Exploring Automated Essay Scoring Models for Multiple Corpora and Topical Component Extraction from Student Essays

by

Haoran Zhang

B.S. in Computer Science and Technology, Hong Kong Baptist University United International College, 2012

M.S. in Computer Science, Chinese University of Hong Kong, 2013

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Haoran Zhang

It was defended on

December 2, 2020

and approved by

Dr. Diane Litman, Department of Computer Science & LRDC, University of Pittsburgh

Dr. Adriana Kovashka, Department of Computer Science, University of Pittsburgh

Dr. Erin Walker, Department of Computer Science, University of Pittsburgh

Dr. Richard Correnti, School of Education, University of Pittsburgh

Exploring Automated Essay Scoring Models for Multiple Corpora and Topical Component Extraction from Student Essays

Haoran Zhang, PhD

University of Pittsburgh, 2021

Since it is a widely accepted notion that human essay grading is labor-intensive, automatic scoring method has drawn more attention. It reduces reliance on human effort and subjectivity over time and has commercial benefits for standardized aptitude tests. Automated essay scoring could be defined as a method for grading student essays, which is based on high inter-agreement with human grader, if they exist, and requires no human effort during the process. This research mainly focuses on improving existing Automated Essay Scoring (AES) models with different technologies. We present three different scoring models for grading two corpora: the Response to Text Assessment (RTA) and the Automated Student Assessment Prize (ASAP). First of all, a traditional machine learning model that extracts features based on semantic similarity measurement is employed for grading the RTA task. Secondly, a neural network model with the co-attention mechanism is used for grading sourced-based writing tasks. Thirdly, we propose a hybrid model integrating the neural network model with hand-crafted features. Experiments show that the feature-based model outperforms its baseline, but a stand-alone neural network model significantly outperforms the feature-based model. Additionally, a hybrid model integrating the neural network model and hand-crafted features outperforms its baselines, especially in a cross-prompt experimental setting. Besides, we present two investigations of using the intermediate output of the neural network model for keywords and key phrases extraction from student essays and the source article. Experiments show that keywords and key phrases extracted by our models support the feature-based AES model, and human effort can be relieved by using automated essay quality signals during the training process.

Table of Contents

List of Tables

List of Figures

Preface

It is a fantastic journey from a zero research experience student to getting a Ph.D. degree. I want to thank everyone who helped me during my journey and made it joyful and achievable.

First of all, I would like to express my greatest gratitude to my advisor, Dr. Diane Litman, who has guided me from the beginning till the end of my journey. It has been a great experience to be at the PETAL (ITSPOKE) lab and receiving endless support and encouragement during my study. In a gradual process, my advisor introduced me to a challenging but interesting research topic and mentored me step by step. It makes me feel not stressed but joyful to pursue my degree. Besides, my advisor also supported me as I am an international student. She tried her best to help and give me extra time to deal with the administration problem, such as applying for a visa and changing my study plan based on the pandemic circumstances. Without her help, I cannot finish my Ph.D. study.

I would also like to thank my dissertation committee: Dr. Richard Correnti, Dr. Adriana Kovashka, and Dr. Erin Walker. I have benefited deeply from your constructive comments and tremendous support. I would like to thank Dr. Jingtao Wang, Dr. Janyce Wiebe, Dr. Shi-Kuo Chang, and Dr. Rebecca Hwa for advising me in the early stage of my research. I also would like to thank Dr. Taieb Znati, Dr. Weifeng Su, Dr. Hui Zhang, Dr. Jing Zhao, Dr. Xin Feng, Dr. Seonphil Sunny Jeong, Chunyan Ji, Yanyan Ji, Dr.Yafei Li, Dr. Zhiyuan Li, Jing He, Yu Liu for helping me even before I started my study.

I am grateful to all former and current members of the PETAL lab and the RTA group who helped me with my research and gave me insights.

I would like to thank all my family members, including my wife (Xue), and my parents (Baohong, Xiulin, Gaimin, and Yanzhen), who love, encourage, and support me. I would also like to thank my brothers Jiaxu Zhang, Dacheng Lv, Jia Ju, Peng Sun, and Hengrui Gu, who are always on my side when I am facing any difficulty.

Finally, I want to say thank you to Yaxi Deng and Kaibin Lei who helped me on an early version of this thesis. I am also grateful to all my good friends: Longhao Li, Mingda

Zhang, Keren Ye, Zinan Zhang, Xiaozhong Zhang, Xiaoyu Liang, Mingzhi Yu, Changsheng Liu, Wei Guo, Fangzhou Cheng, Duncan Yung, Yanbing Xue, Zhipeng Luo, Zhenjiang Fan, Phuong Pham, Jeongmin Lee, Tazin Afrin, Luca Lugini, Ahmed Magooda, Yuhao Song, Jiaqi Yang, Zhanyu Xu, Shuli Lin, Yuanbo Zhou, Lixing Lian, Shukun Xie, Qiao Gao, Shuai Wang, Lixin Liu, Shuai Shi, Yihan Li, Jingchao Wei, Hui Dong, Yanda Li, and Zherui Yang. I could never have completed this degree without them.

Thank you all!!!

1.0 Introduction

Manually grading students' essays is labor-intensive because it requires expert knowledge of the raters. Usually, it takes time for the rater to be trained and for the essays to be graded on a large scale. The subjectivity and time-consuming nature of manual grading may give rise to biases. Therefore, the Automated Essay Scoring (AES) is in demand to provide reliable essay scores without or with the least human effort. Besides, there are more benefits proposed by the AES, such as improved consistency and efficiency as well as minimal cost [Gierl et al., 2014].

AES is one of the most important education applications of natural language processing (NLP). Although research in this area has been ongoing for more than 50 years [Page, 1968], it still draws a lot of attention from the NLP community.

The first step of the AES is getting essay representation. In general, there are two ways to extract essay representation, either by feature engineering or by neural network for automatic feature extraction. Because of the limited availability of annotated corpora and the long history of research in this area, most AES requires feature engineering. By designing hand-crafted features carefully, an AES model can be trained on a small annotated corpus, while maintaining a good performance. Commonly used features include lexical features [Attali and Burstein, 2006], syntactic features [Chen and He, 2013], use of figurative language [Louis and Nenkova, 2013], discourse features [Song et al., 2017], semantic features [Cozma et al., 2018], argument strength features [Persing and Ng, 2015], and rubric-based features [Yamamoto et al., 2019].

