Essay Quality Signals as Weak Supervision for Source-based Essay Scoring

Haoran Zhang
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
colinzhang@cs.pitt.edu

Diane Litman
Department of Computer Science & LRDC
University of Pittsburgh
Pittsburgh, PA 15260
dlitman@pitt.edu

Abstract

Human essay grading is a laborious task that can consume much time and effort. Automated Essay Scoring (AES) has thus been proposed as a fast and effective solution to the problem of grading student writing at scale. However, because AES typically uses supervised machine learning, a human-graded essay corpus is still required to train the AES model. Unfortunately, such a graded corpus often does not exist, so creating a corpus for machine learning can also be a laborious task. This paper presents an investigation of replacing the use of human-labeled essay grades when training an AES system with two automatically available but weaker signals of essay quality: word count and topic distribution similarity. Experiments using two source-based essay scoring (evidence score) corpora show that while weak supervision does not yield a competitive result when training a neural source-based AES model, it can be used to successfully extract Topical Components (TCs) from a source text, which are required by a supervised feature-based AES model. In particular, results show that feature-based AES performance is comparable with either automatically or manually constructed TCs.

1 Introduction

Essay grading is a laborious task that can consume much teacher time and effort, especially for classes with a large number of students. While human essay grading has high reliability and validity, human grading can also be biased and inconsistent over time. Under such conditions, Automatic Essay Scoring (AES) has been proposed as a fast and effective solution to the problem of grading student writing at scale, minimizing effort and assuring consistency. Many AES systems have been developed to evaluate the content, structure, and quality of written essays via natural language processing (NLP) and machine-learning techniques.

Over the more than 50 year history of AES research, the majority of work has used feature-based models (Louis and Higgins, 2010; Persing et al., 2010; Yannakoudakis and Briscoe, 2012; Persing and Ng, 2015; Farra et al., 2015; McNamara et al., 2015; Cummins et al., 2016; Ghosh et al., 2016; Nguyen and Litman, 2018; Amorim et al., 2018; Cozma et al., 2018). However, these models require carefully designed hand-crafted features to represent essays. Recently, neural network models have drawn increasing attention (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Farag et al., 2018; Jin et al., 2018; Li et al., 2018; Tay et al., 2018; Nadeem et al., 2019; Mayfield and Black, 2020) due to their powerful ability to generate essay representations automatically and to generate reliable essay score predictions.

No matter whether a feature-based or neural network model is used, state-of-the-art AES systems are all supervised, which means that the model needs to be trained on a large number of human-graded essays. Unfortunately, such a human-graded corpus often does not exist, and grading a corpus of essays is a laborious task. To address this problem, we investigate using a weakly supervised AES approach, where automatically available essay quality signals replace the use of human-labeled scores when training a state-of-the-art neural network model for source-based essay scoring. We define ***essay quality signals*** in this work as scores that reflect the quality of essays. The human-labeled holistic score is the most common essay quality signal, and is typically used for supervised machine learning. We instead explore the use of two automated essay quality signals, namely word count and topic distribution similarity between student essays and a source article, in a weakly supervised machine learning AES approach.

Our work focuses on the response-to-text assess-

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 85–96
April 20, 2021

ment (RTA) (Correnti et al., 2013), a particular source-based essay writing task. In source-based writing, students read a source article before writing an essay that responds to a prompt related to the source article. In order to guide students on revising an RTA essay after receiving feedback, an automated writing evaluation (AWE) system for the RTA (Zhang et al., 2019) was developed that generated formative feedback based on rubric-based AES features. These AES features were computed by recognizing when a set of manually created ***Topical Components*** (TCs) from the source text were used by students in their essays. The TCs represent a comprehensive list of topics related to evidence in the source article, which include: 1) important keywords indicating the set of topics (denoted as *topic words*), and 2) phrases representing specific examples for each topic (denoted as *specific examples*). The grading rubric encourages students to mention and elaborate topics and specific examples from the source article as much as possible in their essays. Therefore, the RTA AES and AWE systems use rubric-based features based on the TCs in order to follow the grading rubric.

Previously, human effort was required to manually create TCs based on expert knowledge of the source text. To eliminate this effort, Zhang and Litman (2020) used the attention layer output of a co-attention neural network to automatically extract TCs. Their experimental results showed that automated TCs performed equally well as manual TCs on a downstream, rubric-based AES task. However, to eliminate the human effort for creating TCs, they introduced a different human evidence score grading effort to train the co-attention neural network. Unfortunately, we believe that the effort to grade student essays is more than the original effort to create the TCs. Our goal is to remove both sources of human effort.

