AUTOMATED ESSAY GRADING FROM SCRATCH

WON I. LEE, QIN LYU, AND ZELONG QIU

ABSTRACT. Automated essay grading is an important problem for advancing curricular assessments in K-12 education, as it allows for rapid evaluation of student-constructed responses and opens the way for the assessment of more sophisticated analytical and reasoning skills. In this work, we investigate the performance of various deep architectures on the task of automated essay grading. We employ no hand-crafted features and explore both character-level and word-level models for this task, using both convolutional and recurrent architectures. We find that ... . In addition, we provide a more nuanced analysis of the strengths of various architectural and training choices by considering performance across a variety of metrics, showing that ... .

INTRODUCTION AND RELATED WORK

One prominent impediment preventing the enhancement of curricula in K-12 education to focus more on critical reasoning ability and analytical skills is the difficulty of scoring tests to measure these abilities. It is generally the case that measuring such skills requires inviting the student for open-ended answers rather than simple, multiple-choice ones; unfortunately, tests that require essays or other individually-written responses are difficult to score automatically and necessitate scoring by human graders. Such grading involves substantial time and expense from both testing companies and government agencies (for standardized exams), and as a result, most such tests have relied more on multiple-choice questions, limiting the ability to assess more sophisticated skills.

This has long motivated the development of automated system for essay grading. Much effort has been put into building a satisfactory system that could reliable generate scores in line with those of human judges. The Intelligent Essay Assessor (IEA) was first used to score essays for large undergraduate courses in 1994 [1], and the e-Rater, an automated reader developed by the Educational Testing Service was used to grade large-scale essays on standardized exams [2].

All of these examples used on the order of hundreds of hand-crafted features to generate the essay scores. Recently, however, the development of more automatic methods in natural language processing based on deep neural networks has suggested that systems for such tasks boasting comparable or even superior performance can be constructed "from scratch" without manually crafting features for the algorithm. For example, the work of Collobert et al. [3], which helped to inspire the current wave of deep learning applications in NLP, demonstrated that appropriately-trained neural network models incorporating no hand-crafted features could achieve performance rivaling that of state-of-the-art systems based on extensive feature crafting.

More recently, a variety of deep architectures have emerged for various tasks in NLP. In particular, the sequence-to-sequence model, incorporating two LSTMs (encoder and decoder), has emerged as the paradigm for sequential tasks such as machine translation [?]. Certain architectures often consider character-level inputs instead of or in addition to word-level inputs, often employing convolutional neural networks (CNNs) as a representation-generating layer before feeding into a fully-connected model or recurrent network [?].

All of these networks offer important features and benefits that are suitable for certain tasks. For the task of automated essay grading, however, we note that many of the advantages offered by each respective model are desirable. For example, as described below, it would be beneficial to employ character-level inputs for essay grading; in another vein, the sequential information captured by recurrent networks are often vital to judging the quality of essays. Thus, in this work, we report on a series of experiments testing which model architectures work best for the task of automated essay grading. Of course, we cannot be exhaustive in this search, but we aim to provide an illustrative and meaningful look at the advantages of certain architectures and training methods over others.

Essay Set	Score Range	Train Set Size	Valid Set Size	Test Set Size
1	2-12	1426	179	178
2	1-6	1416	193	191
3	0-3	1384	166	176
4	0-3	1407	183	180
5	0-4	1458	178	169
6	0-4	1458	157	185
7	2-24	1257	151	161
8	10-60	612	52	59
Total	0-60	10419	1260	1300

TABLE 1. Statistics for the data set of essays and scores.

DATA

Our dataset consists of essays selected from 8 different essay prompts across a variety of exams. These essay sets differ in the length of essays with which they are associated, their possible score ranges, and the number of examples, among other qualities. We provide an illustration of the statistics associated with each essay set in Table 1, demonstrating the large variability in the score ranges and number of examples.

The dataset contains both the essay set ID and essay ID, as well as the entire text of the essay. For evaluation, the dataset contains one or more human scores, and the final resolved human score (generally an average of scores if there are multiple available for a given essay). The final resolved score is the measure of interest for this task, and in our experiments we only consider this measure as the quantity of interest.

MODELS

Character-Level CNN. A prominent issue with essay grading is that many of the words in candidate essays are misspelled, so that they lie outside a fixed vocabulary. Most models with fixed vocabulary word embeddings can only deal with such words by treating them as "UNK" tokens, which removes any information that such words may convey about the quality of the essay. Unlike most natural language tasks for which out-of-vocabulary words are infrequent and often unimportant, misspellings can provide valuable information regarding the quality of the writing for grading student essays.