Recently, more and more neural network models are introduced into this area. One major benefit is that they no longer need feature engineering. Model essay with RNN layer is standard because it captures long-distance dependencies of the words in the essay [Taghipour and Ng, 2016, Alikaniotis et al., 2016]. Besides, with the development of the neural network model in other research areas, more advanced structures are employed in the neural AES model, such as the hierarchical model [Dong and Zhang, 2016], the BERT embedding model [Liu et al., 2019], the attention model [Dong et al., 2017, Li et al., 2018], the SkipFlow

mechanism [Tay et al., 2018], and multi-task learning [Farag et al., 2018].

Most existing AES systems are supervised learning based systems. Three major learning algorithms are widely used in this area. The first is regression, which is used by most existing systems [Persing and Ng, 2015, Phandi et al., 2015, Taghipour and Ng, 2016, Dong and Zhang, 2016, Dong et al., 2017, Cozma et al., 2018]. The goal of such systems is predicting essay scores directly. Second is classification, which is used by some works to label an essay with a small number of classes [McNamara et al., 2015, Vajjala, 2018, Farra et al., 2015, Nguyen and Litman, 2018, Rahimi et al., 2014, Rahimi et al., 2017, Rudner and Liang, 2002]. The classes could be low, medium, or high class, or a small range of scores. Third, ranking, which is also employed to rank essays based on their quality [Yannakoudakis et al., 2011, Chen and He, 2013, Cummins et al., 2016].

Primarily, the AES problem could be divided into two directions [Ke and Ng, 2019]: holistic scoring and dimension scoring. The holistic score represents the overall quality of the essay, while dimension score measure a specific aspect of the essay. The vast majority of previous works have focused on holistic scoring [Yannakoudakis and Briscoe, 2012, Cozma et al., 2018, Vajjala, 2018, Taghipour and Ng, 2016, Alikaniotis et al., 2016, Dong and Zhang, 2016, Dong et al., 2017, Tay et al., 2018, Phandi et al., 2015, Jin et al., 2018, Nadeem et al., 2019], mainly because of two reasons. For one thing, most AES systems use supervised learning algorithms to predict essay scores, which requires human-annotated corpora to serve as training data. Unfortunately, most publicly available corpora are annotated with only holistic score, such as the Cambridge Learner Corpus-First Certificate in English exam corpus (CLC-FCE) [Yannakoudakis et al., 2011], the Automated Student Assessment Prize corpus (ASAP), and the TOEFL11 corpus [Blanchard et al., 2013]. Among them, the ASAP corpus has the largest number of essays, including 17450 essays over 8 different prompts. This corpus is discussed more thoroughly in Chapter 2. For another reason, holistic scoring has commercial benefits in automatically scoring essays from standardized aptitude tests such as SAT, GRE, TOFEL, and IELTS. These tests require enormous human effort to score a large number of essays within a compressed timeline, which could be reduced by the AES system.

However, a holistic score is not enough in the classroom setting. For example, holistic

scores only tell students about the overall quality of essays without providing further feedback, whereas feedback is essential for essay revision. One possible way to provide feedback is providing dimension scores. Possible dimensions of measuring essay quality are shown in Table 1 [Ke and Ng, 2019]. With dimension scores, it is easy to know which aspect(s) of the essay has room for improvement. There are less publicly available corpora for dimension scoring, examples would be the International Corpus of Learner English (ICLE) [Granger et al., 2009] and the Argument Annotated Essays (AAE) [Stab and Gurevych, 2014], which leads to a limited number of researches in this direction [Persing et al., 2010, Persing and Ng, 2013, Persing and Ng, 2015, Nguyen and Litman, 2018, Louis and Higgins, 2010, Burstein et al., 2010, Somasundaran et al., 2014]. In this research, we use a corpus named the Response to Text Assessment (RTA) [Correnti et al., 2013] for assessing writing skills in Analysis, Evidence, Organization, Style, and MUGS (Mechanics, Usage, Grammar, and Spelling) dimensions. We focus on the evidence dimension, which evaluates students' ability to find and use evidence from a source article to support their position. More details about this corpus will be given in Chapter 2.

Dimension	**Description**
Grammaticality	Grammar
Usage	Use of prepositions, word usage
Mechanics	Spelling, punctuation, capitalization
Style	Word choice, sentence structure variety
Relevance	Relevance of the content to the prompt
Organization	How well the essay is structured
Development	Development of ideas with examples
Cohesion	Appropriate use of transition phrases
Coherence	Appropriate transitions between ideas
Thesis Clarity	Clarity of the thesis
Persuasiveness	Convincingness of the major argument

Table 1: Different dimensions of essay quality [Ke and Ng, 2019].

In this research, essay corpora are divided into two categories, depending on the form of the writing task. In the simplest task, students need to write an essay to respond to a

prompt, and no other information is offered. In another form (the source-based writing task), a source article is provided. It requires students to read a source article before writing an essay to respond to the prompt. Usually, the prompt is highly related to the source article.

In order to evaluate the performance of AES models, the in-prompt experiment setting is widely used [Yannakoudakis and Briscoe, 2012, Persing and Ng, 2014, Tay et al., 2018, Vajjala, 2018, Farag et al., 2018]. A model is trained and tested on essays from the same prompt. This straightforward method is suitable for evaluating the prompt-specific model. However, a cross-prompt experiment setting could further evaluate the ability of prompt adaptation of AES models [Phandi et al., 2015, Cozma et al., 2018, Liu et al., 2019, Cao et al., 2020]. In this experimental setting, a model is trained on essays from the source prompt combine with a limited number of essays from the target prompt, and tested on essays from the target prompt. In this research, we mainly focus on the in-prompt experiment, while exploring the cross-prompt experiment in Chapter 3 and Chapter 5.

1.1 Research Statements

In this research, we mainly focus on two directions. In Chapter 3 to Chapter 5, we developed three AES models with different technologies. In Chapter 6 and Chapter 7, we explored a way of using a neural network AES model to extract keywords and key phrases from the source article of source-based writing tasks.

