Fine-grained essay scoring of a complex writing task for native speakers

Andrea Horbach[1], Dirk Scholten-Akoun[2], Yuning Ding[1], Torsten Zesch[1]
[1] Language Technology Lab, Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Germany
[2] Center of Teachers' Education, University of Duisburg-Essen,
{andrea.horbach|dirk.scholten|torsten.zesch}@uni-due.de
yuning.ding@stud.uni-due.de

Abstract

Automatic essay scoring is nowadays successfully used even in high-stakes tests, but this is mainly limited to holistic scoring of learner essays. We present a new dataset of essays written by highly proficient German native speakers that is scored using a fine-grained rubric with the goal to provide detailed feedback. Our experiments with two state-of-the-art scoring systems (a neural and a SVM-based one) show a large drop in performance compared to existing datasets. This demonstrates the need for such datasets that allow to guide research on more elaborate essay scoring methods.

1 Introduction

Automatic essay scoring is the task of automatically rating free-form writings. The scores assigned are often holistic and are based both on content and form. Automatic essay scoring is nowadays successfully used to reduce human scoring workload (Dikli, 2006), for example for the assessment of language proficiency (Weigle, 2013). Automatically assigned scores are considered reliable enough that they have replaced one out of two human annotators even in high-stakes language proficiency tests such as TOEFL for many years now (Attali and Burstein, 2006).

Essay scoring approaches in recent years have mainly focused on a small number of publicly available datasets, especially the ASAP dataset from the Kaggle competition. On this dataset, many approaches reach very competitive results, comparable to human scoring performance (Shermis and Hamner, 2012), so that the impression might arise that automatic essay scoring is a solved problem.

In this paper, we present experiments on a new dataset that we consider to be more challenging than currently available ones. We score essays written by prospective teachers, before starting their university education in Germany. These essays in German language are collected to assess whether these native-speaking students might need additional language training in order to become a teacher. While other datasets either measure the full range of language proficiency from novice learners to (near-)natives, or measure the writings of high-school students, our dataset shows much less variety in language proficiency. As almost all test-takers are native speakers and possess a general qualification for university entrance, differences between good and a not so good essays are much less pronounced.

When applying state-of-the-art essay scoring systems on this dataset, we find that a feature set working well on a standard dataset shows a considerably worse performance on our data. This makes it very questionable whether automatic scoring techniques could currently be applied in a real-life scenario, thus confirming the need for deeper methods able to handle such datasets.

We first present an overview of related work, especially publicly available datasets and present our corpus in detail. We then asses the scorability of the corpus by a series of experiments using a supervised machine learning system with a standard feature set. We first confirm that our system reaches state-of-the-art performance by evaluating it on the ASAP corpus and scores in our corpus assessing the writing globally. Subsequently, we assess how well such a feature set is suited to model the different scoring variables annotated in our data and find that the global scores are modeled best. Concentrating on these scores, we investigate the influence of various feature settings and different amounts of training data on the scor-

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 357–366
Copenhagen, Denmark, September 8, 2017.

ing performance.

2 Related Work

Automatic essay scoring is almost always tackled as a machine learning task (Dikli, 2006; Valenti et al., 2003). A wide range of features representing different aspects contributing to a good essay have been proposed such, as n-grams (Chen and He, 2013) or LSA (Foltz et al., 1999), length (Mahana et al., 2012; Östling, 2013), linguistic correctness in terms of spelling and grammar (Mahana et al., 2012; Östling, 2013), or cohesion and coherence of a text through identifying overlap between sentences and usage of connective devices (Lei et al., 2014). Recently, also neural methods have been proposed and successfully used for essay scoring (Taghipour and Ng, 2016).

Most essay scoring approaches in recent years have been evaluated either on proprietary datasets or on a few publicly available ones. Not publicly available data include datasets used by Klebanov et al. (2016) with large amounts of college level exam data, or data from music teacher proficiency test (Madnani et al., 2016). Responses in this last dataset are in length on the borderline between short answers and essays and are interesting because they, as well as our corpus, target writings by generally language-proficient population. The dominating publicly available dataset for essay scoring in recent year has been the data of the ASAP essay scoring challenge.[1] It contains both source-based and opinion tasks targeting US students from grade 7 to 10 for 8 different prompts with up to 3000 responses per prompt. Since its release in 2012, the dataset has been widely used in a number of approaches (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Cummins et al., 2016). Another dataset, the CLC-FCE corpus (Yannakoudakis et al., 2011) contains essays written by ESOL test takers, but relatively little data per individual prompt (1,244 essays across 10 prompts), making it not the first choice for prompt-specific approaches. Because of its extensive error annotations, it has also been used for the task of grammatical error detection and correction (e.g. Cahill et al. (2013) and Seo et al. (2012)).

