Automatically Scoring Freshman Writing: A Preliminary Investigation

Courtney Napoles[1] and **Chris Callison-Burch**[2]
[1]Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD
[2]Computer and Information Science Department
University of Pennsylvania, Philadelphia, PA

Abstract

In this work, we explore applications of automatic essay scoring (AES) to a corpus of essays written by college freshmen and discuss the challenges we faced. While most AES systems evaluate highly constrained writing, we developed a system that handles open-ended, long-form writing. We present a novel corpus for this task, containing more than 3,000 essays and drafts written for a freshman writing course. We describe statistical analysis of the corpus and identify problems with automatically scoring this type of data. Finally, we demonstrate how to overcome grader bias by using a multi-task setup, and predict scores as well as human graders on a different dataset. Finally, we discuss how AES can help teachers assign more uniform grades.

1 Introduction

Automatic essay scoring (AES) is the task of automatically predicting the scores of written essays. AES has primarily focused on high-stakes standardized tests and statewide evaluation exams. In this paper, we consider a classroom application of AES to evaluate a novel corpus of more than 3,000 essays written for a first-year writing program.

Many colleges have first-year writing programs, which are typically large courses divided into multiple sections taught by different teachers. These essays are more representative of college writing than assessment-based datasets used for AES, and we wish to examine how AES can help students and teachers in the classroom. These preliminary experiments could help teachers evaluate students and colleges gain insight into variance across instructors.

This corpus may be more difficult to model compared to previous datasets because it lacks multiple grades to establish validity and the essays are not constrained by a prompt. Foltz et al. (2013) reported that prompt-independent scoring generally had 10% lower reliability than prompt-specific scoring.

We address several issues surrounding automatically scoring essays of this nature:

1. Is it possible to model essays graded by several different teachers with no overlapping grades?
2. ...even when scores given by each teacher have different distributions?
3. Can a single model predict the scores of long essays that are (a) not constrained by an essay prompt and (b) written in different styles?
4. How can AES provide constructive feedback to teachers and administrators?

In this work, we describe how multi-task learning can accommodate the differences in teacher scoring patterns by jointly modeling the scores of individual teachers, while sharing information across all teachers. Our multi-task model correlates strongly with actual grades. We also provide an example of how to provide feedback to help teachers grade more uniformly, using the weights learned by a linear model.

Our corpus is described in Section 3. In Section 4 we describe our experimental setup and the features used. Section 5 presents results from our system that achieve human-like levels of correlation. Section 6 discusses our results and proposes a new way to provide feedback to teachers about their grading.

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 254–263,
Denver, Colorado, June 4, 2015.

Project	Target word count	Description
1	600-770	A personal narrative that describes an experience and uses that experience to tell readers something important about the writer.
2	600	A bibliographic essay that asks you to understand the conversation surrounding your chosen topic by examining four relevant sources. Two of these sources must be at least ten years apart so that you can see how interpretations of an event, concept, or person evolve over time and that textual scholarship is an ongoing conversation.
3	600–800	A reflection that asks you to think carefully about how audience and purpose, as well as medium and genre, affect your choices as composers and reflect carefully on a new dimension of your topic.
4	1000–1200	A polished essay that asserts an arguable thesis that is supported by research and sound reasoning.

Table 1: Brief description of the assignments in the FWC, as provided by the syllabus.

2 Related Work

While AES has traditionally been used for grading tests, there are some previous applications of AES in a non-testing environment. For example, Elliot et al. (2012) used AES to assist with placement and Chali and Hasan (2012) automatically graded essays written for an occupational therapy course by comparing them to the course material.

Corpora for AES include English-language learner writing, specifically the First Certification Exam corpus (FCE), a portion of the Cambridge Learner Corpus consisting of 1,244 essays written for an English-language certification exam (Yannakoudakis et al., 2011), and the International Corpus of Learner English (ICLE), 6,085 essays written by university students across the world (Granger, 2003). The Kaggle ASAP–AES dataset has primarily native-English writing, with 22,000 short essays written by middle- and high-school students the United States (Shermis and Hamner, 2013). The FCE and Kaggle data were collected during examinations while the ICLE data was written during an exam or as part of a class assignment.