One way to deal with such cases is to consider character-level models rather than word-level models. Thus, we consider a character-level convolutional neural network (CNN), which embeds each character into a d-dimensional vector ($d \approx 30$) and employs a series of 1D convolutions and max-pooling layers to extract efficient representations for inference. We stack fully-connected layers on top of the convolutional layers for final processing, and use either dense or one-hot encodings of the characters as input. We provide a schematic of our overall architecture in Figure 1. We consider a 6 convolutional layer model inspired by the work of Zhang, Zhao, and LeCun, 2015 [?] as well as a simpler, 1 convolutional layer model inspired by the work of Kim, 2014 [?].

Training. In this formulation, we primarily treated the essay score as a percentage of the score range (i.e. for essay set 1, which has range 2-12, an essay with score 6 would be treated as 0.4), and used logistic regression as our final output. We used standard stochastic gradient descent (SGD) with momentum for training, using binary cross-entropy loss as our objective.

Word-Level LSTM. An issue encountered in character-level models on this task, however, is that the models are unable to leverage semantic relationships between the words in order to make better predictions regarding essay quality.

Ranking Model.

METRICS

For the purposes of investigating the respective strengths of the various models considered above, we also consider a number of different evaluation metrics, which highlight different performance properties. We note that these are *not* the metrics used during training, but rather metrics used to evaluate the performance of
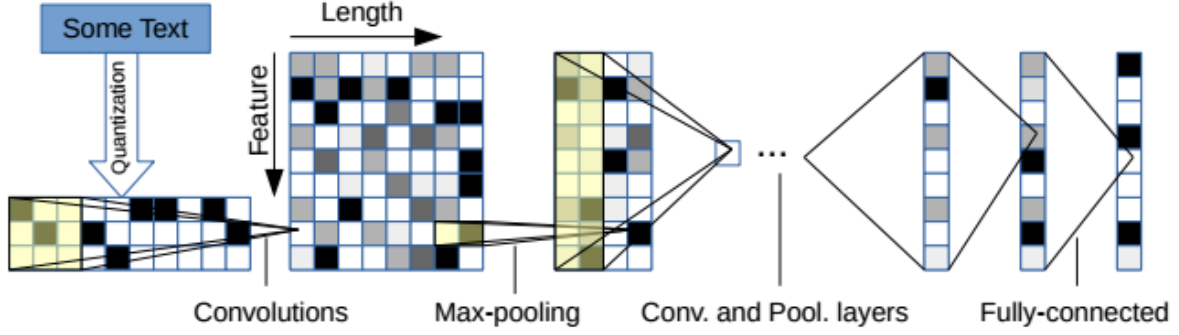
FIGURE 1. Architecture of character-level CNN model; figure credit to Zhang, X., J. Zhao, and Y. LeCun (2015).

Model	Accuracy	MSE	nDCG	Quadratic Kappa
Char-level CNN (1-layer, one-hot)				
Char-level CNN (1-layer, embed)				
Char-level CNN (6-layer, one-hot)				
Char-level CNN (6-layer, embed)				

TABLE 2. Performance on test set evaluated via the metrics discussed above.

the models on the test set. For all formulas, we let p_i to be the predicted score on the essay i and y_i to be the actual score, with a total test size of N.

Classification Accuracy. As the essay grading problem is often treated as a classification problem (each discrete score being a "category" of essays), we consider evaluating on the simplest classification metric, namely that of accuracy on the test set:

$$ACC \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{p_i = y_i}$$

where $\mathbb{I}_{x=y}$ is an indicator of whether $x = y$.

Mean-Squared Error (MSE). One issue with a raw classification accuracy metric is that it does not provide any indication of the discrepancy between the actual and predicted scores. For example, on a scale of 2-12, a model predicting a score of 3 would incur the same loss as a model predicting 9 for an actual score of 10. We would clearly prefer the latter model over the former model, however; indeed, for grading purposes, it is not as important to achieve the exact same score so much as it is to obtain predictions that are on average close to the actual scores.

Thus, we consider the traditional mean-squared error (MSE) as one of our metrics, namely:

$$MSE \equiv \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2$$

Discounted Cumulative Gain (nDCG).

Quadratic Kappa.

Results

Discussion

Member Contributions

Won Lee. Prior to the switch in our project from music-to-lyric generation to automated essay grading, I conducted all data generation (i.e. cleaning, formatting, compiling music-lyric pairs, etc.) and finalized models and training for training/evaluation on Harvard's clusters. With regard to essay grading, I developed all of the character-level models and conducted exhaustive grid search and training for these models. I also contributed to the evaluation script as well as various data-related auxiliary scripts for loading onto the models. Finally, I did most of the write-up for the current report as well as the proposal.

Qin Lyu.

Zelong Qiu.

References

[1] Foltz, P. W., et al. (2013). "Implementation and applications of the Intelligent Essay Assessor." *Handbook of Automated Essay Evaluation*: 68-88.

[2] Attali, Y. and J. Burstein (2006). "Automated Essay Scoring with e-Rater V2." *The Journal of Technology, Learning and Assessment* 4.

[3] Collobert, R.