In Swedish, a corpus of high school essays has been released by Östling (2013) with an overall number of 1,702 essay for 19 different prompts. This means, also this dataset contains few essays per prompt, such that their automatic scoring mainly focuses on form (which can be assessed across prompts) rather than content (which is to a higher degree prompt-specific).

Some other corpora were originally not designed for the task of essay scoring, but each sample comes with a language proficiency level of its writer, therefore allowing to use them for language proficiency assessment. That means their labels do not necessarily reflect the proficiency of the current essay, but rather the general language proficiency of the writer. For example, the ETS corpus of non-native written English (Blanchard et al., 2013) contains 12,100 TOEFL test essays and has originally been published for the task of native-language identification (Tetreault et al., 2013), but also comes with coarse proficiency levels and has been used for the task of proficiency classification (Klebanov et al., 2016; Vajjala, 2017). Similarly the ICNALE corpus (Ishikawa, 2011) contains English essays from Asian writers where each essay has an assigned proficiency level. Beyond the English language, proficiency classification has been performed on the Swedish SweLL corpus (Volodina et al., 2016; Pilán et al., 2016), and for Estonian (Vajjala and Lėo, 2014).

3 A more Challenging Essay Dataset

As described above, most datasets are either small or target a wide range of proficiency levels, so that relatively shallow features are sufficient to achieve quite good performance. To overcome this problem, we have created a new dataset from essays written in German by prospective university students, mostly native speakers. The essays are one part of a large-scale assessment project at the University of Duisburg-Essen, SkaLa (Bremerich-Vos and Scholten-Akoun, 2016). All students who intend to enroll in a degree program for future teachers have to participate in a compulsory language assessment. A major constituent of this assessment is an open writing task with two parts. First, students are asked to summarize a newspaper article dealing with an education-related topic (which we call the *source text*), in our datasets, an article about the pros and cons of study fees. This part of the response is the *summary* part of the essay. Second, the students shall briefly discuss a particular

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

statement from the prompt (the *discussion* part). The time limit for this task is 120 minutes and the produced text is supposed to consist of at least 350 words.

The aim of the fine-grained evaluation is to identify the participants' strengths and weaknesses as precisely as possible. After a manual evaluation of the essays, students receive detailed feedback about their performance in each of the manually scored variables and –if need be– are informed about relevant available training programs at the university designed to foster written language competencies.

In this way, 2,020 essays with an average of around 600 tokens per essay were collected and scored as described next.

Scoring Rubric While many essay-scoring corpora provide only a holistic score, this dataset has been scored using a fine-grained rubric, targeting different aspects of writing.

The raters were asked to evaluate the students' texts with regard to a total of 41 variables. The writing skills ratings are based upon analytical descriptors (cf. Weigle (2002, p. 114) and Weir (2005, p. 183)). Table 1 provides an overview of the annotated variables. 11 variables measured content-related aspects, i.e. whether a certain argument regarding the topic of the source text is mentioned in the essay, 3 formal, 5 structural and 10 measured linguistic aspects. In addition, there are 6 dimension variables and one overall variable.

Before the annotators scored the texts according to the fine-grained rubric, they evaluated the texts in a subjective-holistic overall rating (*G1 – written language competence*). This evaluation was always carried out immediately after the first reading of the text, hence before the extensive analytical evaluation.

The rating scheme includes three types of variables: a) *descriptors* are variables for the evaluation of specific individual aspects of an essay (e.g. whether a certain argument from the source text is covered in the summary, whether the central thesis is correctly identified or whether grammar is proficiently used). The descriptors are directly annotated. b) *Dimension ratings (G2–G7)* are weighted aggregations of individual descriptors, i.e. they are not annotated but computed based on the descriptor annotations (e.g. G4–Discussion is an aggregation of the descriptors for the discussion part D1 to D4). c) Finally, a *superordinate rating* (*G8* – informed overall judgment) emerges from the weighted aggregation of the dimension ratings and therefore relates to the entire text in all of its aspects covered by the rating scheme. (Annotators were allowed to change the aggregated G8 score, if they felt it did not adequately represent the essay.)