Student writing collections not suitable for AES include the Michigan Corpus of Upper-level Student Papers, with 829 academic papers that received an A grade, written by college seniors and graduate students across several disciplines (Mic, 2009). A separate corpus of freshman writing was collected at University of Michigan containing 3,500 ungraded pre-entrance essays (Gere and Aull, 2010).

Methods previously used for AES include linear regression (Attali and Burstein, 2006), rank algorithms (Yannakoudakis et al., 2011; Chen and He, 2013), LSA (Pearson, 2010; Chali and Hasan, 2012), and Bayesian models (Rudner and Liang, 2002). Recent approaches focus on predicting specific aspect of the score by using targeted features such as coherence (McNamara et al., 2010; Yannakoudakis and Briscoe, 2012).

Draft	Tokens	Sentences	Paragraphs
Intermed.	840.3	35.6	5.2
Final	938.5	39.6	5.7

Table 2: Average length of essays from the Fall 2011 semester.

Multi-task learning jointly models separate tasks in a single model using a shared representation. It has been used in NLP for tasks such as domain adaptation (Finkel and Manning, 2009), relation extraction (Jiang, 2009), and modeling annotator bias (Cohn and Specia, 2013).

3 Data

The Freshman Writing Corpus (FWC) is a new corpus for AES that contains essays written by college students in a first-year writing program. The unique features of this corpus are multiple essay drafts, teacher grades on a detailed rubric, and teacher feedback. The FWC contains approximately 23,000 essays collected over 6 semesters. To our knowledge, this is the first collection of take-home writing assignments that can be used for AES.

In this work, we consider one semester of es-

Category	Weight	Level	Possible Points	Brief Description
Focus	25%	Basics	0–4	Meeting assignment requirements
		Critical thinking	0–4	Strength of thesis and analysis
Evidence	25%	Critical thinking	0–4	Quality of sources and how they are presented
Organization	25%	Basics	0–4	Introduction, supporting sentences, transitions, and conclusion
		Critical thinking	0–4	Progression and cohesion of argument
Style	20%	Basics	0–4	Grammar, punctuation, and consistent point of view
		Critical thinking	0–4	Syntax, word choice, and vocabulary
Format	5%	Basics	0–4	Paper formatting and conformance with style guide

Table 3: The rubric for grading essays. The teachers used a more detailed rubric that provided guidelines at each possible score.

says from the FWC, for a total of 3,362 essays written by 639 students during the Fall 2011 semester.[1] Students were enrolled in the same Composition I course, which was divided into 55 sections taught by 21 teachers. All sections had the same curriculum and grading rubric.

The course had four writing projects, and for each project students could hand in up to three drafts: Early, Intermediate, and Final. Each project focused on a different type of essay, specifically a personal narrative, a bibliographic essay, a remediation, and a thesis-driven essay, but the topic was open-ended. A description of the requirements for each essay is found in Table 1.

Submission and grading was done on My Reviewers.[2] Students uploaded PDF versions of their essays to the site, where teachers graded them. Teachers could also comment on the PDFs to provide feedback to the students.

We downloaded the essays in PDF format from MyReviewers, extracted text from PDFs using the PDFMiner library[3], and automatically labeled text by document section based on its (x, y) position on the page. Document sections include header, title, paragraph, page number, and teacher annotation.

To anonymize the data, we replaced student and teacher names with numeric IDs. We ran sentence segmentation on the paragraphs using Splitta (Read et al., 2012) and added several layers of annotation to the sentences: constituent and dependency parses, named entities, and coreference chains using Stanford Core NLP (Manning et al., 2014); 101 discourse markers with the Explicit Discourse Connectives Tagger[4]; and 6,791 opinion words defined by Hu and Liu (2004).

In this work, we only consider the Intermediate and Final drafts. We leave out Early drafts because less than half of Final essays have an Early draft (80% have an Intermediate draft) and Early drafts are typically short outlines or project proposals, while Intermediate drafts generally have a similar form to the Final draft. The average essay has 899 words, 38 sentences, and 5.5 paragraphs (Table 2 has lengths by draft).