First, in Chapter 3, we developed [Zhang and Litman, 2017] a feature-based model that employs word embedding for feature extraction, and evaluates a specific dimension of student essay from a source-based writing task, called evidence dimension. A previous research [Rahimi et al., 2014, Rahimi et al., 2017] extracted features based only on the lexical form of words. However, it assessed additional information beyond the rubric, such as grammar mistakes. Besides, it could not recognize words that are out of the vocabulary. Therefore, this model extracts features derived from the lexical form of words as well as the semantic meaning of words. Since the extracted features are closely connected to the content of the source article, this model only works for source-based writing tasks. Unfortunately,

this model could only work to score the evidence dimension, limiting the scope of usage of this model.

Second, in Chapter 4, we developed a model that introduces the co-attention mechanism into the neural network model [Zhang and Litman, 2018]. Since the neural network model demonstrates a strong ability for modeling word sequence, whose requirements cannot be fulfilled by feature engineering, we switch our model from the feature-based model to the neural network model. Experiments show that the co-attention neural network model outperforms the feature-based models significantly. Besides, experiments show that this model is not only effective for evidence dimension scoring but also for holistic scoring. Unfortunately, the potential of this model is still limited in the source-based writing tasks due to the design of the neural network.

Third, in Chapter 5, we proposed a hybrid model that integrates hand-crafted features into the neural network model. The neural network model uses hand-crafted features as external knowledge and guides the training process. Unlike the co-attention neural network model, the source article is unnecessary for this model due to the design of this model. Depending on the hand-crafted features we integrate, it is suitable for both holistic scoring and dimension scoring. Since this model integrates hand-crafted features that may adapt across prompts, this model also works in the cross-prompt situation.

Fourth, although the neural networks model outperforms the feature-based model in terms of score prediction, hand-crafted features are still needed for an Automated Writing Evaluation (AWE) system [Zhang et al., 2019] as hand-crafted features provide more interpretable information than features extracted by neural network models. However, our feature-based model still requires human effort to extract topic words and specific example phrases from the source article. Therefore, in Chapter 6, we presented a model that generates topic words and specific example phrases automatically with intermediate outputs of the co-attention neural network model [Zhang and Litman, 2020].

Lastly, in Chapter 7, we extended the previous work even further. The previous work trains the co-attention neural network model 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 investigated 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.

1.2 Contributions

For the educational community, we developed various kinds of models (feature-based model, neural network model, and hybrid model) that assess student essays for either holistic score, or evidence score. These models can be used for educational purposes and relieves human burden.

As for the AES community, we presented multiple models that assess student essays more accurately. On top of that, we introduced the co-attention mechanism into this research area, as well as a hybrid model with more hand-crafted features and more advanced ways of integration from lower levels (word level or sentence level).

For the NLP community, we proposed multiple contributions. First, we proposed a method to use the word embedding model. Rather than using the word embedding as word representation, our model uses it for feature extraction in order to match words in their semantic meaning rather than in lexical form. Second, we showed that the co-attention mechanism that was originally developed for machine comprehension can also be implemented on automated source-based essay scoring. Third, we presented a hybrid model that combines the neural network model with hand-crafted features. However, our model integrates hand-crafted features from lower levels and models hand-crafted features as sequences of inputs. Fourth, we show that other than the final output of the neural network model, its intermediate output also provides useful information for downstream applications, such as keyword and keyphrase extraction. At last, we showed that although the weakly supervised AES approach is insufficient for essay scoring, it is still useful for generating keywords or key phrases.

2.0 Datasets

2.1 RTA Dataset

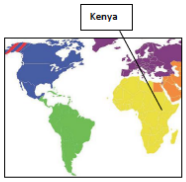
The Response to Text Assessment (RTA) [Correnti et al., 2013] assesses student's analytic response-to-text writing skills. The RTA was designed to evaluate writing skills in Analysis, Evidence, Organization, Style, and MUGS (Mechanics, Usage, Grammar, and Spelling) dimensions. In this research, we only focus on the evidence dimension.

The RTA essay corpora were all collected from classrooms using the following procedure. The teacher first read aloud an article while students followed along with their copy. After the teacher explained some predefined vocabulary and discussed standardized questions at designated points, there is a prompt at the end of the text which asks students to write an essay in response to the prompt.

Two forms of the RTA have been developed, based on different articles that students read before writing essays in response to a prompt. The first form is RTA_{MVP} and is based on an article from *Time for Kids* about the Millennium Villages Project, an effort by the United Nations to end poverty in a rural village in Sauri, Kenya. The other form is RTA_{Space}, based on a developed article about the importance of space exploration. Figure 1 and Figure 2 show source articles and prompts of RTA_{MVP} and RTA_{Space}, respectively.

Two corpora of RTA_{MVP} from lower and higher age groups were introduced in [Correnti et al., 2013]. One group included grades 4-6 (denoted by MVP_L), and the other group included grades 6-8 (denoted by MVP_H). The students in each age group represent different levels of writing proficiency. We also combined these two corpora to form a larger corpus, denoted by MVP. The corpus of the RTA_{Space} is collected only from students of grades 6-8 (denoted by *Space*).

Based on the rubric criterion shown in Table 2, the essays in each corpus were annotated by two raters on a scale of 1 to 4 (from low to high). Raters are experts and trained undergraduates. Table 3 shows the distribution of Evidence scores. For MVP_L, MVP_H, MVP, and *Space*, scores are from the first rater because the first rater graded more essays.



Today we are going to read an article about a project of the United Nations called the Millennium Villages project in Kenya, a country in Africa. The United Nations is an organization that helps people around the world to have a safer, healthier and better life.
The person who wrote this article was 12 years old when she visited Kenya with her father in 2010 when the project was just beginning. At that time the people of Sauri lived on less than $1 dollar a day. She also tells us in the article about her return to Kenya and how the Millennium Villages project has helped many people have a better life already.

A Brighter Future
Hannah Sachs

The unpaved dirt road made our car jump as we traveled to the Millennium Village in Sauri (sah-ooh-ree), Kenya. We passed the market where women sat on the dusty ground selling bananas. Little kids were wrapped in cloth on their mothers' backs, or running around in bare feet and tattered clothing. When we reached the village, we walked to the Bar Sauri Primary School to meet the people. Welcoming music and singing had almost everyone dancing. We joined the dancing and clapped along to the joyful, lively music.