The essays were annotated by one out of 6 annotators each. The annotators received extensive training on a subset of randomly selected 120 essays. After training, annotators reached an inter-annotator agreement between 52 and 100% ModAgree (cf. Harsch and Martin (2012, p. 228-250) and Harsch and Martin (2013)) for the different descriptor variables. Percentage ModAgree is computed by measuring per essay and variable what percentage of all ratings assigned by the different annotators for this essay agrees with the mode, i.e. the value assigned most often. These values are then aggregated across all essays. For the subjective-holistic G1 score, annotators reached 59% ModAgree, for the aggregated G8 score 60% ModAgree was reached . Note that higher agreement values for the descriptor variables are partially due to fewer categories available for annotation.

In very few cases (up to four essays per variable), it was not possible for the annotator to encode a certain variable for a category (e.g. discussion variables could not be annotated if the discussion part was missing). We do not represent those essays in the label distributions and exclude them from the training and test data when performing machine learning experiments for that variable.

4 Experimental Setup

We split the data randomly into 90% training data and reserve 10%, i.e. 202 essays, as held-out test set. If not reported otherwise, our results are based on ten-fold cross-validation on the training section. In accordance with previous work, we evaluate using quadratically weighted kappa (QWK). For a more intuitive interpretation of the results, we also report accuracy. All data is preprocessed using a DKPro pipeline (Eckart de Castilho and Gurevych, 2014) consisting of segmentation, POS-tagging (both OpenNLP[2]), lemmatization using the MateLemmatizer (Anders et al., 2010) and parsing using the StanfordParser (Rafferty and Manning, 2008).

[2] https://opennlp.apache.org

	Score	Description	Range	% Mod-Agree	Relevant Essay Part	Distribution
	G1	Written language competence based on first impression	1/2/3/4/5/6	59	Both	
Form	F1	Appropriateness: Does the text address the task?	1/2/3/4	100	Both	
	F2	Plagiarism: Does the text copy the prompt?	1/2	100	Both	
	F3	Running text vs. bullet points	1/2	100	Both	
Content	C1	Central question	1/2/3	86	Summary	
	C2	Central thesis	1/2/3	79	Summary	
	C3	Political background	1/2/3	85	Summary	
	C4	Effect of study fees	1/2/3	90	Summary	
	C5	Securing academic educ. financially	1/2/3	92	Summary	
	C6	Beneficiaries	1/2/3	93	Summary	
	C7	Primary education (finances)	1/2/3	89	Summary	
	C8	Primary education (career)	1/2/3	85	Summary	
	C9	Academics vs. non aademics	1/2/3	91	Summary	
	C10	Overtaxing the poor	1/2/3	86	Summary	
	C11	Paying later	1/2/3	88	Summary	
	G2	Coherence – overall score	1/2/3/4/5/6	65	Summary	
	G3	Summary –aggregated score	1/2/3/4/5/6	74	Summary	
Discussion	D1	Are there own contributions (aspects not mentioned in source)?	1/2	92	Discussion	
	D2	Is an own point of view present and is it motivated?	1/2/3	75	Discussion	
	D3	Quality of argumentation	1/2/3	85	Discussion	
	D4	Rigor of discussion	1/2/3/4/5/6	59	Discussion	
	G4	Discussion – aggregated score	1/2/3/4/5/6	55	Discussion	
Structure	S1	Introduction present and marked?	1/2/3	94	Both	
	S2	Summary present and marked?	1/2/3	79	Both	
	S3	Discussion present and marked?	1/2/3	88	Both	
	S4	Conclusion present and marked?	1/2/3	78	Both	
	S5	Formatting	1/2/3/4	68	Both	
	G5	Structure – aggregated score	1/2/3/4/5/6	67	Both	
Language	L1	Spelling: no / up to 5 / 6 to 10 /more	1/2/3/4	76	Both	
	L2	Typos: no / up to 5 / more	1/2/3	86	Both	
	L3	Grammar: no / up to 5 / more	1/2/3	79	Both	
	L4	Punctuation errors: no / up to 5 / more	1/2/3	80	Both	
	L5	Word usage (correctness)	1/2/3	68	Both	
	L6	Word usage (variance)	1/2	91	Both	
	L7	Is there conceptually oral language?	1/2/3	76	Both	
	L8	Sentence structure (variability)	1/2	98	Both	
	L9	Citations (formal aspects): use of quotation marks, references…	1/2	75	Summary	
	L10	Citations (content): Are direct citations used for central points?	1/2	86	Summary	
	G6	Stilistic skills – aggregated score	1/2/3/4/5/6	52	Both	
	G7	Verbal skills – aggregated score	1/2/3/4/5/6	64	Both	
	G8	Overall Impression, aggregated from G2 to G8	1/2/3/4/5/6	60	Both	

