Automated Essay Scoring
Gupta, Amit

Problem Statement: Essay grading is a time consuming effort and is often doubled scored involving two human raters to ensure grading consistency. There is a need for fast, effective and affordable solutions for automated grading of student-written essays. Traditional automation approaches still require a significant human involvement for complex feature engineering to identify grammar correctness, text coherence, logical flow and time consuming implementation for these features with supervised shallow learning networks (SVM, Random Forest et al.)

Technology: With deep learning models we will show the models learn to discriminate important features in the essays automatically, and come close to predicting the scores assigned by human raters. In this project we employ various NLP data augmentation techniques to expand our limited training samples dataset to better train a stacked CNN based model.

Benefits: Automated essay scoring systems are targeted at both alleviating the workload of teachers, and accelerating the feedback cycle time in educational systems (schools, standardized tests like SAT, GRE). The techniques used in this project can be expanded to adjacent applications like resume scoring.

Challenges: The dataset for automated essay scoring is tiny with ~1800 samples for each of the essay prompts. Further, the samples in the observed essay prompt are unbalanced with 2% of the essays having essays with a score of zero. This posed a challenge for the models to correctly predict essays with a score of zero, and this was reflected in our results with non-augmented data.

Result: Our deep learning models achieve a kappa score ("alignment with human rater") of 0.48-0.51 with the original dataset, and 0.76 with data augmentation. The state of art featured engineered solutions demonstrated a kappa score of 0.66, and the state of art deep learning models have a kappa score of 0.80 (paper). The results show the various data augmentation techniques deployed in this project are effective in boosting the performance of simpler stacked CNN models targeted for text scoring, and show better performance against feature engineered solutions, and comes close to the state of art deep learning models.

Youtube URL:
2 min: **https://youtu.be/53hlsm-CWfA**

Dataset and Data Augmentation
AES Dataset:**https://www.kaggle.com/c/asap-aes/data**
NLP Data Augmentation: **https://arxiv.org/abs/1901.11196**

Key Papers
Neural Coherence : **https://arxiv.org/abs/1711.04981**
Siamese BiLSTM: **https://www.mdpi.com/2073-8994/10/12/682**

Introduction: Automated Essay Scoring (AES) systems are targeted at both alleviating the workload of teachers and improving the feedback cycle in educational systems. AES systems have also seen adoption for several high-stakes assessment, e.g., the e-rater system which has been used for TOEFL and GRE examinations. A successful AES system brings about widespread benefits to society and the education industry. Traditionally, the task of AES has been regarded as a machine learning problem which learns to approximate the marking process with supervised learning. Decades of AES research follow the same traditional supervised text regression methods in which handcrafted features are constructed and subsequently passed into a machine learning based classifier. A wide assortment of features are commonly extracted from essays. Simple and intuitive features may include essay length, sentence length. On the other hand, intricate and complex features may also be extracted, e.g.., features such as grammar correctness, readability and textual coherence. However, these handcrafted features are often painstakingly designed, and require a lot of human involvement and usually require laborious implementation for every new feature.

In this project we have applied deep learning models to learn important features of the essay automatically, and compare the results with the state of art deep learning models.

Dataset Description: There are eight essay sets. Each of the sets of essays was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students ranging in grade levels from Grade 7 to Grade 10. All essays were hand graded and were double-scored. Each of the eight data sets has its own unique characteristics. The variability is intended to test the limits of your scoring engine's capabilities. The training data is provided in three formats: a tab-separated value (TSV) file, a Microsoft Excel 2010 spreadsheet, and a Microsoft Excel 2003 spreadsheet. The current release of the training data contains essay sets 1-6.

Each of these files contains 28 columns:
essay_id: A unique identifier for each individual student essay
essay_set: 1-8, an id for each set of essays
essay: The ascii text of a student's response
rater1_domain1: Rater 1's domain 1 score; all essays have this
rater2_domain1: Rater 2's domain 1 score; all essays have this
rater3_domain1: Rater 3's domain 1 score; only some essays in set 8 have this.
domain1_score: Resolved score between the raters; all essays have this
rater1_domain2: Rater 1's domain 2 score; only essays in set 2 have this
rater2_domain2: Rater 2's domain 2 score; only essays in set 2 have this
domain2_score: Resolved score between the raters; only essays in set 2 have this
rater1_trait1 score - rater3_trait6 score: trait scores for sets 7-8

Note: We are using the tsv training data file, and this project is focussed on training and evaluation of one essay prompt namely "essay set 3" as each essay is different (argumentative, narrative, source response), has a different marking structure, and taken by different grade level students (7th thru 10th). “Essay Set 3” is one of the harder essays to train judging by various papers (see Key Papers). It is “source-response” type of essay. Please see the appendix for more information on “Essay Set 3”, and the rubric.