To do so, we introduce two simple essay quality signals, word count and topic distribution similarity, which can be generated automatically and used as weak supervision for training the neural AES. Although the learned AES model outperforms simple baselines, weak supervision is not enough to yield a state-of-the-art AES model. Nonetheless, the proposed essay quality signals can be successfully used to generate TCs for a downstream rubric-based AES task. By using auto-generated essay quality signals, we can thus eliminate all human effort for generating TCs. We evaluate the generated TCs using a rubric-based AES requiring TCs, for two RTA source articles. Results show that our simple signals for training a neural AES to create TCs automatically do not degrade performance compared to using manually constructed TCs, whether for 1) representing essays as rubric-based features, 2) grading essays.

2 Related Work

The majority of research in the AES area uses supervised machine learning techniques that require a large number of human-graded essays for training. However, graded essay corpora are usually missing in real classroom scenarios, and annotating a corpus to train an AES model is labor-intensive. A prior proposal to address this problem used an unsupervised-learning approach based on a voting algorithm (Chen et al., 2010). The area of short answer scoring has also faced a similar problem. Zesch et al. (2015a) presented a semi-supervised method to reduce the size of the required human-labeled corpus. Meanwhile, Ramachandran et al. (2015) proposed a graph-based lexico-semantic text matching for pattern identification. These works reduce human effort, but do not eliminate them. In contrast, our AES work fully replaces human graded evidence scores with essay quality signals that are easy to extract automatically and to use during training. Although our results show that the signals are not effective for the AES task itself, they are useful for extracting Topical Components (TCs).

Previously, human expert effort was required to extract TCs. Specifically, experts read through the source article and created lists of topic words and of specific examples that students were expected to use in their essays (Rahimi et al., 2017). In order to eliminate this human effort, three systems were later developed. An LDA-based system (Rahimi and Litman, 2016) used a LDA topic model (Blei et al., 2003) and TurboTopic algorithm (Blei and Lafferty, 2009) for TC extraction. Zhang and Litman (2020) proposed another system based on the PositionRank (Florescu and Caragea, 2017) algorithm. While these two TC extraction systems did not require any human coding, they also did not match prior performance. The state-of-the-art system (Zhang and Litman, 2020) extracted TCs by exploiting the attention weights of a neural AES model. However, human grading effort was needed for model training. In our work, we replace hu-

Prompt	RTA_{MVP}	RTA_{Space}
Score 1	852	538
	(29%)	(26%)
Score 2	1197	789
	(40%)	(38%)
Score 3	616	512
	(21%)	(25%)
Score 4	305	237
	(10%)	(11%)
Total	2970	2076

Table 1: The Evidence score distribution of RTA.

man scores with automated essay quality signals for training, while still achieving state-of-the-art TC extraction.

We believe that many predictive features used in the traditional feature-based AES systems can be useful signals for our weak supervision approach to TC extraction. For example, length-based features (Attali and Burstein, 2006; Chen and He, 2013; Östling et al., 2013; Phandi et al., 2015; Zesch et al., 2015b), prompt-relevant features (Louis and Higgins, 2010; Klebanov et al., 2016), or semantic features (Klebanov and Flor, 2013; Persing and Ng, 2013) all weakly relate to the quality of an essay's content. In this paper, we examine two such signals, word count and topic distribution similarity, and show that with these simple essay quality signals, human-labeled essay scores are unnecessary for TC extraction.

3 Corpora

Our work focuses on a specific source-based essay writing task named the response-to-text assessment (RTA) (Correnti et al., 2013). The corpora were all collected from upper elementary level classrooms. Based on different articles that students read before writing their essays, two forms of the RTA were developed. RTA_{MVP} is based on an article from *Time for Kids* about an effort by the United Nations to end poverty in a rural village in Sauri, Kenya, named the Millennium Villages Project (MVP). RTA_{Space} is based on an article which discusses the importance of space exploration. After reading an RTA article, there is a prompt that asks students to write an essay in response to the prompt. The prompt encourages students to use evidence from the source article to support their claims. Table 2 shows a source article excerpt, the corresponding prompt, and a student essay of RTA_{MVP}. The bolded phrases in the source excerpt are evidence examples manually labeled by human experts, which are essential to be discussed in the student essay.