Tattered: Torn or ragged

The year was 2010, the first time I had ever been to Sauri. With the help of the Millennium Villages project, the place would change dramatically in the coming years. The Millennium Villages project was created to help reach the Millennium Development Goals.

Poverty: Poor; having little or no money

The plan is to get people out of poverty, assure them access to health care and help them stabilize the economy and quality of life in their communities. Villages get technical advice and practical items, such as fertilizer, medicine and school supplies. Local leaders take it from there. The goals are supposed to be met by 2025; some other targets are set for 2035. We are halfway to 2025, and the world is capable of meeting these goals. But our first glimpse of Sauri showed us that there was plenty of work to do.

Fertilizer: A substance spread into soil that helps and supports plant growth

The Fight for Better Health
On that day in 2010, we followed the village leaders into Yala Sub-District Hospital. It was not in good shape. There were three kids to a bed and two adults to a bed. The rooms were packed with patients who probably would not receive treatment, either because the hospital did not have it or the patients could not afford it. There was no doctor, only a clinical officer running the hospital. There was no running water or electricity. It is hard for me to see people sick with preventable diseases, people who are near death when they shouldn't have to be. I just get scared and sad.

Malaria (mah-lair-eeh-ah) is one disease, common in Africa, that is preventable and treatable. Mosquitoes carry malaria, and infect people by biting them. Kids can die from it easily, and adults get very sick.

Mosquitoes that carry malaria come at night. A bed net, treated with chemicals that last for five years, keeps malarial mosquitoes away from sleeping people. Each net costs $5. There are some cheap medicines to get rid of malaria too. The solutions are simple, yet 20,000 kids die from the disease each day. So sad, and so illogical. Bed nets could save millions of lives.

Water, Fertilizer, Knowledge
We walked over to see the farmers. Their crops were dying because they could not afford the necessary fertilizer and irrigation. Time and again, a family will plant seeds only to have an outcome of poor crops because of lack of fertilizer and water. Each year, the farmers worry: Will they harvest enough food to feed the whole family? Will their kids go hungry and become sick?

Many kids in Sauri did not attend school because their parents could not afford school fees. Some kids are needed to help with chores, such as fetching water and wood. In 2010, the schools had minimal supplies like books, paper and pencils, but the students wanted to learn. All of them worked hard with the few supplies they had. It was hard for them to concentrate, though, as there was no midday meal. By the end of the day, kids didn't have any energy.

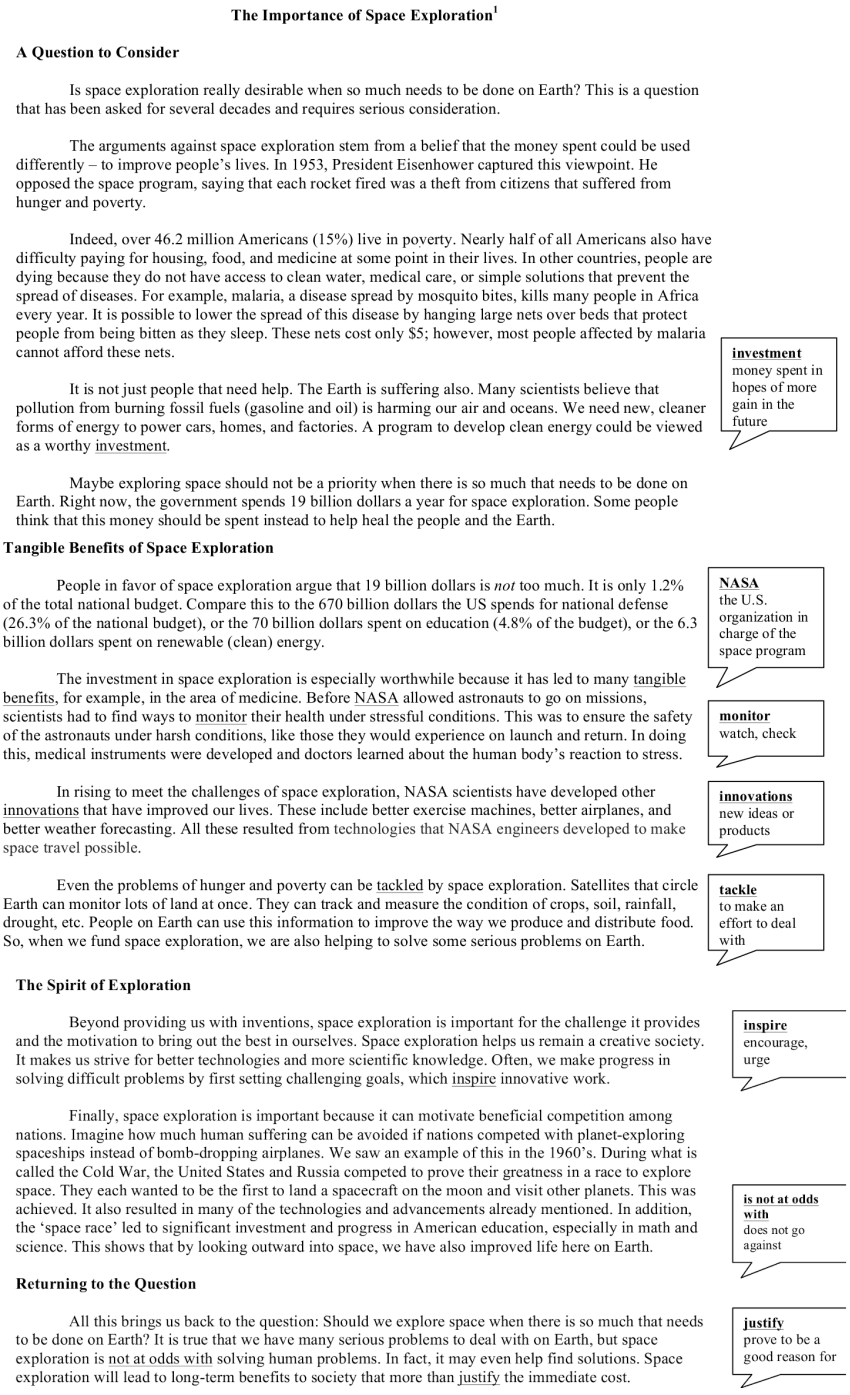
A Better Life—2018
The people of Sauri have made amazing progress in just eight years. The Yala Sub-District Hospital has medicine, free of charge, for all of the most common diseases. Water is connected to the hospital, which also has a generator for electricity. Bed nets are used in every sleeping site in Sauri. The hunger crisis has been addressed with fertilizer and seeds, as well as the tools needed to maintain the food supply. There are no school fees, and the school now serves lunch for the students. The attendance rate is way up.

