. This is the best-case scenario for the automated learners. In this situation, corpus size does not seem to matter as much as might be expected: The RF learner performs as well for CREE and CREG as it does for the binarised version of ASAP, which is roughly 16 times larger.

An additional factor in this scenario may be that CREE and CREG appear to be easier to machine-grade than many content-assessment corpora (Padó, 2016). Padó (2017) hypothesizes that this effect is due in part to the fact that the majority of questions in language learner corpora are text comprehension questions, which require reproduction and tend to produce answers that are very close to the reading text as well as the reference answer taken from this text. Additionally, language learners' limited proficiency may keep them from paraphrasing freely, which compounds the effect.

In sum, the language learner corpora we used for our experiments are very well suited to train reliable automated graders, and this is mirrored in the evaluation results: The machine ensemble predictions (in full and partial agreement) were correct at about the level of human performance (85% of labels correct) without any human review. This can optimally be raised to 91% correct categories when reviewing just 20% of the data. The remaining error (where the ensemble fully agrees on the wrong category) is at 9% of the data, which is the same level as human agreement on the wrong category for a subset of CSSAG.

This makes our strategy especially promising for free text grading in language instruction: On the one hand, the question type is frequent and grading therefore adds substantially to the teachers' workloads; on the other hand, machine-supported manual grading yields grades that are definitely reliable enough for formative testing and possibly even summative testing (after the reservations about possible distortions in grade distribution are considered).

Lessons for ad-hoc manual grading Comparing the CSSAG, Mohler and ASAP data set, we also observe that the amount of training data available per category and the quality of human annotation clearly matters. Although the categories are imbalanced in all data sets, the absolute number of examples for each of the five categories in the ASAP data is considerably higher than for Mohler and CSSAG, where some categories are very sparse. Also, the human κs are much higher and more consistent for the ASAP data. Consequently, the machine graders learn to make high-quality predictions. In light of this observation, we recommend using as few categories as possible in automatically supported grading to avoid sparseness issues and to prioritise clear category definitions (resulting in high human grader agreement). As the analysis of the manual grading shows, too many categories lead to unclear representations in humans as well, which in turn do not allow for clear models using machine learning. Fewer categories are also easier to interpret, as the differences of individual steps on a 10-point scale are less clear than differences for example on a 6-point scale.

In some NLP tasks, reducing the scale is not an option. Therefore, information about easily modelled categories could be used to focus the annotation effort on the harder categories to ensure consistent models.

8 Conclusions and Future Work

We have shown the practical usefulness of unoptimised automatic grading tools for reducing human grading effort, based on seven different corpora and two different grading scenarios. For the binary grading scenario, where only correct vs. incorrect is distinguished, effort can be reduced by at least 75%. In more complex grading scenarios of assigning grades based on various levels of granularity, effort can be reduced by at least 40% – depending on the scale complexity. This reduction in effort retains an acceptable error rate, which is comparable to or even below human error rate. In the literature reviewed in Section 2, a reduction in grading effort of of 60% is possible while the human error levels.

Although our suggested strategy involves various evaluation steps, it is nevertheless technically simple to use for the human grader: Individual automatic graders have to be analysed with respect to their performance using Precision and Recall in order to determine their biases. The automatic grader predictions are then compared using Inter-Annotator Agreement, which gives a detailed picture of the grading quality of individual categories and items. This enables the human grader to focus the correction effort on the most important cases, ignoring automatic annotations that are most likely correct.

The cases to be revised can be chosen according to available grading time and required level of remaining error: First, only items that did not receive an automatic grade have to be corrected. In the binary case this even means no manual effort at all, but this strategy also leaves the highest error rate. Second, only items where the automatic graders are not unanimous and predict weakly performing categories are manually checked. This results in an error rate of 9-25% while reviewing 13% of the data for the binary scenario and 25-50% of the data for the multiclass scenario. The most detailed strategy involves reviewing all items that did not receive an unanimous vote. This results in an error rate at or below human level, while reviewing 7-28% of the data for the binary scenario and 40-60% of the data for the multiclass scenario. Further reductions of effort and error are possible if evaluation is slightly relaxed in the multiclass case.