All collected essays (2970 essays for RTA_{MVP} and 2076 essays for RTA_{Space}) were manually graded on a scale of 1 to 4 (low to high) in five dimensions: Analysis, Evidence, Organization, Style, and MUGS (Mechanics, Usage, Grammar, and Spelling). Each dimension evaluates one aspect of students' writing skills. We focus on the Evidence dimension , which evaluates students' ability to find and use evidence from the source article to support their ideas. This dimension is also directly related to the final output of this work, Topical Components. Table 1 shows the distribution of Evidence scores for both forms of the RTA. Table 2 shows an essay with an evidence score of 3. Phrases semantically related to the TCs from the source article are shown in **boldface**. The average essay word counts for each prompt are 180 and 220, respectively.

4 Prior AES Systems for the RTA

Two approaches to AES have been developed for the RTA: 1) a feature-based supervised learning approach (with the features aligned to the Evidence rubric criteria), which we refer to as AES_{rubric}, and 2) a neural network approach, which we refer to as AES_{neural}.

AES_{rubric} (Rahimi et al., 2017; Zhang and Litman, 2017) used a traditional supervised learning framework with hand-crafted, rubric-based features that require knowledge of TCs to compute. That is, a set of interpretable features were carefully designed to capture the relation between student essays and the two aspects of TCs described above (*topic words* and *specific examples*):

Number of Pieces of Evidence (NPE): an integer feature based on the list of *topic words* for each topic that indicates the number of topics (semantically) mentioned in the essay.

Concentration (CON): a binary feature that indicates if an essay elaborates on topics, again based on the list of *topic words*.

Specificity (SPC): a vector of integer values indicating the number of *specific examples* (semantically) mentioned in the essay per topic.

Word Count (WOC): number of words.

To support the needs of the AWE system for the RTA (Zhang et al., 2019), the feature values and predicted evidence scores from AES_{rubric} are passed to the AWE system to select formative feed-

Source Excerpt: Today, Yala Sub-District **Hospital has medicine**, **free of charge**, **for all of the most common diseases**. **Water is connected to the hospital**, which also has a **generator for electricity**. **Bed nets are used** in every sleeping site in Sauri...
Essay Prompt: The author provided one specific example of how the quality of life can be improved by the Millennium Villages Project in Sauri, Kenya. Based on the article, did the author provide a convincing argument that winning the fight against poverty is achievable in our lifetime? Explain why or why not with 3-4 examples from the text to support your answer.
Essay: In my opinion I think that they will **achieve it in lifetime**. During the years threw **2004 and 2008 they made progress**. People didn't have the money to buy the stuff in 2004. **The hospital was packed with patients** and they didn't have alot of treatment in 2004. In 2008 it changed the **hospital had medicine**, **free of charge**, and **for all the common dieases**. **Water was connected to the hospital** and has a **generator for electricity**. **Everybody has net** in their site. **The hunger crisis has been addressed** with **fertilizer and seeds**, as well as the **tools needed to maintain the food**. **The school has no fees** and **they serve lunch**. To me that's sounds like it is going achieve it in the lifetime.

Table 2: A source excerpt for the RTA_{MVP} prompt and an essay with an evidence score of 3.

Topic	Keywords
Hospital	care, health, hospital, treatment, doctor, electricity, disease, water, ...
Malaria	bed, net, malaria, infect, bednet, mosquito, bug, sleeping, die, cheap, ...
Farming	farmer, fertilizer, irrigation, dying, crop, seed, water, harvest, hungry, ...
School	school, supplies, fee, student, midday, meal, lunch, supply, book, paper, ...

Table 3: The partial list of topic words of RTA_{MVP}.

back messages such as "Use more evidence from the article" (based on NPE values) or "Provide more details for each piece of evidence you use" (based on NPE and SPC values).

Although AES_{rubric} thus provides useful information for the AWE system, in order to improve the stand-alone AES performance, AES_{neural} (Zhang and Litman, 2018) was later developed. AES_{neural} used a hierarchical neural network model with a self-attention mechanism in the phrase level and a co-attention mechanism in the sentence level. The self-attention layer captures the importance of each individual phrase, while the co-attention layer captures the relationship between the source article and the essay. In terms of the essay score prediction task, AES_{neural} significantly outperforms AES_{rubric}, and no human effort for either TC extraction or feature engineering is required. However, the essay representations created by AES_{neural} cannot be directly used by the AWE system, which depends on the rubric-based features to provide formative feedback in terms of the grading criteria.

Category 1	Category 5	Category 7
unpaved roads	crops dying	progress just four years
tattered clothing	not afford fertilizer irrigation	medicine free charge
bare feet	outcome poor crops	no midday meal lunch
less than 1 dollar day	lack fertilizer water	kids go school now
	...	...

Table 4: The partial list of specific examples of RTA_{MVP}.