Impoverished Poverty-stricken; emptied of strength or richness

Dramatic changes have occurred in 80 villages across sub-Saharan Africa. The progress is encouraging to supporters of the Millennium Villages project. There are many solutions to the problems that keep people impoverished. What it will really take is for the world to work together to change poverty-stricken areas for good. When my kids are my age, I want this kind of poverty to be a thing of history. It will not be an easy task. But Sauri's progress shows us all that winning the fight against poverty is achievable in our lifetime.

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.

Figure 1: Source text and prompt of RTA_{MVP}.



The Importance of Space Exploration[1]

A Question to Consider

Is space exploration really desirable when so much needs to be done on Earth? This is a question that has been asked for several decades and requires serious consideration.

The arguments against space exploration stem from a belief that the money spent could be used differently – to improve people's lives. In 1953, President Eisenhower captured this viewpoint. He opposed the space program, saying that each rocket fired was a theft from citizens that suffered from hunger and poverty.

Indeed, over 46.2 million Americans (15%) live in poverty. Nearly half of all Americans also have difficulty paying for housing, food, and medicine at some point in their lives. In other countries, people are dying because they do not have access to clean water, medical care, or simple solutions that prevent the spread of diseases. For example, malaria, a disease spread by mosquito bites, kills many people in Africa every year. It is possible to lower the spread of this disease by hanging large nets over beds that protect people from being bitten as they sleep. These nets cost only $5; however, most people affected by malaria cannot afford these nets.

It is not just people that need help. The Earth is suffering also. Many scientists believe that pollution from burning fossil fuels (gasoline and oil) is harming our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. A program to develop clean energy could be viewed as a worthy investment.

investment
money spent in hopes of more gain in the future

Maybe exploring space should not be a priority when there is so much that needs to be done on Earth. Right now, the government spends 19 billion dollars a year for space exploration. Some people think that this money should be spent instead to help heal the people and the Earth.

Tangible Benefits of Space Exploration

People in favor of space exploration argue that 19 billion dollars is *not* too much. It is only 1.2% of the total national budget. Compare this to the 670 billion dollars the US spends for national defense (26.3% of the national budget), or the 70 billion dollars spent on education (4.8% of the budget), or the 6.3 billion dollars spent on renewable (clean) energy.

NASA
the U.S. organization in charge of the space program

The investment in space exploration is especially worthwhile because it has led to many tangible benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health under stressful conditions. This was to ensure the safety of the astronauts under harsh conditions, like those they would experience on launch and return. In doing this, medical instruments were developed and doctors learned about the human body's reaction to stress.

monitor
watch, check

In rising to meet the challenges of space exploration, NASA scientists have developed other innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. All these resulted from technologies that NASA engineers developed to make space travel possible.

innovations
new ideas or products

Even the problems of hunger and poverty can be tackled by space exploration. Satellites that circle Earth can monitor lots of land at once. They can track and measure the condition of crops, soil, rainfall, drought, etc. People on Earth can use this information to improve the way we produce and distribute food. So, when we fund space exploration, we are also helping to solve some serious problems on Earth.

tackle
to make an effort to deal with

The Spirit of Exploration

Beyond providing us with inventions, space exploration is important for the challenge it provides and the motivation to bring out the best in ourselves. Space exploration helps us remain a creative society. It makes us strive for better technologies and more scientific knowledge. Often, we make progress in solving difficult problems by first setting challenging goals, which inspire innovative work.

inspire
encourage, urge

Finally, space exploration is important because it can motivate beneficial competition among nations. Imagine how much human suffering can be avoided if nations competed with planet-exploring spaceships instead of bomb-dropping airplanes. We saw an example of this in the 1960's. During what is called the Cold War, the United States and Russia competed to prove their greatness in a race to explore space. They each wanted to be the first to land a spacecraft on the moon and visit other planets. This was achieved. It also resulted in many of the technologies and advancements already mentioned. In addition, the 'space race' led to significant investment and progress in American education, especially in math and science. This shows that by looking outward into space, we have also improved life here on Earth.

is not at odds with
does not go against

Returning to the Question

All this brings us back to the question: Should we explore space when there is so much that needs to be done on Earth? It is true that we have many serious problems to deal with on Earth, but space exploration is not at odds with solving human problems. In fact, it may even help find solutions. Space exploration will lead to long-term benefits to society that more than justify the immediate cost.

justify
prove to be a good reason for

Prompt: Consider the reasons given in the article for why we should and should not fund space exploration. Did the author convince you that "space exploration is desirable when there is so much that needs to be done on earth"? Give reasons for your answer. Support your reasons with 3-4 pieces of evidence from the text.

Figure 2: Source text and prompt of RTA_{Space}.

	1	2	3	4
Number of Pieces of evidence	Features one or no pieces of evidence (NPE)	Features at least 2 pieces of evidence (NPE)	Features at least 3 pieces of evidence (NPE)	Features at least 3 pieces of evidence (NPE)
Relevance of evidence	Selects inappropriate or irrelevant details from the text to support key idea (SPC); references to text feature serious factual errors or omissions	Selects some appropriate and relevant evidence to support key idea, or evidence is provided for some ideas, but not actually the key idea (SPC); evidence may contain a factual error or omission	Selects pieces of evidence from the text that are appropriate and relevant to key idea (SPC)	Selects evidence from the text that clearly and effectively supports key idea
Specificity of evidence	Provides general or cursory evidence from the text (SPC)	Provides general or cursory evidence from the text (SPC)	Provides specific evidence from the text (SPC)	Provides pieces of evidence that are detailed and specific (SPC)
Elaboration of Evidence	Evidence may be listed in a sentence (CON)	Evidence provided may be listed in a sentence, not expanded upon (CON)	Attempts to elaborate upon evidence (CON)	Evidence must be used to support key idea / inference(s)
Plagiarism	Summarize entire text or copies heavily from text (in these cases, the response automatically receives a 1)			

Table 2: Rubric for the evidence dimension of RTA. The abbreviations in the parentheses identify the corresponding feature group discussed in Chapter 3 that is aligned with that specific criteria [Rahimi et al., 2014, Rahimi et al., 2017].