3.1 Scores

All essays were graded on the same rubric, which has five categories broken into eight sub-categories, with bulleted requirements for each. The overall score is a weighted combination of the individual category scores that ranges from 0–4, which corresponds to a letter grade. (A condensed version of the rubric is shown in Table 3, and the correspondence between score and grade is shown in Figure 1.) This grading scheme has two immediate advantages, the first that students have a clear sense of how different aspects of their paper contributes to the grade,

[1] There were 3,745 graded essays in total, but we were unable to automatically extract text from 383 of the PDFs.

[2] www.myreviewers.com/

[3] http://www.unixuser.org/∼euske/python/pdfminer/index.html

[4] http://www.cis.upenn.edu/∼epitler/discourse.html

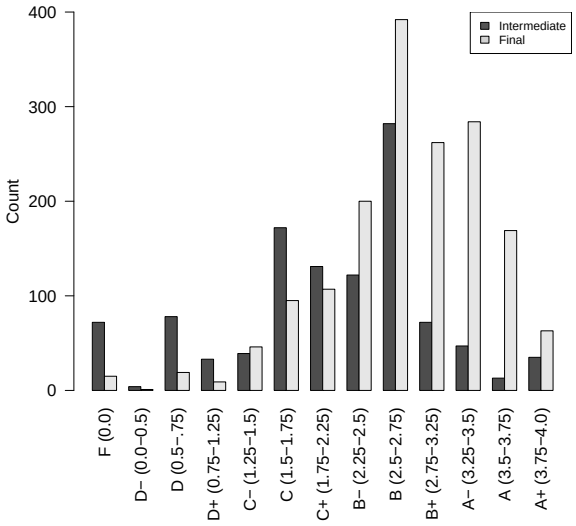

Figure 1: Number of essays by grade. Each letter grade corresponds to a range of numeric scores, in parentheses.

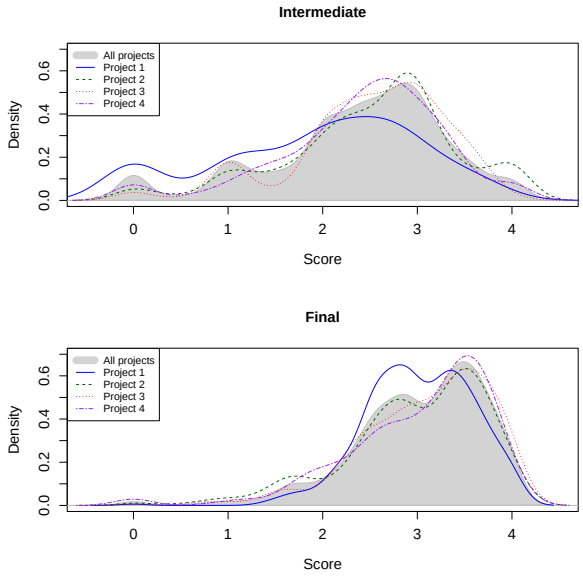

Figure 2: Distribution of scores by project and draft.

Project	Intermediate	Final	Change
1	1.94	3.02	+1.08
2	2.51	2.98	+0.70
3	2.31	3.09	+0.87
4	2.35	3.02	+0.69
All	2.35	3.03	+0.86

Table 4: Average score for each draft by project, including the average change in score between the Intermediate and Final drafts. The standard deviation of the Intermediate and Final draft scores are 0.92 and 0.68, respectively.

and the second to promote consistent grading across teachers (Graham et al., 2012).

The grade "curve" is different for Intermediate and Final drafts (Kolmogorov-Smirnov test, $D = 0.332$, $p < 10^{-10}$) and the scores of neither draft are normally distributed by the Shapiro-Wilk test (Intermediate: $W = 0.948$, $p < 10^{-10}$, Final: $W = 0.932$, $p < 10^{-10}$). Figure 2 illustrates the distribution of grades across projects and drafts. Intermediate scores have higher variance and tend to be below 2.5 (corresponding to a B grade), while Final scores are more tightly distributed, the majority of them at least a B grade (Figure 5 and Table 4).

3.2 Teachers

Since each essay is graded by one teacher, we cannot guarantee that teachers grade consistently. To illustrate the differences between teacher grades, we randomly selected nine teachers who graded at least 150 Intermediate and Final drafts and graphically represented the score distribution assigned by each one (Figure 3).