Table 1: Scoring categories in our corpus. Note that a lower score corresponds to a better essay.

For our experiments, we rely on two state-of-the-art systems: A classical supervised system based on hand-crafted features, and an SVM classifier, and an LSTM neural model based on embeddings.

4.1 SVM Classifier

We use Weka's (Hall et al., 2009) Support Vector classifier (SMO) in standard configuration as provided through DKPro TC (Daxenberger et al., 2014). We utilize a number of state-of-the-art features: As the essays in our dataset were written within a certain time limit, the **length** of an essay is an indicator of its quality. We measure length by the number of sentences, tokens and characters per essay. Additionally, we measure average sentence length in tokens and average token length in characters. **N-gram** features model words and phrases or constructions – in the case of POS n-grams – in an essay. We use boolean occurrence features for token and POS uni- to trigrams and token skip bi- to 5-grams. We count the **occurrence** of linguistic features, such as certain punctuations (commas, exclamation marks, quotation marks) as well as formal references to the source text and occurrences of reported speech.

Another set of features is based on **syntax**. We count the number of subordinate clauses in general, as well as the number of temporal and causal subordinate clauses using lists of indicator words for the latter two. We model syntactic variability through the average and maximal depth of parse trees in an essay and the distribution of individual POS tags. We also cover the linguistic variance in an essay through type-token-ratio. Also **language errors** are usually considered informative. We use the rule-based LanguageTool[3] checker to identify the number of spelling mistakes, punctuation errors and other grammatical errors. The number of **cohesive devices**, i.e. connectives, normalized by the essay length in tokens. Additionally, the average similarity between adjacent sentences measured both through greedy string tiling and the number of shared nouns between two sentences represents **coherence**.

4.2 Neural System

As the neural system, we use the Neural Essays Assessor (NEA) (Taghipour and Ng, 2016)[4], a LSTM architecture using a mean-over-time layer for aggregation in its reported best configuration exchanging the English word embeddings for German polyglot embeddings (Al-Rfou et al., 2013) and using 50 LSTM units for run-time efficiency. We also perform 10-fold cross-validation using 8 folds for training and 1 fold as development set to determine which of 50 epochs to use per run.

[3] https://languagetool.org/de/

[4] https://github.com/nusnlp/nea

5 Experiments & Results

This section presents our experimental results. We first evaluate state-of-the-art systems on the two global variables G1 and G8 in our data and compare to the performance on ASAP. We then investigate the performance on all variables to measure to what factors our model is sensitive. In subsequent experiments we address the influence of using the essay's summary and discussion part separately, of individual feature groups and the size of training data on the scoring performance.

5.1 Experiment 1: Performance of State-of-the-Art Systems

In our first experiment, we assess the overall performance of the scoring system on the two global variables G1 (holistic) and G8 (aggregated) under different feature settings. We apply the neural system and for the SVM-based system we test several conditions: As n-grams are known to be strong features (Yannakoudakis et al., 2011), we evaluate a baseline taking only token n-grams into consideration. We use two versions of the feature, one where we consider the top 1,000 most frequent n-grams (*n-gram 1,000*) and one where we consider the top 10,000 n-grams (*n-gram 10,000*). We next evaluate the full system with and without stacking of the three groups of n-gram features individually in order to avoid that these feature groups might overpower the other features.

In Table 2, we report the performance of the different setups for the variables G1 and G8. We see that we always reach a higher performance when predicting the informed overall score G8 than the holistic G1. Remember that G1 is assigned before scoring the other essay variables while G8 is a score based on the other variables. It seems plausible that G8 is more consistent and easier to predict automatically, although we do not find that reflected in agreement scores between human annotators.