Figure 3 shows an excerpt from the article of RTA_{MVP} and an essay with a score of 3; evidence from the text that raters want to see in students' essays are in bold.

Since these corpora have not been released, there is only one work that focuses on assessing the evidence dimension [Rahimi and Litman, 2016]. The common evaluation method is Quadratic Weighted Kappa (QWK) for RTA corpus. The state-of-the-art performance beside this research is shown in Table 4.

Given the fact that the feature-based AES model (Chapter 3) is a traditional machine learning model which does not require large datasets for training, Chapter 3 uses MVP_L, MVP_H, MVP, and *Space* for its experiments. However, in the rest of chapters, only MVP and *Space* will be used, since they are using neural network and require larger datasets for training. All experimental performances are measured by Quadratic Weighted Kappa.

	MVP_L	MVP_H	MVP	*Space*
Score 1	535	317	852	538
	(30%)	(27%)	(29%)	(26%)
Score 2	709	488	1197	789
	(39%)	(42%)	(40%)	(38%)
Score 3	374	242	616	512
	(21%)	(21%)	(21%)	(25%)
Score 4	186	119	305	237
	(10%)	(10%)	(10%)	(11%)
Total	1804	1166	2970	2076

Table 3: The distribution of evidence scores.

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: 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.

Figure 3: An excerpt from the article of RTA_{MVP} and an example essay with score of 3.

Method	MVP_L	MVP_H	MVP	*Space*
[**Rahimi and Litman, 2016**]	0.628	0.599	0.624	0.606

Table 4: The state-of-the-art performances of recent models for RTA corpus. The best QWK score for each prompt is highlighted in bold.

2.2 ASAP Dataset

The Automated Student Assessment Prize (ASAP)[1] was released in 2012, and consists of written responses to 8 prompts (denoted by $ASAP_1$ to $ASAP_8$). All responses were written by students ranging in grade levels from Grade 7 to Grade 10. Each prompt has its own unique characteristics, which intends to test the limits of capabilities of an AES model. Since the scores assigned to essays are holistic, assessment evaluates an essay's overall quality rather than a specific dimension. Table 5 shows all eight prompts of ASAP. Among them, $ASAP_3$, $ASAP_4$, $ASAP_5$, and $ASAP_6$ are source-based which means students read an article before writing their essays. Appendix A show source articles of $ASAP_3$, $ASAP_4$, $ASAP_5$, and $ASAP_6$, respectively.

Based on the grading rubrics in Appendix B, all essays were hand graded and were double-scored. Finally, a holistic score is always assigned to an essay. Table 6 shows the score range, number of essays, and average length of each ASAP prompt. Figure 4 shows an essay with score of 4 for $ASAP_5$.

The release of ASAP corpus has renewed interest on the AES topic. Most work in this area uses ASAP corpus for evaluation [Chen and He, 2013, Phandi et al., 2015, Taghipour and Ng, 2016, Dong and Zhang, 2016, Alikaniotis et al., 2016, Dong et al., 2017, Cozma et al., 2018, Tay et al., 2018, Liu et al., 2019]. The common evaluation method is Quadratic Weighted Kappa (QWK) for ASAP corpus. The state-of-the-art performances of recent models are shown in Table 7.

In this research, Chapter 4 only focus on $ASAP_3$, $ASAP_4$, $ASAP_5$, and $ASAP_6$, because

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

Task	Prompt
1	More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.
2	Censorship in the Libraries "All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf – that work I abhor – then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us." –Katherine Paterson, Author Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive? Support your position with convincing arguments from your own experience, observations, and/or reading.
3	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.
4	Read the last paragraph of the story. "When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again." Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas.
5	Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir.
6	Based on the excerpt, describe the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. Support your answer with relevant and specific information from the excerpt.
7	Write about patience. Being patient means that you are understanding and tolerant. A patient person experience difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.
8	We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.

Table 5: Prompts of ASAP.

they are source-based responses. They have similar setting to the RTA corpus, except that the scores were assigned to essays based on not only use of evidence, but also other aspects, although students were asked to use evidence from the source article to support their claims. In contrast, Chapter 5 focuses on all prompts because the model does not require a source article.

Essay: The author of the memoir, Narciso Rodriguez creates a caring, happy, and thoughtful mood. By mentioning the Cuban traditions shared in the neighborhood between close friends, and cooking in the kitchen to share a great meal with one another the mood is happy. When Narciso talks about the great friends he made from different heritages and knowing the entire community like family the mood is thoughtful and caring because it shows that the people really appreciated each other's company. It is also caring in the story when Narciso talks about how his parents devoted their lives to making sure that their children and the people they knew had good lives to. When Narciso describes the way his parents struggled during the cold winters, yet they always let others in, shows a very caring mood in the memoir. I also think that the fact that a small, simple apartment they lived in is very important to Narciso because he repeats it several times. I think he does this to show a thoughtful for mood, in that the house was small but through creativity in bringing culture in made it seem much bigger.

Figure 4: An essay with score of 4 for $ASAP_5$.

Prompt	**Lowest**	**Highest**	**# Essays**
$ASAP_1$	2	12	1783
$ASAP_2$	1	6	1800
$ASAP_3$	0	3	1726
$ASAP_4$	0	3	1772
$ASAP_5$	0	4	1805
$ASAP_6$	0	4	1800
$ASAP_7$	0	30	1569
$ASAP_8$	0	60	723

Table 6: The score range and number of essays of each ASAP prompt.