A one-way ANOVA on the Intermediate draft scores revealed a significant difference between at least one pair of teachers' scores (17 teachers, $F(16, 1079) = 51.9$, $p < 10^{-10}$), and Tukey's post-hoc analysis revealed significant differences between 66 pairs of teachers ($p < 0.001$). Similar results were found for the Final drafts (20 teachers, $F(19, 1642) = 15.57$, $p < 10^{-10}$; 44 pairs significantly different $p < 0.001$). Even with a detailed rubric, teachers appear to grade differently.

In Figure 4, we compare the correlation of four features to the scores assigned by different teachers. This figure provides an example of how teachers exhibit a considerable amount of variance in how they unconsciously weight different criteria.

3.3 Students

We do not have access to specific demographic information about the students, but we can make es-

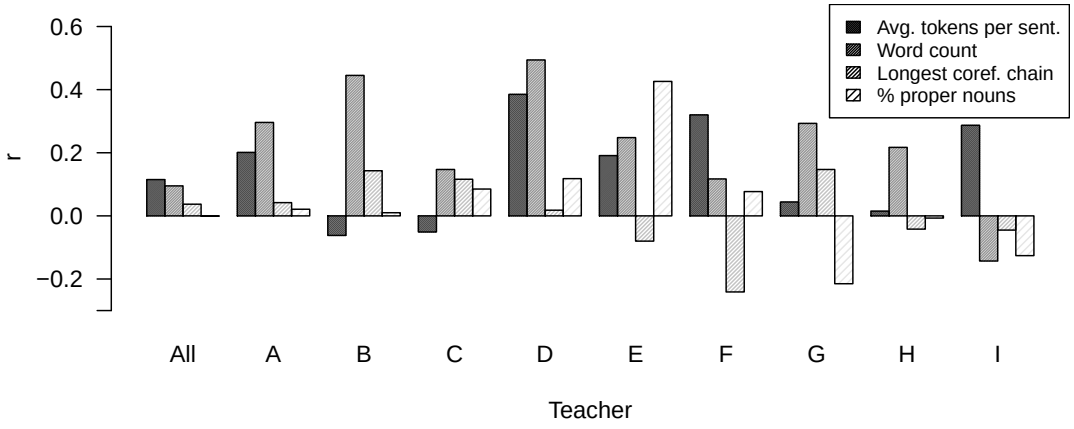

Figure 4: The correlation of four different features with Final draft scores, compared across nine teachers.

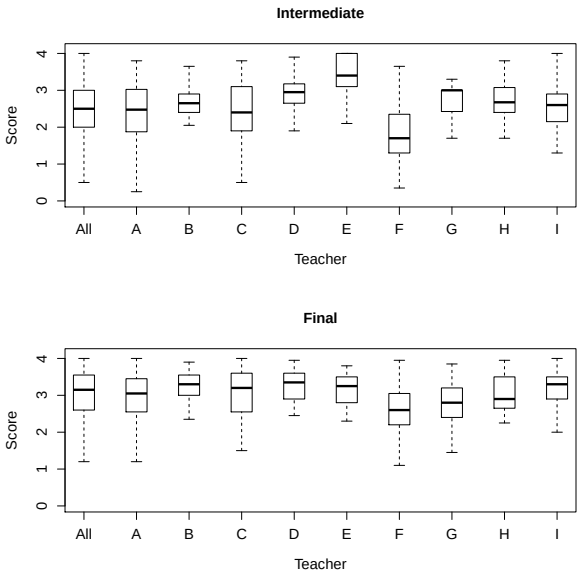

Figure 3: Distribution of scores given by nine teachers.

timates of their writing ability and native language. The writing course is a university requirement that students can place out of if they have completed a comparable course or received a sufficient grade in any number of pre-college standardized tests.[5] Therefore, we assume that the students in this course require additional support to develop college-level writing skills.

We also assume that the majority of students in this course are native English speakers. Because native English speakers and English language learners generally have different difficulties with writing, we wished to estimate how many of the students in the course were native English speakers. 96% the student body as a whole are American citizens, whom we assume are native English speakers. If the demographics of the writing course are the same as the university as a whole, then at most 4% of the students are non-native English speakers, which is our lower-bound estimate.