Further we have considered only 4 columns namely **essay_id**, **essay_set**, **essay** and **domain1_score** in our models. Many of the fields in other columns were found to have no data.

Data Analysis:

	Min	Max	Median
Essay Length (words)	2	189	48

	Score of 0	Score of 1	Score of 2	Score of 3	Total
Number of essays	39 (unbalanced class)	607	657	423	1726

Sentence Length correlation to score is high	0.7
90% of essays have an essay length of	91 words

Deep Learning Model: Following are the list of techniques applied in this project

Vectorization: TF-IDF, and jointly learned embeddings

Models evaluated : Dense, CNN+Dense, LSTM+Dense, BiLSTM+Dense, CNN+LSTM+Dense, CNN+BiLSTM+Dense

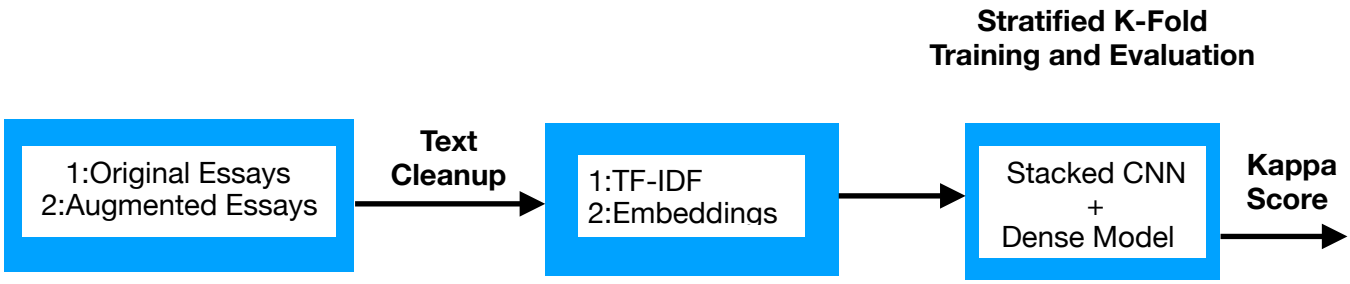
NLP data augmentation techniques applied (Courtesy: Jason Wei and Kai Zou):

Please see example table below taken from Jason Wei's blog post

1. Synonym Replacement (SR): Randomly choose n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random.
2. Random Insertion (RI): Find a random synonym of a random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence. Do this n times.
3. Random Swap (RS): Randomly choose two words in the sentence and swap their positions. Do this n times.
4. Random Deletion (RD): For each word in the sentence, randomly remove it with probability p.

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

Augmentation operations for NLP proposed in **[this paper]**. SR=synonym replacement, RI=random insertion, RS=random swap, RD=random deletion. The Github repository for these techniques can be found **[here]**.



Data Flow

Text Cleanup: Keras and NLTK routines were used for tokenization and stop words removal.

Data Augmentation: Code for Synonym Replacement, Random insertion, Random Swap and Random Delete was leveraged from here, and applied to essays with "zero" scores to increase the number of zero scoring sample essays.

Models: For all models, stratified k-fold approach was adopted for training and evaluation on both the original and the augmented dataset. Stratified K-fold ensures an even distribution of unbalanced classes in training and evaluation phase. Several models were considered including stacked CNN+stacked LSTM+Dense, stacked CNN+stacked BiLSTM+Dense, stacked LSTM+Dense and stacked BiLSTM+Dense. However the results were inferior compared to simple stacked CNN+Dense models. It appears the recurrent models do not do a good job of remembering the history as word length of essays may be too long. Tweaks to LSTM and BiLSTM models are needed as shown in the promising work in the listed papers on Neural Coherence, and Siamese BiLSTM, and Attention approaches to get better results.

Loss Function:Sparse categorical cross-entropy function was adopted as it was giving better results compared to mean-squared-error. Another benefit was the availability of confusion matrix to analyze the model prediction capability.