Method	$ASAP_1$	$ASAP_2$	$ASAP_3$	$ASAP_4$	$ASAP_5$	$ASAP_6$	$ASAP_7$	$ASAP_8$	Overall
[**Phandi et al., 2015**]	0.761	0.606	0.621	0.742	0.784	0.775	0.730	0.617	0.705
[**Dong and Zhang, 2016**]	NA	NA	NA	NA	NA	NA	NA	NA	0.734
[**Dong et al., 2017**]	0.822	0.682	0.672	0.814	0.803	0.811	0.801	0.705	0.764
[**Tay et al., 2018**]	0.832	0.684	0.695	0.788	0.815	0.810	0.800	0.697	0.764
[**Cozma et al., 2018**]	0.845	0.729	0.684	0.829	0.833	**0.830**	0.804	**0.729**	0.785
[**Liu et al., 2019**]	**0.852**	**0.736**	0.731	0.801	0.823	0.792	0.762	0.684	0.773
[**Cao et al., 2020**]	0.824	0.699	0.726	0.859	0.822	0.828	0.840	0.726	0.791
[**Uto et al., 2020**]	**0.852**	0.651	**0.804**	**0.888**	**0.885**	0.817	**0.864**	0.645	**0.801**

Table 7: The state-of-the-art performances of recent models for ASAP corpus. The best QWK score for each prompt is highlighted in bold.

3.0 Word Embedding for Response-To-Text Assessment of Evidence

3.1 Introduction

Manually grading the RTA is labor-intensive. Therefore, an automated scoring method was developed [Rahimi et al., 2014, Rahimi et al., 2017], which defined a set of interpretable features based on the grading rubric shown in Table 2. Although these features significantly improve over competitive baselines, the feature extraction approach is primarily based on lexical matching and can be enhanced.

In this chapter, we introduced word embedding to improve the existing AES model [Rahimi et al., 2014, Rahimi et al., 2017]. The major contributions of this chapter are employing a new way of using the word embedding model and showing the word embedding could be used to deal with noisy data given the disparate writing skills of students at the upper elementary level. This work is illustrated in [Zhang and Litman, 2017].

3.2 Related Work

Most research studies in automated essay scoring have focused on holistic rubrics [Attali and Burstein, 2006, Shermis and Burstein, 2003]. In contrast, our work focuses on evaluating a single dimension to obtain a rubric score for students' use of evidence from a source text to support their stated position. To evaluate the content of students' essays, Louis and Higgins [Louis and Higgins, 2010] presented to detect if an essay is off-topic. Xie et al. [Xie et al., 2012] presented a method to evaluate content features by measuring the similarity between essays. Burstein et al. [Burstein et al., 2001], and Ong et al. [Ong et al., 2014] both presented methods to use argumentation mining techniques to evaluate the students' use of evidence to support claims in persuasive essays. However, those studies are different from this work in that they did not measure how the essay uses material from the source article. Furthermore, young students find it difficult to use sophisticated argumentation structures

in their essays.

Rahimi et al. [Rahimi et al., 2014, Rahimi et al., 2017] presented a set of interpretable rubric features that measure the relatedness between students' essays and a source article by extracting evidence from the students' essays based on lexical matching. However, evidence from students' essays could not always be extracted by their word matching method. For example, different vocabularies other than words from the article or spelling error (which is not assessed by the rubric) affect the lexical matching method. There are some potential solutions using the word embedding model. Rei and Cummins [Rei and Cummins, 2016] presented a method to evaluate topical relevance by estimating sentence similarity using weighted-embedding. Kenter and de Rijke [Kenter and de Rijke, 2015] evaluated short text similarity with word embedding. Kiela et al. [Kiela et al., 2015] developed specialized word embedding by employing external resources. However, none of these methods address essays written by young students.

Most recently, one of the state-of-the-art AES models presented by Cozma et al. [Cozma et al., 2018] also introduced the word embedding model into this area, which combined the bag-of-super-word-embeddings (BOSWE) [Butnaru and Ionescu, 2017] with string kernels. They used the BOSWE to obtain document embedding by computing the occurrence count of each super word embedding in the respective document. In contrast, we use word embedding for word matching and extract interpretable features. Furthermore, the BOSWE does not contribute to the model stand-alone, and improvement only can be observed when the BOSWE is combined with the string kernel. However, our method only uses word embedding to improve model performance.

Besides, neural network models also play an essential role in this area. There are multiple neural network models were developed for assessing students' essays more accurate [Taghipour and Ng, 2016, Dong and Zhang, 2016, Dong et al., 2017, Nadeem et al., 2019, Liu et al., 2019]. Unfortunately, none of them provides additional feedback besides the final score because the features extracted by neural network models are not interpretable. On the other side, our model improves model performance by extracting interpretable features more accurately. Therefore, the extracted feature is also useful for other downstream applications, such as the Automated Writing Evaluation system.

3.3 Rubric Features

Based on the rubric criterion for the evidence dimension, Rahimi et al. [Rahimi et al., 2014, Rahimi et al., 2017] developed a set of interpretable features related to the use of *Topical Components* (TCs) in an essay. By using this set of features, a predicting model can be trained for automated essay scoring in the evidence dimension. Before extracting features, the expert effort was first required to create the TCs. For each source, the TCs consist of a comprehensive list of topics related to evidence which include: 1) important words indicating the set of evidence topics in the source, and 2) phrases representing specific examples for each topic that students need to find and use in their essays. Table 35 and Table 39 are topic words list and specific example phrases list of MVP article, respectively.

Number of Pieces of Evidence (NPE). A good essay should mention evidence from the article as much as possible. Then, they use a simple window-based algorithm with a fixed size window to extract this feature. If a window contains at least two words from the *topic words* list, they consider this window to contain evidence related to a topic. To avoid redundancy, each topic is only counted once. Words from the window and crafted list will only be considered a match if they are exactly the same. This feature is an integer to represent the number of topics that are mentioned by the essay.

Concentration (CON). Rather than list all the topics in the essay, a good essay should explain each topic with details. The same *topic words* list and simple window-based algorithm are used for extracting the CON feature. An essay is concentrated if the essay has fewer than 3 sentences that mention at least one of the topic words. Therefore, this feature is a binary feature. The value is 1 if the essay is concentrated, otherwise it is 0.