We arrive at an upper bound if we assume that every international student in the freshman class (168 out of 4,200 total students) is in the writing class, or at most 26% of the writing class are non-native speakers. In reality, the number is probably somewhere between 4–26%.

4 Experiments

We separated 3,362 essays by draft, Intermediate and Final (1,400 and 1,962 essays, respectively, skipping 31 Intermediate drafts that had no grade assigned). We randomly selected 100 essays for development and 100 for testing from each draft type and

[5]For example, students need a 4 in an English Language/Literature AP course, or a 5 in an IB English course to place out.

represented all essays with feature vectors.[6]

4.1 Model

In this work, we establish a single-task model and explain how it can be extended for multi-task learning. The single-task model represents essays graded by every teacher in the same feature space.

We have n essays graded by T teachers and m features. In the single-task setup, we represent each essay by a vector containing the values of these m features calculated over that essay. An essay $\boldsymbol{x}$ is represented as an m-dimensional vector:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$$

For multi-task learning, we make a copy of the entire feature set for each of the T teachers. Each of the original features has a global feature and one feature specific to each teacher, for a total of $(1 + T) \times m$ features. For example, an essay graded by teacher A has a set of global features that are equal to the teacher-A-specific feature values. The features specific to other teachers are assigned zero value.

Specifically, we have an n-dimensional teacher vector $\boldsymbol{t}$, such that t_i is the teacher that graded essay i. In the multi-task framework, each essay is represented by a $(1 + T) \times m$-dimensional vector, $\boldsymbol{x}^*$. The new vector $\boldsymbol{x}^*$ contains twice as many non-zero features as the original vector $\boldsymbol{x}$,

$$\boldsymbol{x}^* = (x_1, x_2, \ldots, x_m, x_{t_i 1}, x_{t_i 2}, \ldots, x_{t_i m}, \ldots) \quad \text{s.t. } x_j = x_{t_i j} \tag{1}$$

We favor linear models in this work because the contribution of each feature is transparent, which allows us to provide teachers with feedback based on the weights learned by the model. In the multi-task setup, we used principal component analysis to transform the features into a lower dimension to reduce computational burden. scikit-learn was used for dimensionality reduction and model learning.

Since there is a mapping between scores and letter grades, we experimented with closed-class classification as well as ranking classification, but linear regression yielded the best results on the development set. We predicted scores using linear regression over a number of features, described in Section 4.2 below.

[6] Analysis in Section 3 was done over the training set only.

For evaluation, we report the correlation between predicted and actual scores as Pearson's r and Kendall's τ, as well as the mean squared error. We round all predictions to the nearest 0.05, to conform with the actual scores. We also report the exact agreement and quasi-adjacent agreement, which we define as a predicted score within 0.25 points of the actual score (approximately the difference between a grade G and a G+ or G-).

Using the same experimental setup, we learn different models to predict

- the overall score of Intermediate and Final drafts,
- the score of individual rubric components, and
- the score improvement from an Intermediate to Final draft.

4.2 Features

We broadly categorize features as surface, structural, lexical, syntactic, and grammatical.

Surface features include average word, sentence, and paragraph lengths; lengths of the longest and shortest sentences; and number of tokens, sentences, and paragraphs. Another feature indicates the ratio of unique first three words of all sentences to the total number of sentences, to loosely capture sentence variety. (9 features)

Structural features include the frequency of discourse markers and the number of sentences containing discourse markers, as well as measures of cohesion, specifically the average and longest coreference chain lengths and the number of coreference chains (representing the number of entities discussed in the essay). Finally, we calculate the following statistics over the first, last, and body paragraphs: number of polarity words, number of "complex" words (with more than 3 syllables), and Flesch–Kincaide grade level. (25 features)

Lexical features are token trigrams skipping singletons and bag of words without stop words. We also include ratios of each of the following to the number of tokens: stop words, out-of-vocabulary words, proper nouns, and unique token types. (5 + # tokens - # stopwords + # token trigrams features)

Syntactic features include the average and longest lengths between the governor and dependent in all dependency relations; the number of clauses in an essay, specifying subordinating clauses, direct