5 Prior TC Extraction Methods

To develop AES_{rubric}, human expert effort was first required to manually extract TCs (TC_{manual}) in the form of two lists related to evidence in the source text: 1) a *topic words list* of important keywords that indicate the main set of article topics, and 2) a *specific examples* list that includes phrases representing specific examples for article topics. Table 3 shows a partial topic words list for RTA_{MVP}, where the four topics ("hospital", "malaria", "farming", and "school") and the associated keywords were manually created by a human expert. Table 4 shows a partial specific examples list for RTA_{MVP}. The full list has 8 categories. Some categories are similar to Category 1, which is not related to the 4 main topics, but the human expert thought they were important to be mentioned in the essay. Other categories are similar to Category 5, in being directly related to one of the main topics. Other categories are similar to Category 7, in being directly related to multiple main topics.

To replace the need for the human expert in creating such TCs, Zhang and Litman (2020) developed a method for TC extraction using AES_{neural}. Their algorithm was based on the observations that the co-attention layer on the sentence level assigned higher attention scores to important sentences, while the self-attention layer on the word (phrase) level assigned higher attention scores to important words (phrases). Therefore, their system extracted important words from important

Prompt	WC	TDS
RTA_{MVP}	0.480	0.359
RTA_{Space}	0.489	0.253

Table 5: Pearson's r comparing different essay quality signals with evidence score.

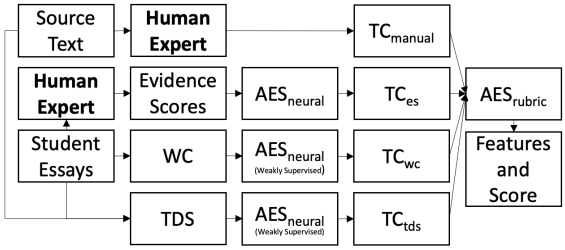

Figure 1: An overview of four TC extraction systems.

sentences based on attention scores and used k-medoids to cluster all words. Finally, it extracted TCs from each cluster. Since human-labeled evidence scores of each essay were required for the neural network training, we denote this method by TC_{es}. Note that TC_{es} replaced the human effort needed to extract TCs with the human effort needed to create the AES_{neural} training supervision signal.

6 Weak Essay Quality Signals

Currently, TC_{es} reaches the top performance for automated TC extraction (Zhang and Litman, 2020) when compared to the LDA-based and PositionRank methods discussed in the Related Work section. However, TC_{es} requires extra human effort for essay grading, a barrier to making the system useful in real classroom scenarios. Therefore, in this work, we aim to explore essay quality signals other than gold-standard evidence scores in order to eliminate the remaining human effort in the TC extraction process.

6.1 Word Count (WC)

Most intuitively, word count is usually highly positively correlated with essay quality, especially with the holistic score of an essay. Most feature-based AES systems use word count as one of the features (Attali and Burstein, 2006; Chen and He, 2013; Östling et al., 2013; Phandi et al., 2015; Zesch et al., 2015b). Since the word count is highly predictive of essay score on its own, some models even manually assign a lower weight to this feature to prevent it from dominating the final model (Burstein et al., 2004). Therefore, we believe the word count is a good indicator of overall essay quality. In addition, per the grading rubric of the evidence dimension, an essay with a higher evidence score should mention more topics and elaborate more specific examples. Therefore, we also believe that the word count should be correlated with the RTA evidence score as well. Table 5 shows that the correlations between word count and evidence score on our two corpora are 0.480 and 0.489 for RTA_{MVP} and RTA_{Space}, respectively.

6.2 Topic Distribution Similarity (TDS)

Although the LDA-based TC extraction system (Rahimi and Litman, 2016) did not outperform TC_{es} on a downstream AES_{rubric} task, the generated TCs still seemed to be of reasonable quality. One possible reason is that the quality of the LDA model trained on student essays is good enough to extract important information. Since an essay with a higher evidence score should mention topics and specific examples from the source article as much as possible, we hypothesize that the topic distribution of a good essay should be similar to the source article. Therefore, the second weak essay quality signal we explore is the similarity between the topic distribution of the student essay and the source article.

More specifically, we first train an LDA model on both student essays and the source article. We believe that including the source article into the LDA training process will provide more information to learn from, even if the influence is minor. We then use the LDA model to infer the topic distribution of each essay and the source article. Finally, we calculate the similarity between the topic distribution of a student essay and the source article as the essay's quality signal for the proposed weakly-supervised approach for co-attention neural network training.