Kappa Score: This is a statistic that measures inter-annotator or inter-rater agreement. In this case, it is measuring the agreement between the model and the human rater. For additional information, please see https://en.wikipedia.org/wiki/Cohen%27s_kappa and https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html. The score ranges from 0 to 1. "1" - perfect agreement. "0" - no agreement.

Results:

	CNN+Dense (TF-IDF)	CNN+Dense (jointly learned embeddings)	State of art Siamese BiLSTM model	State of art feature-engineered Shallow networks
Kappa Score (original dataset)	0.48	0.51	0.80	0.66
Kappa Score (data augmented)	0.76	0.76		

Confusion Matrix (Original Dataset)

We see significant amount of mis-predictions for zero scoring essays. Out of 8 zero scoring essays, only 1 essay was correctly predicted with a zero score. Remaining “7” zero score essays were assigned a score of 1.

	Score 0	Score 1	Score 2	Score 3
Score 0	1	7	0	0
Score 1	3	78	33	7
Score 2	1	77	101	12
Score 3	0	7	64	14

Confusion Matrix (Augmented Dataset)

We see significant improvement in model performance. Out of 78 zero scoring essays, 75 were correctly predicted with a zero score. There were only 3 mis-predictions. There is also improvement seen in essays with scores of 2 and 3. None of the score 2 essays were mis-predicted with a score of 0. We also see better performance with score 3 essays with more them correctly predicted (24) with score of 3.

	Score 0	Score 1	Score 2	Score 3
Score 0	75	3	0	0
Score 1	3	68	46	4
Score 2	0	11	102	18
Score 3	0	5	56	24

Summary:
Overall the results of data augmentation with stacked CNN model architecture comes close to state of art results, and beat feature engineered models with a large margin. Stacked CNN models can perform as well as advanced models (attention, neural coherence) given a large sample size to train on. No performance differences were observed between TF-IDF and learnt word embeddings with the augmented data set.

Installation Steps:
1. Download the dataset from here.
2. Create a /data directory, and place the dataset file “training_set_rel3.tsv” in it. (This file is also included in the zip directory)
3. Execute the jupyter notebook code in finalproject.ipynb

Appendix

Essay Set #3

Type of essay:	Source Dependent Responses
Grade level:	10
Training set size:	1,726 essays
Final evaluation set size:	575 essays
Average length of essays:	150 words
Scoring:	1st Reader Score, 2nd Reader Score, Resolved CR Score
Rubric range:	0-3
Resolved CR score range:	0-3

Source Essay

ROUGH ROAD AHEAD: Do Not Exceed Posted Speed Limit
by Joe Kurmaskie

FORGET THAT OLD SAYING ABOUT NEVER taking candy from strangers. No, a better piece of advice for the solo cyclist would be, “Never accept travel advice from a collection of old-timers who haven’t left the confines of their porches since Carter was in office.” It’s not that a group of old guys doesn’t know the terrain. With age comes wisdom and all that, but the world is a fluid place. Things change.

At a reservoir campground outside of Lodi, California, I enjoyed the serenity of an early-summer evening and some lively conversation with these old codgers. What I shouldn’t have done was let them have a peek at my map. Like a foolish youth, the next morning I followed their advice and launched out at first light along a “shortcut” that was to slice away hours from my ride to Yosemite National Park.

They’d sounded so sure of themselves when pointing out landmarks and spouting off towns I would come to along this breezy jaunt. Things began well enough. I rode into the morning with strong legs and a smile on my face. About forty miles into the pedal, I arrived at the first “town.” This place might have been a thriving little spot at one time—say, before the last world war—but on that morning it fit the traditional definition of a ghost town. I chuckled, checked my water supply, and moved on. The sun was beginning to beat down, but I barely noticed it. The cool pines and rushing rivers of Yosemite had my name written all over them.

Twenty miles up the road, I came to a fork of sorts. One ramshackle shed, several rusty pumps, and a corral that couldn’t hold in the lamest mule greeted me. This sight was troubling. I had been hitting my water bottles pretty regularly, and I was traveling through the high deserts of California in June.

I got down on my hands and knees, working the handle of the rusted water pump with all my strength. A tarlike substance oozed out, followed by brackish water feeling somewhere in the

neighborhood of two hundred degrees. I pumped that handle for several minutes, but the water wouldn't cool down. It didn't matter. When I tried a drop or two, it had the flavor of battery acid.