	Intermediate Drafts					Final Drafts				
Model	r	τ	**MSE**	**Exact**	**Adj.**	r	τ	**MSE**	**Exact**	**Adj.**
Baseline	0.045	-0.008	1.995	0.094	0.323	0.101	0.098	0.876	0.180	0.450
Single-task	0.399	0.274	0.980	0.198	0.469	0.252	0.157	0.997	0.130	0.440
Multi-task	*0.755*	*0.558*	*0.474*	*0.323*	*0.708*	*0.558*	*0.408*	*0.397*	*0.250*	*0.760*

Table 5: Correlation between predictions and teacher scores, measured by Pearson's r and Kendall's τ, as well as the mean squared error (MSE) and exact and adjacent agreements. The baseline is a random balanced sample.

questions, and inverted declarative sentences and questions; the number of passive and active nominal subjects; the tallest and average parse-tree heights; and the ratios of adjective, prepositional, and verb phrases to noun phrases. (14 features)

Grammatical features are trigram counts of part-of-speech (POS) tags and the number of POS 5-grams unseen in a 24-million-token portion of the English Gigaword corpus. We also include the perplexity assigned to the text by three language models: a 500k-token Gigaword LM, and LMs estimated over the correct and incorrect learner text from the NUCLE 3.2 corpus. (4 + # POS trigrams features)

5 Results

5.1 Predicting the overall score by draft

We learned two single-task models using the features described above, one for Intermediate drafts and one for Final drafts, and the correlation between the predicted and actual scores was well below human levels. By introducing a multi-task approach (Section 4), the model made significant gains, with the correlation increasing from $r = 0.422$ to $r = 0.755$ and from $r = 0.252$ to $r = 0.558$ for the Intermediate and Final drafts, respectively. The Intermediate model predicts scores that very strongly correlate with the human score, and does as well as a human grader. Results are summarized in Table 5.

Using the same setup, we trained separate models for each of the projects, and found that the individual models did not do as well as a composite model (Table 6).

5.2 Predicting specific rubric scores

Next, we predicted individual rubric scores with multi-task learning. The rubric scores that correlate most with overall score are Organization, Evidence, and Focus ($r >= 0.84$), and we were curious whether our model would do better predicting

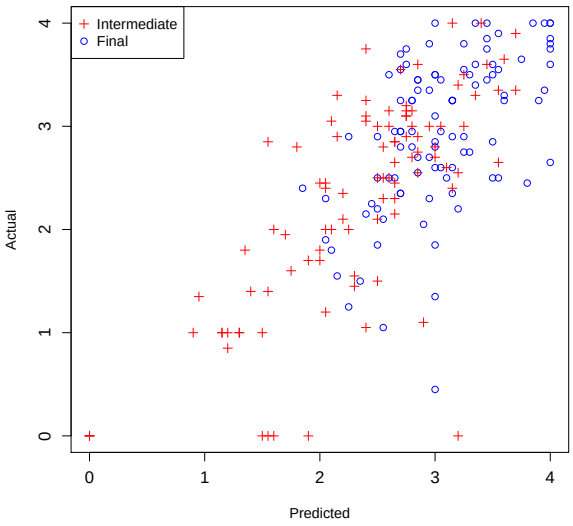

Figure 5: Predicted versus actual essay scores.

those rubric categories than the others. Focus and Evidence predictions correlated very strongly, but the Organization predictions had weaker correlation with the actual scores (Table 7).

5.3 Predicting score change

In a preliminarily experiment to predict the improvement between draft pairs, we represent each draft pair by a vector that was the difference between the feature vector of the Intermediate and the Final drafts. Less than 10% of Final drafts show a decrease in score and on average the score increases 0.86 between the Intermediate and Final draft, so a binary classification of whether the score improved would be trivial. Instead we aim to predict the *amount* of the score change.

Training single-task and multi-task models over 794 draft pairs from the same training set above, we tested 50 pairs of essays. The single-task model pre-

Project	Intermediate	Final
P1	0.859	0.511
P2	0.706	0.483
P3	0.571	0.463
P4	0.591	0.382
P1–4	0.704	0.454

Table 6: The correlation (Pearson's r) of actual scores to predictions made by individual models for each project/draft pair. P1–4 represents predictions of all project models.