Since LDA is an unsupervised method and it is hard to know how many topics exist in a corpus, we use the Topic Coherence score (Röder et al., 2015) to select the best number of topics in an automated manner. Topic Coherence measures whether a topic is semantically interpretable by computing the semantic similarity between important words in the topic. We use C_V measurement because it reaches the best performance in the original paper. C_V measurement is based on a sliding window, and combines the indirect cosine measure with the normalized pointwise mutual information (NPMI).

Layer	Parameter Name	Value
Embedding	Embedding dimension	50
Word-CNN	Kernel size	5
	Number of filters	100
Sent-LSTM	Hidden units	100
Modeling	Hidden units	100
Dropout	Dropout rate	0.5
Others	Epochs	100
	Batch size	100
	Initial learning rate	0.001
	Momentum	0.9

Table 6: Hyper-parameters for neural training.

Prompt	Component	Parameter	TC_{es}	TC_{wc}	TC_{tds}
RTA_{MVP}	Topic Words	# of Topics	16	13	5
		# of Words	25	10	25
	Examples	# of Topics	18	14	20
		# of Examples	15	15	30
RTA_{Space}	Topic Words	# of Topics	10	18	16
		# of Words	20	30	25
	Examples	# of Topics	9	19	16
		# of Examples	20	15	20

Table 7: Selected parameters for different models.

Since a good topic model should have as many semantically interpretable topics as possible, a good topic model should receive high topic coherence scores. We train multiple LDA models with different numbers of topics for each individual form of RTA, and select the number of topics resulting in the best coherence scores. The best number of topics for RTA_{MVP} is 7, and the best number of topics for RTA_{Space} is 14.

Once we use the pre-trained LDA model to infer topic distributions for each essay and the source article, we calculate the similarity between them to get topic distribution similarity. We select cosine similarity rather than dot product similarity since the grading rubric encourages students to mention more topics rather than go deep into one topic. A full elaboration of evidence is only required for essays with a high evidence score, although the rubric encourages all students to elaborate evidence as much as possible. Therefore, in geometrical terms, we care about angle difference more than magnitude difference. In other words, we measure how many topics from the source article are mentioned in an essay. Table 5 shows that the correlations between topic distribution similarity and evidence scores are 0.359 and 0.253 for RTA_{MVP} and RTA_{Space}, respectively.

7 Experimental Setup

Figure 1 shows an overview of usage of AES_{neural} and four TC extraction systems to be evaluated. TC_{manual} lets human experts extract TCs from each source article, and is thus the upper bound for evaluating the other (automated) TC extraction systems. TC_{es} is our baseline automated model, which builds on AES_{neural} and a clustering algorithm to extract TCs from student essays and the source article, using the gold-standard evidence score of each essay for AES_{neural} training. TC_{wc} and TC_{tds} are methods proposed by this work that are instead based on weakly-supervised AES_{neural} training. TC_{wc} replaces evidence score with the word count of each essay, while TC_{tds} uses topic distribution similarity with the number of topics.

Our experiments are designed to test two hypotheses related to the alternative AES and TC methods shown in Figure 1. The first hypothesis, denoted by H1, is that while weakly supervised training might not yield state-of-the-art, AES_{neural} performance when evaluated as an end in itself, the use of automated essay quality signals nonetheless can outperform weak baselines such as random and majority score prediction. Our second hypothesis, denoted by H2, is that weakly supervised training can nonetheless yield versions of AES_{neural} that are still useful for automated TC extraction. All neural network models are built with TensorFlow 2.2.0, and trained on an RTX 5000 GPU.

7.1 $\text{AES}_{\text{neural}}$ Performance (H1)

Our experiment for H1 tests the impact of replacing human-labeled evidence scores with our proposed weak essay quality signals when training the AES_{neural} model. Specifically, we train AES_{neural} models on human-labeled evidence score, word count, and topic distribution similarity. Then, we calculate the Quadratic Weighted Kappa (QWK) between predicted scores of AES_{neural} and human evidence scores. We also compare these scoring results to random and majority prediction baselines.