We further observe that the performances of

Paradigm	Configuration	Our corpus G1 (holistic) acc.	Our corpus G1 (holistic) QWK	Our corpus G8 (aggregated) acc.	Our corpus G8 (aggregated) QWK	ASAP acc.	ASAP QWK
Neural	NEA (Taghipour and Ng, 2016)	.43	**.45**	**.59**	**.53**	n/a	**.76**
SVM	n-grams – top 1000	.36	.36	.47	.40	.44	.64
	n-grams – top 10000	.45	.44	.56	.48	.49	.67
	full + n-grams top 1000	.40	.39	.54	.45	.46	.66
	full + n-grams top 10000	.45	**.45**	.57	.48	.49	.68
	full + stacked n-grams 1000	.47	.39	.58	.47	.53	.72
	full + stacked n-grams 10000	**.48**	.42	**.59**	.46	**.54**	.72

Table 2: Scoring performance for G1 (intuitive holistic) and G8 (aggregated holistic). For comparison, we also provide the performance of a comparable English model for the ASAP dataset (averaged over all 8 prompts).

both variables benefit from a larger number of n-grams. However, additional features in the full model are only beneficial if we have lower numbers of n-grams. These findings suggest that there is some redundancy between the n-grams and the remaining features.

We observe that using the out-of-the-box neural system is at least on par with the best supervised configuration. While this shows the potential of neural approaches on the global variables, in the following experiments, we concentrate on the SVM system that can be more easily targeted towards the individual variables.

Verification Using ASAP Data For comparison, we also evaluate our feature set (with minor adaptations from German to English) on the ASAP corpus, a well-known dataset for essay scoring (see Section 2). As labeled test data is not available, we evaluate using 5-fold cross validation on the training data – Table 2, the two rightmost columns. For the neural system, we report results by Taghipour and Ng (2016). Both the neural system .76 QWK and the SVM .72 QWK are on par with the best open-source system participating in the ASAP shared task that reached .71 QWK.[5]

These results show that the applied systems are state of the art on established datasets and are thus probably also state of the art on our new dataset. However, the performance level is much lower, as the task is more challenging.

5.2 Experiment 2: Scoring Performance for Different Variables

Next, we want to assess how well our essay scoring system is able to predict the different variables. We repeat Experiment 1 in the best-performing feature setting using the full model with 10,000 n-grams for each scoring variable separately, i.e. we use always the same features to train different models.

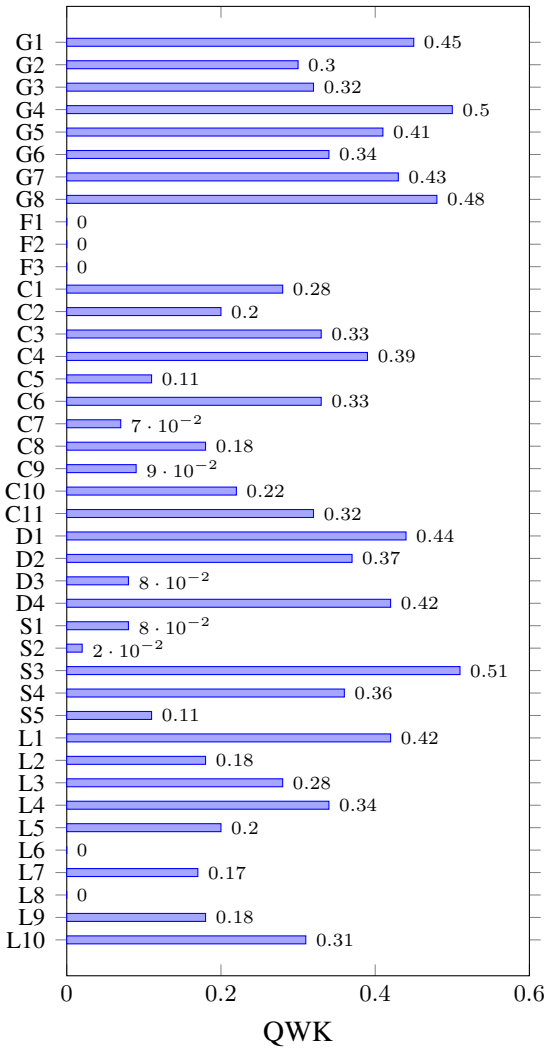

Figure 1: Scoring performance in quadratically weighted kappa for models trained separately using the same features on each scoring variable.