Model	r	MSE
Baseline	0.067	0.815
Single-task	0.346	4.304
Single-task, no content	0.087	0.399
Multi-task	-0.027	5.841
Multi-task, no content	*0.356*	*1.702*

Table 8: Correlation between the predicted and actual change between Intermediate and Final draft scores.

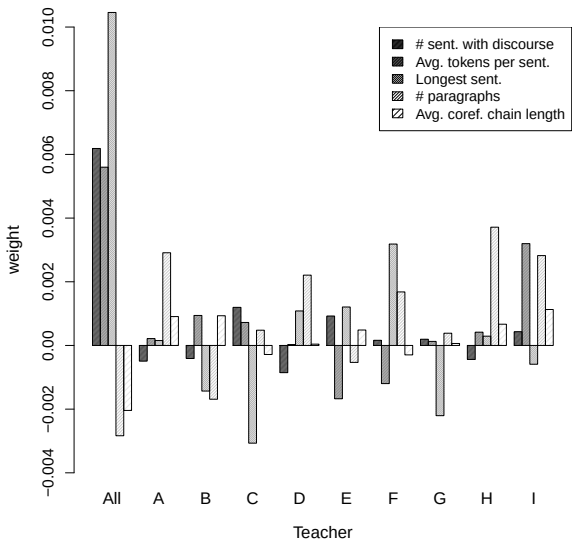

Figure 6: A comparison of feature weights learned in individual, teacher-specific models.

dicted the change much better than the multi-task, ($r = 0.346$ versus $r = -0.027$, which is worse than a random balanced baseline). When we removed content features (unigrams and trigrams), the multi-task model outperformed the single-task model with content, both by correlation and MSE. Removing content features significantly degraded the performance of the single-task model (Table 8).

5.4 Potential for providing feedback

We trained individual models for each of 17 teachers over Intermediate drafts, without dimensionality reduction. The predicted scores correlated strongly with the instructor scores ($r = 0.650$). We isolated the features with the heaviest average weights across all 17 models to examine whether teachers weighted these features differently in the individual models, and found that these weights varied by magnitude and polarity (Figure 6).

A graphical representation of this type could provide useful feedback to teachers. For example, the longest sentence feature has a high negative weight for teachers C and G, but is positively weighted for the other teachers. Given this information, teachers C and G could slightly alter their grading practices to better match the other teachers. However, before such a technology is deployed, we would need to develop more reliable models, examine the essays to check that the features is not a proxy for some other aspect of the text, and perform pilot testing.

6 Discussion and Future Work

One of the primary challenges of our dataset is the lack of multiple annotations. We only have one score for each essay, and the scores are provided by 21 different teachers whose grades are from different distributions. Modeling scores from different distributions in a single task yields predictions that only weakly correlate with the actual scores.

A joint model across all teachers and all projects does better than individual models for predicting essay scores. The multi-task setup enables us to jointly model characteristics of individual teachers while taking advantage of shared information across all teachers, and the models' predictions strongly correlate with human scores. On the Intermediate drafts, the correlation is very strong and within the range of human–human correlation (inter-human correlations ranged from 0.61 to 0.85 on the Kaggle ASAP–AES data (Shermis and Hamner, 2013)).

Unlike the Kaggle data, these essays are open ended, and open-ended topics are thought to be more difficult to score (Foltz et al., 2013). Furthermore,

Draft	Overall	Focus	Evidence	Organization	Style	Format
Intermediate	0.755	0.720	0.789	0.666	0.594	0.787
Final	0.558	0.340	0.324	0.329	0.350	0.432

Table 7: Correlation (Pearson's r) of predicted to actual scores for individual rubric categories.

the form of each project is different (personal narrative, bibliographic essay, remediation, thesis-driven essay), and we are able to score these different types of open-ended essays using a single model.

Our model predicts Intermediate scores better than Final scores, possibly because Intermediate drafts have higher variance than Final drafts, which are more tightly clustered, with more than 50% of the scores between 2.5 and 3.5. The adjacent agreement and MSE are better for Final drafts than Intermediate, suggesting that even though the correlation of Final drafts is weaker, the predictions are within a close range of the true scores.

We have shown that multi-task learning makes better predictions, and in the future we will apply multi-task learning to grading new teachers.