Following Zhang and Litman (2018), we use 5-fold cross-validation in this experiment. We split both corpora into 5 folds, the partition is the same as the setting presented by Zhang and Litman (2018). In each fold, 60% of the data are used for training, 20% of the data are the development set, and 20% of the data are used for testing. All essay

Prompt	Majority (1)	Random (2)	Evidence Score (3)	WC (4)	TDS (5)
RTA_{MVP}	0.000	0.016	**0.697 (1,2,4,5)**	0.366 (1,2)	0.440 (1,2)
RTA_{Space}	0.000	0.016	**0.684 (1,2,4,5)**	0.380 (1,2)	0.386 (1,2)

Table 8: The performance (QWK) of AES_{neural} using different essay quality signals for training. The numbers in the parentheses show the model numbers over which the current model performs significantly better ($p \leq 0.05$). The best results in each row are in bold.

scores and WC are scaled to the range [0, 1] for training. Since TDS is topic distribution similarity between essays and the source article, the score range is [0, 1], so we do not scale it in the experiments. The scaled essay scores or essay quality signals are only used for training. When calculating the Quadratic Weighted Kappa, we scaled the predicted score back to the original score range, which is [1, 4]. All hyper-parameters for the AES_{neural} training are shown in Table 6. Please note that we used essay scores for the development set to determine early stopping. Therefore, we assume a small amount of graded essays is available for this purpose in experiment of H1. Since all hyper-parameters are inherited from prior work, they are not selected based on the development set.

7.2 Extracted TCs (H2)

We configure experiments to evaluate the four TC extraction methods in Figure 1 both extrinsically and intrinsically. We thus break H2 into two sub-hypotheses: H2a) the AES_{rubric} model for scoring Evidence will perform comparably when extracting features using TC extraction methods involving either human (TC_{manual}, TC_{es}) or automated (TC_{wc}, TC_{tds}) methods; H2b) the correlation between the human evidence score and the TC-dependent feature values will be comparable when extracting features using either TC_{manual}, TC_{es}, TC_{wc}, and TC_{tds}. Extrinsically, the experiment for H2a examines the impact of using our proposed TC extraction methods on the downstream AES_{rubric} task. Intrinsically, the experiment for H2b measures the impact on the essay representation itself. For H2a, we calculate the Quadratic Weighted Kappa (QWK) between predicted scores of AES_{rubric} and human evidence scores. For H2b, we compare the correlation between human evidence score with NPE feature and sum of SPC features, because both features are integer features and are extracted based on TCs.

For these experiments, we stratify essay corpora following Zhang and Litman (2020): 40% for training word embeddings and extracting TCs, 20% for selecting the best embedding and parameters, and 40% for testing. We use the same hyper-parameters from Zhang and Litman (2018) for the co-attention neural network training as shown in Table 6. Table 7 show all other parameters selected using the development sets for all models.

8 Results and Discussion

Table 8, which addresses H1, shows the Quadratic Weighted Kappa between human evidence scores and predicted scores by AES_{neural} using different essay quality signals for training, as well as random prediction and majority prediction. Unsurprisingly, the models trained on human scores significantly outperform our proposed weaker essay quality signals on both prompts, even though we assume a small amount of graded essays is available for early stopping. Although QWK of WC and TDS are lower than Evidence Score, they still significantly outperform random and majority prediction baselines. The results support H1 that while weak supervision signals such as word count and topic distribution similarity are not enough for training AES_{neural} to reach a state-of-the-art QWK, they still provide some predictive utility.

Although both WC and TDS underperform the human-generated Evidence Scores, TDS constantly outperforms WC, despite the fact that WC has higher correlations with Evidence Score than TDS (recall Table 5). One possible reason is that the human evidence score assesses if an essay mentions and elaborates evidence from the source article, which measures the relationship between the essay and the source article. TDS is topic distribution similarity between student essays and the source article, so the AES model learns more relations between student essays and the source article. However, WC only contains length information of essays but no relation between essays and the source article.

Table 9, which addresses H2a, shows the Quadratic Weighted Kappa between human evidence scores and predicted scores by AES_{rubric} when using different TCs. On RTA_{MVP}, TC_{wc}

Prompt	TC_{manual} (1)	TC_{es} (2)	TC_{wc} (3)	TC_{tds} (4)
RTA_{MVP}	0.643	0.648 (1)	0.645	**0.652 (1,2,3)**
RTA_{Space}	0.609 (4)	**0.622 (1,4)**	**0.622 (1,4)**	0.599

Table 9: The performance (QWK) of AES_{rubric} using different TCs extraction methods for feature creation. The numbers in the parentheses show the model numbers over which the current model performs significantly better ($p < 0.05$). The best results between automated methods in each row are in bold.