[5]Results are not directly comparable, as the official test data from the challenge is not publicly available.

feature set	G1 (holistic)	G8 (aggregated)
merged	**.449**	.481
split	.441	**.522**

Table 3: Scoring performance measured in quadratically weighted kappa for G1 and G8 with features computed on the complete essay text (*merged*) and with features computed on the summary and discussion part separately (*split*).

It is clear that our one-fits-all approach can be improved by using feature sets tailored towards the individual variables. With this experiment we rather want to investigate which variables are sensitive to our model which uses features used for predicting global scores. This could help to answer the question which aspects of a global score an essay scoring system actually measures. Figure 1 shows the results.

We see that the feature set predicts at a very moderate level for many of the variables. For the very skewed variables F1 to F3, L6 and L8, performance is particularly bad. We also see that variables from each of the four categories (content discussion, structure, and language) can be learnt to a very limited degree. Interestingly, the model performs a bit better on the aggregated scores G2 to G7 and the two global variables G1 and G8 although still on a level that prohibits a practical use of the system. In the following, we concentrate on G1 an G8. In doing so, we can also compare to scores in other essay datasets that use only holistic scores.

5.3 Experiment 3: Splitting Essays into Summary and Discussion Part

The prompt in our task asks for essays with a specific structure: a summarization and a discussion part. Some of the variables are measured on only one of those two parts (cf. Table 1). Therefore it seems reasonable to measure not only if a feature occurs, but also in which part of the essay. For example, the trigram *in my opinion* might be an indicator for a good essay if it occurs in the discussion, but not in the summary. Therefore, we also determine n-gram features (token, pos, and skip) separately for both essay parts.[6] In the *split* condition in Table 3, we duplicate each n-gram feature and compute it individually on the summary and the discussion part, the *merged* condition repeats values from Table 2 for the best-performing SVM, the full model with 10,000 n-grams.

We see that for G8 we profit from that split, while for G1 we do not. We do not have a good intuition why this is the case, but suspect that the more informed G8 score takes this additional information better into account. What we learn from this experiment is that it is helpful to take additional prompt-specific structure in the data into consideration. Identifying further automatically detectable sub-parts of the essay and treating them separately is a promising step for future work.

[6]There are essays where only one part was present. In such cases all features for the other part have been set to 0.

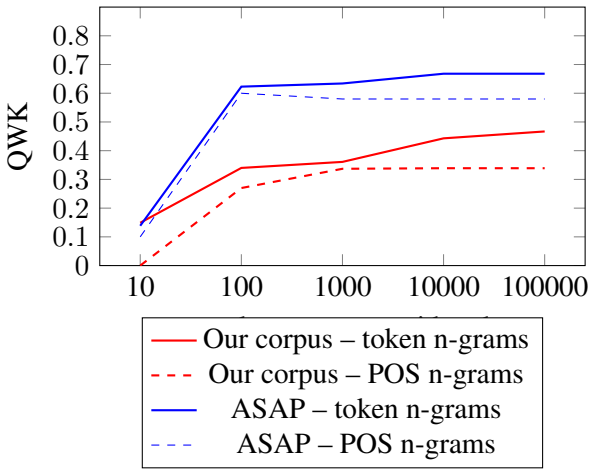

Figure 2: Assessing various numbers of token or POS n-grams as features for the scoring performance of G8.

5.4 Experiment 4: Number and Type of N-grams

We have seen that a major contribution to the performance for both ASAP and our dataset comes from token n-gram features and that we benefit from a higher number of n-grams. To further assess this influence, we take the number of available n-grams to their extremes and perform experiments using token n-grams and POS n-grams individually while varying the number of k top-frequent n-grams to extract from 10 to 10,000. Note that it can happen for POS n-grams and for token n-grams on ASAP that k is bigger than the actual number of n-grams present in the data. In that case, we take all available n-grams.