The old guys had sworn the next town was only eighteen miles down the road. I could make that! I would conserve my water and go inward for an hour or so—a test of my inner spirit.

Not two miles into this next section of the ride, I noticed the terrain changing. Flat road was replaced by short, rolling hills. After I had crested the first few of these, a large highway sign jumped out at me. It read: ROUGH ROAD AHEAD: DO NOT EXCEED POSTED SPEED LIMIT.

The speed limit was 55 mph. I was doing a water-depleting 12 mph. Sometimes life can feel so cruel.

I toiled on. At some point, tumbleweeds crossed my path and a ridiculously large snake—it really did look like a diamondback—blocked the majority of the pavement in front of me. I eased past, trying to keep my balance in my dehydrated state.

The water bottles contained only a few tantalizing sips. Wide rings of dried sweat circled my shirt, and the growing realization that I could drop from heatstroke on a gorgeous day in June simply because I listened to some gentlemen who hadn't been off their porch in decades, caused me to laugh.

It was a sad, hopeless laugh, mind you, but at least I still had the energy to feel sorry for myself. There was no one in sight, not a building, car, or structure of any kind. I began breaking the ride down into distances I could see on the horizon, telling myself that if I could make it that far, I'd be fi ne.

Over one long, crippling hill, a building came into view. I wiped the sweat from my eyes to make sure it wasn't a mirage, and tried not to get too excited. With what I believed was my last burst of energy, I maneuvered down the hill.

In an ironic twist that should please all sadists reading this, the building—abandoned years earlier, by the looks of it—had been a Welch's Grape Juice factory and bottling plant. A sandblasted picture of a young boy pouring a refreshing glass of juice into his mouth could still be seen.

I hung my head.

That smoky blues tune "Summertime" rattled around in the dry honeycombs of my deteriorating brain.

I got back on the bike, but not before I gathered up a few pebbles and stuck them in my mouth. I'd read once that sucking on stones helps take your mind off thirst by allowing what spit you have left to circulate. With any luck I'd hit a bump and lodge one in my throat.

It didn't really matter. I was going to die and the birds would pick me clean, leaving only some expensive outdoor gear and a diary with the last entry in praise of old men, their wisdom, and their keen sense of direction. I made a mental note to change that paragraph if it looked like I was going to lose consciousness for the last time.

Somehow, I climbed away from the abandoned factory of juices and dreams, slowly gaining elevation while losing hope. Then, as easily as rounding a bend, my troubles, thirst, and fear were all behind me.

GARY AND WILBER'S FISH CAMP—IF YOU WANT BAIT FOR THE BIG ONES, WE'RE YOUR BEST BET!

"And the only bet," I remember thinking.

As I stumbled into a rather modern bathroom and drank deeply from the sink, I had an overwhelming urge to seek out Gary and Wilber, kiss them, and buy some bait—any bait, even though I didn't own a rod or reel.

An old guy sitting in a chair under some shade nodded in my direction. Cool water dripped from my head as I slumped against the wall beside him.

"Where you headed in such a hurry?"

"Yosemite," I whispered.

"Know the best way to get there?"

I watched him from the corner of my eye for a long moment. He was even older than the group I'd listened to in Lodi.

"Yes, sir! I own a very good map."

And I promised myself right then that I'd always stick to it in the future.

Prompt

Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.

Rubric Guidelines

Score 3: The response demonstrates an understanding of the complexities of the text.

- Addresses the demands of the question
- Uses expressed and implied information from the text
- Clarifies and extends understanding beyond the literal

Score 2: The response demonstrates a partial or literal understanding of the text.

- Addresses the demands of the question, although may not develop all parts equally
- Uses some expressed or implied information from the text to demonstrate understanding
- May not fully connect the support to a conclusion or assertion made about the text(s)

Score 1: The response shows evidence of a minimal understanding of the text.

- May show evidence that some meaning has been derived from the text
- May indicate a misreading of the text or the question
- May lack information or explanation to support an understanding of the text in relation to the question

Score 0: The response is completely irrelevant or incorrect, or there is no response.

Adjudication Rules

- If Reader-1 Score and Reader-2 Score are exact or adjacent, adjudication by a third reader is not required.
- If Reader-1 Score and Reader-2 Score are not adjacent or exact, then adjudication by a third reader is required.