Our results match insights from the general machine learning domain: a) Grader performance correlates to the number of training instances for a category. b) By using three flawed automated graders, we make use of the power of error independence in the machine ensemble (Kuncheva, 2004).

Finally, based on our analysis we can give recommendations for Computer-Aided Language Learning (CALL), especially regarding the development of corpora, which serve as the basis of many approaches. In order to optimise machine grading performance, first, the grading scale should not be too fine-grained, as rarely occurring categories are problematic even for humans. Second, the grading categories should be clearly defined. But thirdly, even relatively small-sized corpora are sufficient to create good models for automatic pre-grading if the first two points are true.

Our strategy seems especially promising for short answer grading in language instruction. On the one hand, the question type is frequent and grading therefore adds substantially to the teachers' workloads; on the other hand, machine-supported manual grading in our experiments yields grades that are definitely reliable enough for formative testing at a fraction of the manual effort.

Given the encouraging results of the present study, a logical future step to extend this work would be a user study with real-world human graders, since our work so far has been carried out only on existing corpora.

References

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics (TACL)*, 1:391–402.

Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy

Vanderwende. 2014. Divide and correct: Using clusters to grade short answers at scale. In *Proceedings of L@S '14,* Atlanta, Georgia, 4–5 March 2014, pages 89–98.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

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 *Proceedings of SemEeval-2013,* Atlanta, Georgia, 14–15 June 2013, pages 263–274.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Derrick Higgins, Chris Brew, Michael Heilman, Ramon Ziai, Lei Chen, Aoife Cahill, Michael Flor, Nitin Madnani, Joel R. Tetreault, Daniel Blanchard, Diane Napolitano, Chong Min Lee, and John Blackmore. 2014. Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *Computing Research Repository, Computation and Language*.

Andrea Horbach and Alexis Palmer. 2016. Investigating active learning for short-answer scoring. In *Proceedings of BEA-11,* San Diego, California, 16 June 2016, pages 301–311.

Andrea Horbach, Alexis Palmer, and Magdalena Wolska. 2014. Finding a tradeoff between accuracy and rater's workload in grading clustered short answers. In *Proceedings of LREC 2014,* Reykjavik, Iceland, 26–31 May 2014, pages 588–595.

Andrea Horbach and Manfred Pinkal. 2018. Semi-Supervised Clustering for Short Answer Scoring. In *Proceedings of LREC 2018,* Miyazaki, Japan, 7–12 May 2018.

Ludmila I. Kuncheva. 2004. *Combining Pattern Classifiers – Methods and Algorithms*. Wiley, Hoboken, NJ.

Nitin Madnani, Anastassia Loukina, and Aoife Cahill. 2016. A large scale quantitative exploration of modeling strategies for content scoring. In *Proceedings of BEA-12,* Copenhagen, Denmark, 8 September 2017, pages 457–467.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011a. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *Special Issue on Free-text Automatic Evaluation. International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL)*, 21(4):355–369.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011b. 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.

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 ACL-HLT 2011,* Portland, Oregon 19–24 June 2011, pages 752–762.

Ulrike Padó. 2016. Get semantic with me! The usefulness of different feature types for short-answer grading. In *Proceedings of COLING 2016,* Osaka, Japan, 13–16 December 2016.

Ulrike Padó. 2017. Question difficulty – How to estimate without norming, how to use for automated grading. In *Proceedings of BEA-12,* Copenhagen, Denmark, 8 September 2017.

Ulrike Padó and Cornelia Kiefer. 2015. Short answer grading: When sorting helps and when it doesn't. In *4th NLP4CALL Workshop at Nodalida*, pages 42–50, Vilnius, Lithuania.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of BEA-10,* Denver, Colorado, 4 June 2015, pages 213–223.

Torsten Zesch, Michael Heilmann, and Aoife Cahill. 2015. Reducing annotation efforts in supervised short answer scoring. In *Proceedings of BEA-10,* Denver, Colorado, 4 June 2015, pages 124–132.

Torsten Zesch and Andrea Horbach. 2018. ESCRITO - An NLP-Enhanced Educational Scoring Toolkit. In *Proceedings of LREC 2018,* Miyazaki, Japan, 7–12 May 2018.