Prompt	Feature	TC_{manual}	TC_{es}	TC_{wc}	TC_{tds}
RTA_{MVP}	NPE	0.542	**0.639**	**0.560**	0.533
	SPC (sum)	0.689	0.679	0.653	0.674
RTA_{Space}	NPE	0.484	**0.625**	**0.615**	**0.599**
	SPC (sum)	0.601	0.598	0.485	0.438

Table 10: Pearson's r comparing feature values computed using each TCs extraction method with human (gold-standard) Evidence essay scores. All correlation values are significant ($p \leq 0.05$). Bolding indicates that the automated method is better than TC_{manual}.

yields statistically similar performance compared to TC_{manual} and TC_{es}, while TC_{tds} significantly outperforms all other methods. The story is different when switching to RTA_{Space}, where TC_{tds} is now outperformed by all other methods. Considering that the two proposed methods based on weak supervision do not require human expert effort for either TC extraction (TC_{manual}) or for grading evidence score for neural training (TC_{es}), we believe the results support H2a.

Table 10 shows Pearson's r comparing feature values computed using each TC extraction method with human (gold-standard) Evidence essay scores, and partially support H2b. For NPE feature, TC_{wc} always yields better performance than TC_{manual}. TC_{tds} yields better performance than TC_{manual} on RTA_{Space} only. However, for SPC features, there is no automated method that outperforms TC_{manual}. On RTA_{MVP}, the proposed methods yield similar performance as TC_{es}.

A very interesting finding is that both WC and TDS underperform Human Score on AES_{neural} task, while TC_{wc} and TC_{tds} help AES_{rubric} reach an even higher QWK. This result shows that while learning using weak supervision is not enough for AES_{neural} training, with post-processing the intermediate output, the neural predictions can still help to generate useful TCs for the AES_{rubric} task.

Since word count is highly positively correlated with evidence score for both RTA_{MVP} and RTA_{Space}, TC_{wc} works well on average compared to TC_{tds}. Extrinsically, it outperforms TC_{manual} on both corpora. It also matches TC_{es} on RTA_{Space}, and has similar performance on RTA_{MVP}. Intrinsically, TC_{wc} yields higher correlations for the NPE feature when comparing to TC_{manual}. Although correlations for SPC are worse than TC_{es}, considering word count is the most intuitive signal without the needs of human effort, it performs surprisingly well.

Moving to topic distribution similarity, TC_{tds} shows worse extrinsic performance on RTA_{Space} comparing to RTA_{MVP}. To figure out the reason, we take a deep dive into the TCs generated by both methods. We consider that good automated TCs should cover topics in TC_{manual} as many as possible. Therefore, we manually label a topic for each of the manual topic words. For RTA_{MVP}, TC_{tds} has 4 related topics out of 5 (**80%**), while there are 10 related topics out of 16 (62.50%) for RTA_{Space}. Obviously, TC_{tds} preserves more related topics in RTA_{MVP}. Similarly, we also manually compare specific examples of both automated TCs with TC_{manual}. For examples rather than keywords, TC_{tds} has 16 out of 20 related categories (**80%**) for RTA_{MVP}, while there are 11 out of 16 related categories (68.75%) for RTA_{Space}. TC_{tds} again preserves more related categories in RTA_{MVP}.

We also observe that we can always find a better QWK using an automated TC method compared to TC_{manual} (Table 9). It is typically assumed that humans are the upper bound, but they do not seem to do an optimal job when creating TCs. One possible reason is that the human expert is subjective when creating TCs, and they add words and examples they thought necessary. However, some words or examples may not be as important as humans thought. Meanwhile, AES_{neural} is objective. TCs generated by TC_{es}, TC_{wc}, and TC_{tds} directly extract important words and examples that

AES_{neural} considers essential, and they are highly related to its essay score or essay quality signals. Therefore, TC_{es}, TC_{wc}, and TC_{tds} are more suitable for AES_{rubric}, which heavily relies on feature values extracted based on TCs.

A concern that might be raised about our work is that our essay quality signals have the potential to be gamed by students. For example, our signals could assign high scores to long essays with no relation to the prompt or to essays with replicated words of the source article. However, in our experiments, both essay quality signals are only used for training AES_{neural} and extracting Topical Components. For experiments of H1, essay quality signals are used as the label for training the AES_{neural} model instead of used as features during testing (when gaming would be expected). The trained AES_{neural} is in fact tested on the human-graded evidence scores, not word count nor topic distribution similarity. For experiments of H2, essay quality signals are only used as the label for training AES_{neural} and extracting Topical Components. AES_{rubric} extracts features based on the Topical Components, and are then trained and tested on evidence scores. The purpose of this experiment is to show that the Topical Components extracted from the AES_{neural} are useful for another downstream AES model, even though the QWK of the AES_{neural} trained on essay quality signals is low. In sum, since essay quality signals are not used as predictive features in our system and only used as the label for training, we believe that our proposed method for Topical Components extraction is hard to be gamed – assuming adverserial essays only appear in the testing set. However, if such essays appear in the training data the risk of the manipulation of both AES_{neural} and AES_{rubric} based on word count, topic distribution similarity, or other essay quality signals is still possible (Lochbaum et al., 2013). A future deep analysis of both AES models would be necessary to address this problem.