In addition to predicting the overall essay scores, we applied the same setup to two other tasks facilitated by this dataset: predicting individual rubric scores and predicting the score change from Intermediate to Final drafts. We found room for improvement in both tasks. To predict isolated rubric scores, future work will include investigating different features tailored to specific aspects of the rubric.

Our experiments in predicting improvement from Intermediate to Final draft revealed that content features confound a multi-task model but a single-task model does better with content features. This suggests that the single-task, no-content model underfits the data while the multi-task, with-content model overfits, illustrating the potential benefit of a multi-task setup to low-dimensional space.

There are inconsistencies in the paired-essay data, which may confound the model. 23 essays did not change between the Intermediate and Final drafts. Of these essays, the score decreased for 9, remained unchanged for 5, and increased for 9 essays–in two instances, the score increase was 2 points or more. Further analysis is warranted to determine whether there was a rationale for how the scores of unchanged essays were assigned.

Future work includes having the essays re-scored by another grader to establish validity. Until then, we cannot claim to have developed a reliable system, only to have robustly modeled the grading tendencies of this particular set of teachers for this class.

7 Conclusion

Consistent grading across teachers is difficult to achieve, even with training and detailed rubrics (Graham et al., 2012). Automatic tools to provide constant feedback may help promote consistency across teachers. This work is the first step aiming to identify when and how teachers grade differently. In the future, we hope to drill down to separate rubric scores so that we can provide specific feedback when teachers use different internal criteria.

In this work we introduced a new set of essays for evaluating student writing that is more representative of college writing than previous AES datasets. We developed a single, robust system for automatically scoring open-ended essays of four different forms (personal narrative, bibliographic, reflective and thesis driven), graded by 21 different teachers. Our predictions correlate strongly with the actual scores, and predicts the scores of Intermediate drafts as well as human raters on a different set of essays. We present a method for handling a dataset labeled by multiple, non-overlapping annotators.

This is an exciting new dataset for educational NLP, and this paper presents just a sample project facilitated by its unique characteristics. At this time we cannot release the corpus due to privacy concerns, but we hope it will be available to the community at some point in the future.

Acknowledgments

We thank Joseph Moxley for his assistance in obtaining the corpus, Burr Settles for his ideas in developing a multi-task approach, and Benjamin Van Durme and the reviewers for their feedback.

References

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

Yllias Chali and Sadid A. Hasan. 2012. Automatically assessing free texts. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 9–16, Mumbai, India, December. The COLING 2012 Organizing Committee.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA, October. Association for Computational Linguistics.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria, August. Association for Computational Linguistics.

Norbert Elliot, Perry Deess, Alex Rudniy, and Kamal Joshi. 2012. Placement of students into first-year writing courses. *Research in the Teaching of English*, 46(3):285–313.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610, Boulder, Colorado, June. Association for Computational Linguistics.

Peter W. Foltz, Lynn A. Streeter, Karen E. Lochbaum, and Thomas K. Landauer. 2013. Implementation and applications of the intelligent essay assessor. *Handbook of Automated Essay Evaluation*, pages 68–88.

Anne Ruggles Gere and Laura Aull. 2010. Questions worth asking: Intersections between writing research and computational linguistics. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 51–55, Los Angeles, CA, USA, June. Association for Computational Linguistics.

Matthew Graham, Anthony Milanowski, and Jackson Miller. 2012. Measuring and promoting inter-rater agreement of teacher and principal performance ratings. *Online Submission*.

Sylviane Granger. 2003. The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3):538–546.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Jing Jiang. 2009. Multi-task transfer learning for weakly-supervised relation extraction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1012–1020, Suntec, Singapore, August. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Danielle S. McNamara, Scott A. Crossley, and Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

2009. Michigan corpus of upper-level student papers. The Regents of the University of Michigan.

Pearson. 2010. Intelligent Essay Assessor fact sheet. Technical report, Pearson.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jrgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India, December. The COLING 2012 Organizing Committee.

Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Mark D. Shermis and Ben Hamner. 2013. 19 contrasting state-of-the-art automated scoring of essays. *Handbook of automated essay evaluation: Current applications and new directions*, page 313.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43, Montréal, Canada, June. Association for Computational Linguistics.

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*, pages 180–189, Portland, Oregon, USA, June. Association for Computational Linguistics.