In Figure 2, we see a huge difference between ASAP and our corpus: in ASAP, a steep performance increase can be observed already with low numbers of n-grams and the curve flattens out early. In our corpus, we see a steady increase of

Configuration	G8 (aggregated)
full model	.47
– token n-grams	.41
– skip n-grams	.47
– POS n-grams	.45
– length	.46
– coherence	.47
– cohesion	.46
– syntax	.46
– occurrence	.47
– error	.44

Table 4: Ablation test (QWK) for the global G8 variable.

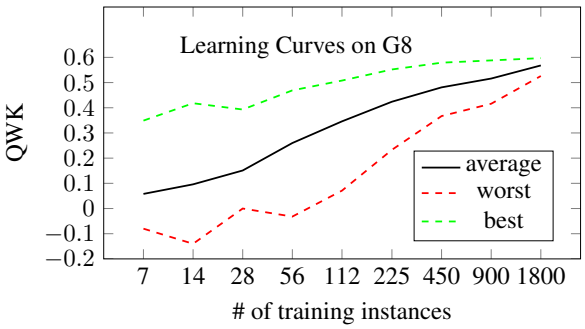

Figure 3: Learning curve experiments using different numbers of training instances, always testing on the same test set for G8.

performance that is less pronounced in the beginning and in general on a much lower level. One corpus variable explaining this effect is the average length of the essay. ASAP essays are shorter (the average number of tokens per prompt varies between 100 and 600) while our essays have a general average around 600 tokens. Even if we add more n-gram features the performance gap never closes and shows the difficulty in our data.

5.5 Experiment 5: Feature Ablation

We perform an ablation test to discern the contribution of individual feature groups. Table 4 shows the performance for the full model (using top 1,000 token, POS and skip n-grams, and stacking) and for the model with individual feature groups ablated. We chose this model because our models with more n-gram features show very similar results for the full model in comparison to n-grams only and the current settings seems most suitable to highlight potential contributions of individual features.

We can see that the feature group with the highest effect are unsurprisingly token n-grams. Most of the other features have only a minor effect. However, we saw in the comparison between the full model and n-grams only, that the additional features have a beneficial effect in our setting. We assume that our feature set is quite redundant, so that e.g. the occurrence of a connective can also be learnt from the respective unigram.

5.6 Experiment 6: Amount of Training Data

In a practical setting it is important to know how many training instances have to be available to reach a certain performance and at which amount of training data the performance levels off. This helps us to decide whether we can already fully assess the performance of our method on the given data or whether more training data would be helpful. We therefore perform a learning curve experiment showing the correlation between the number of training data and the scoring performance. In this experiment, we keep the test data constant and use the 10% held-out data for this purpose. We use the split feature set with 10,000 n-gram features which showed the best performance on the cross-validation experiments. We always double the number of training data, starting from 7 until we reach 1,800. We sample each number of training instances randomly 100 times from the pool of unlabeled data and report average, worst and best performance across those 100 runs. The resulting learning curves are shown in Figure 3 We can see that the performance varies tremendously between the best and worst runs for smaller amounts of training data. This highlights that a careful selection of training data can help when only limited human annotation effort is available. We also see that the curve starts to flatten out in the end for the best case, so that we will not profit much more from more training data.

6 Conclusions and Future Work

We have presented a set of experiments on a new challenging dataset and have shown that standard features that perform well on a standard essay scoring dataset do not perform so well here. We attribute our results to the high proficiency of our writers. Of course, we cannot be sure that some of the differences might be due to the language of the essays being German, not English and we expect that some features of the German language, such as compounds, are indeed an additional challenge. Nevertheless, we have demonstrated that essay

scoring still can be a challenging problem, calling for deeper linguistic analysis. Future work needs to concentrate on finding better representations for this kind of data, e.g. we hypothesize that recognizing argumentative structure might be helpful, as e.g. done in (Stab and Gurevych, 2016).

While the current set of essay data cannot be published for copyright reasons, we are preparing to collect and release a set of essays from the same setting from the next cohort. Essays of a similar type and in similar amounts are being collected at the beginning of each semester and we are preparing ways of getting the students' consent to publishing their anonymized essays in a corpus. In doing so, we aim at providing a challenging dataset to the community and broaden the range of available essay data.

Acknowledgements

This work is funded by the German Federal Ministry of Education and Research under grant no. FKZ 01PL16075.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Sofia, Bulgaria, pages 183–192. http://www.aclweb.org/anthology/W13-3520.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. http://aclweb.org/anthology/P/P16/P16-1068.pdf.