Another limitation of our research is that to date, our approach has only been tested on two RTA corpora. This is because Topical Components have only been used in AES_{rubric} until now. Furthermore, although Word Count shows a high correlation with evidence scores in our RTA corpora, this high correlation could easily disappear in tasks in which the prompt specifies the expected word count (Weiss et al., 2019). Similarly, topic distribution similarity might not a good essay quality signal if a prompt is not source-based, or does not expect similar topics in content.

9 Conclusion and Future Work

This paper presented an investigation of replacing human-labeled evidence scores with other automated essay quality signals, such as word count and topic distribution similarity. These signals are easy to be calculated and integrated into existing systems in order to eliminate human effort. Not surprisingly, these weak supervised signals are not enough for training a useable AES_{neural} model. However, they still help generate TCs, which is required by AES_{rubric}. We observe that even a simple signal like word count does not hurt the state-of-the-art baseline (TC_{es}). Since there is no need for human effort, we believe that our work brings AES technology closer to being useful in real classroom scenarios.

In our future work on using weak supervision for AES training, we would like to explore other possible essay quality signals beyond the two investigated here, again drawing on machine learning features used in prior work. It would also be interesting to examine whether our two existing signals might yield better AES results when training other types of neural algorithms and/or when applied to different datasets. Moreover, the impact of data size for training on this model is worth exploration. Finally, with respect to TC generation, because the specific examples are generated from clustering results, words in a specific example are not in readable orders. This leads to another interesting future investigation into making all examples in the specific examples list more human-understandable, although this would not affect the system performance due to the nature of the AES_{rubric}.

Acknowledgments

We would like thank members of the PETAL group as well as reviewers for comments on an earlier version of the paper. We would also like to thank Adriana Kovashka for originally suggesting that we explore weak supervision.

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160245 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

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 (Volume 1: Long Papers)*, volume 1, pages 715–725.

Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237.

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

David M Blei and John D Lafferty. 2009. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *Ai magazine*, 25(3):27–27.

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.

Yen-Yu Chen, Chien-Liang Liu, Chia-Hoang Lee, Tao-Hsing Chang, et al. 2010. An unsupervised automated essay-scoring system. *IEEE Intelligent systems*, 25(5):61–67.

Richard Correnti, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang. 2013. Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.

Madalina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509.

Ronan Cummins, Meng Zhang, and Edward John Briscoe. 2016. Constrained multi-task learning for automated essay scoring. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 263–271.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74.

Corina Florescu and Cornelia Caragea. 2017. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Beata Beigman Klebanov and Michael Flor. 2013. Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1148–1158.

Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016. Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72.

Xia Li, Minping Chen, Jianyun Nie, Zhenxing Liu, Ziheng Feng, and Yingdan Cai. 2018. Coherence-based automated essay scoring using self-attention. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 386–397. Springer.

Karen E Lochbaum, Mark Rosenstein, Peter Foltz, Marcia A Derr, et al. 2013. Detection of gaming in automated scoring of essays with the iea. In *National Council on Measurement in Education Conference (NCME)*.

Annie Louis and Derrick Higgins. 2010. Off-topic essay detection using short prompt texts. In *proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 92–95.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

Danielle S McNamara, Scott A Crossley, Rod D Roscoe, Laura K Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 484–493.

Huy V Nguyen and Diane J Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Robert Östling, Andre Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439.

Zahra Rahimi and Diane Litman. 2016. Automatically extracting topical components for a response-to-text writing assessment. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 277–282.

Zahra Rahimi, Diane Litman, Richard Correnti, Elaine Wang, and Lindsay Clare Matsumura. 2017. Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, 27(4):694–728.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers. 2019. Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of german essays. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 30–45.

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. Association for Computational Linguistics.

Torsten Zesch, Michael Heilman, and Aoife Cahill. 2015a. Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015b. Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232.

Haoran Zhang and Diane Litman. 2017. Word embedding for response-to-text assessment of evidence. In *Proceedings of ACL 2017, Student Research Workshop*, pages 75–81.

Haoran Zhang and Diane Litman. 2018. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop*

on Innovative Use of NLP for Building Educational Applications, pages 399–409.

Haoran Zhang and Diane Litman. 2020. Automated topical component extraction using neural network attention scores from source-based essay scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8569–8584.

Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsmura, Emily Howe, and Rafael Quintana. 2019. erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9619–9625.