Björkelund Anders, Bohnet Bernd, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*. Coling 2010 Organizing Committee, pages 33–36. http://aclweb.org/anthology/C10-3009.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment* 4(3).

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series* 2013(2):i–15. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x.

Albert Bremerich-Vos and Dirk Scholten-Akoun. 2016. *Schriftsprachliche Kompetenzen von Lehramtsstudierenden in der Studieneingangsphase, Eine empirische Untersuchung*. Schneider Verlag Hohengehren, Baltmannsweiler.

Aoife Cahill, Martin Chodorow, Susanne Wolff, and Nitin Madnani. 2013. Detecting missing hyphens in learner text. *NAACL/HLT 2013* page 300.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *EMNLP*. pages 1741–1752.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. Dkpro tc: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 61–66. http://www.aclweb.org/anthology/P14-5011.

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment* 5(1).

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pages 1–11. http://www.aclweb.org/anthology/W14-5201.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* 1(2):939–944.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.* 11(1):10–18. https://doi.org/10.1145/1656274.1656278.

Claudia Harsch and Guido Martin. 2012. Adapting cef-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing* 17(4):228–250.

Claudia Harsch and Guido Martin. 2013. Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice* 20(3):281–307.

Shinichiro Ishikawa. 2011. A new horizon in learner corpus studies: The aim of the icnale project. *Corpora and language technologies in teaching, learning and research* pages 3–11.

Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016. Topicality-based indices for essay scoring. In (Tetreault et al., 2016), pages 63–72. http://aclweb.org/anthology/W/W16/W16-0507.pdf.

Chi-Un Lei, Ka Lok Man, and TO Ting. 2014. Using learning analytics to analyze writing skills of students: A case study in a technological common core curriculum course. *IAENG International Journal of Computer Science* 41(3).

Nitin Madnani, Aoife Cahill, and Brian Riordan. 2016. Automatically scoring tests of proficiency in music instruction. In (Tetreault et al., 2016), pages 217–222. http://aclweb.org/anthology/W/W16/W16-0524.pdf.

Manvi Mahana, Mishel Johns, and Ashwin Apte. 2012. Automated essay grading using machine learning. *Mach. Learn. Session, Stanford University* .

Robert Östling. 2013. Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. pages 42–47.

Ildik Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks.

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing three german treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*. Association for Computational Linguistics, Stroudsburg, PA, USA, PaGe '08, pages 40–46. http://dl.acm.org/citation.cfm?id=1621401.1621407.

Hongsuck Seo, Jonghoon Lee, Seokhwan Kim, Kyusong Lee, Sechun Kang, and Gary Geunbae Lee. 2012. A meta learning approach to grammatical error correction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 328–332.

Mark D Shermis and Ben Hamner. 2012. Contrasting state-of-the-art automated scoring of essays: Analysis. In Mark D Shermis and Jill C Burstein, editors, *Handbook of automated essay evaluation: Current applications and new directions*, Routledge, pages 313–346.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *CoRR* abs/1604.07370.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. The Association for Computational Linguistics, pages 1882–1891. http://aclweb.org/anthology/D/D16/D16-1193.pdf.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*.

Joel R. Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis, editors. 2016. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*. The Association for Computer Linguistics. http://aclweb.org/anthology/W/W16/.

Sowmya Vajjala. 2017. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education* pages 1–27. https://doi.org/10.1007/s40593-017-0142-3.

Sowmya Vajjala and Kaidi Lèo. 2014. Automatic cefr level prediction for estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*. Linköping University Electronic Press, 107.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research* 2(1):319–330.

Elena Volodina, Ildik Piln, Ingegerd Enstrm, Lorena Llozhi, Peter Lundkvist, Gunlg Sundberg, and Monica Sandell. 2016. Swell on the rise: Swedish learner language corpus for european reference level studies.

Sara C Weigle. 2013. English as a second language writing and automated essay evaluation. *Handbook of automated essay evaluation: Current application and new directions* pages 36–54.

Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge University Press, Cambridge.

Cyril J. Weir. 2005. *Language Testing and Validation*. Palgrave Macmillan, Basingstoke.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 180–189. http://dl.acm.org/citation.cfm?id=2002472.2002496.