Specificity (SPC). A good essay should use relevant examples as much as possible. For each example from *specific example phrases* list, the same window-based algorithm is used for matching. If the window contains at least two words from an example, they consider the window to mention this example. Therefore, the SPC feature is an integer vector. Each value in the vector represents how many examples in this topic were mentioned by the essay. To avoid redundancy, each example is only to be counted at most one time. The length of the vector is the same as the number of categories of examples in the crafted list.

Word Count (WOC). The SPC feature can capture how many evidences were mentioned in the essay, but it cannot represent if these pieces of evidence support key ideas effectively. From previous work, we know longer essays tend to have higher scores. Thus, they use word count as a potentially helpful fallback feature. This feature is an integer.

3.4 Word Embedding Feature Extraction

Based on the previous results [Rahimi et al., 2014, Rahimi et al., 2017], the interpretable rubric-based features outperform competitive baselines. However, there are limitations in their feature extraction method. It cannot extract all examples mentioned by the essay due to the use of simple exact matching.

First, students use their own vocabularies other than words in the crafted list. For instance, some students use the word "power" instead of "electricity" from the crafted list.

Second, according to our corpora, students at the upper elementary level make spelling mistakes, and sometimes they make mistakes in the same way. For example, around 1 out of 10 students misspell "poverty" as "proverty" instead. Therefore, evidence with student spelling mistakes cannot be extracted. However, the evidence dimension of RTA does not penalize students for misspelling words. Although manual spelling corrections indeed improves performance, it is not significantly [Rahimi et al., 2014, Rahimi et al., 2017].

Finally, tenses used by students can sometimes be different from that of the article. Although a stemming algorithm can solve this problem, sometimes there are words that slip through the process. For example, "went" is the past tense of "go", but stemming would miss this conjugation. Therefore, "go" and "went" would not be considered a match.

To address the limitations above, we introduced the Word2vec (the skip-gram (SG) and the continuous bag-of-words (CBOW)) word embedding model [Mikolov et al., 2013a] into the feature extraction process. By mapping words from the vocabulary to vectors of real numbers, the similarity between two words can be calculated. Words with high similarity can be considered a match. Because words in the same context tend to have similar meaning, they would therefore have higher similarity.

We use the word embedding model as a supplement to the original feature extraction process, and use the same searching window algorithm [Rahimi et al., 2014, Rahimi et al., 2017]. If a word in a student's essay is not exactly the same as the word in the crafted list, the cosine similarity between these two words is calculated by the word embedding model. We consider them matching, if the similarity is higher than a threshold.

In Figure 3, the phrases in italics are examples extracted by the existing feature extraction method. For instance, "water was connected to the hospital" can be found because "water" and "hospital" are exactly the same as words in the crafted list. However, "for all the common dieases" cannot be found due to misspelling of "disease". Additional examples that can be extracted by the word embedding model are in bold.

3.5 Experimental Setup

We configure experiments to test several hypotheses:

H1: the model with the word embedding trained on our own corpus will outperform or at least perform equally well as the baseline (denoted by *Rubric*) [Rahimi et al., 2014].

H2: the model with the word embedding trained on our corpus will outperform or at least perform equally well as the model with off-the-shelf word embedding models.

H3: the model with word embedding trained on our own corpus will generalize better across students of different ages. Note that while all models with word embeddings use the same features as the *Rubric* baseline, the feature extraction process was changed to allow non-exact matching via the word embeddings.

We stratify each corpus into 3 parts: 40% of the data are used for training the word embedding models; 20% of the data are used to select the best word embedding model and best threshold (this is the development set of our model); and another 40% of data are used for final testing.

For word embedding model training, we also add essays not graded by the first rater

(*Space* has 229, MVP_L has 222, MVP_H has 296, and MVP has 518) to 40% of the data from the corpus in order to enlarge the training corpus to get better word embedding models. We train multiple word embedding models with different parameters, and select the best word embedding model by using the development set.

Two off-the-shelf word embeddings are used for comparison. The first vectors have 300 dimensions and were trained on a newspaper corpus of about 100 billion words [Mikolov et al., 2013b]. The other vectors have 400 dimensions, with the context window size of 5, 10 negative samples and subsampling [Baroni et al., 2014].

We use 10 runs of 10-fold cross validation in the final testing, with Random Forest (max-depth = 5) implemented in Weka [Witten et al., 2016] as the classifier. This is the setting used by [Rahimi et al., 2014, Rahimi et al., 2017]. Since our corpora are imbalanced with respect to the four evidence scores being predicted (Table 3), we use SMOTE oversampling method [Chawla et al., 2002]. This involves creating "synthetic" examples for minority classes. We only oversample the training data. All experiment performances are measured by Quadratic Weighted Kappa (QWK).

3.6 Results and Discussion

Results for H1. The results shown in Table 8 partially support this hypothesis. The skip-gram embedding yields a higher performance or performs equally well as the rubric baseline on most corpora, except for MVP_H. The skip-gram embedding significantly improves performance for the lower grade corpus. Meanwhile, the skip-gram embedding is always significantly better than the continuous bag-of-words embedding.

Results for H2. Again, the results shown in Table 8 partially support this hypothesis. The skip-gram embedding trained on our corpus outperform Baroni's embedding on *Space* and MVP_L. While Baroni's embedding is significantly better than the skip-gram embedding on MVP_H and MVP.

Results for H3. We train models from one corpus and testing it on 10 disjointed sets of the other test corpus, and we do it 10 times and average the results in order to perform

		Off-the-Shelf		On Our Corpus	
Corpus	**Rubric(1)**	**Baroni(2)**	**Mikolov(3)**	**SG(4)**	**CBOW(5)**
Space	0.606(2)	0.594	0.606(2)	**0.611(2,5)**	0.600(2)
MVP_L	0.628	0.666(1,3,5)	0.623	**0.682(1,2,3,5)**	0.641(1,3)
MVP_H	**0.599(3,4,5)**	0.593(3,4,5)	0.582(5)	0.583(5)	0.556
MVP	0.624(5)	**0.645(1,3,4,5)**	0.634(1,5)	0.634(1,5)	0.614

Table 8: The performance (QWK) of the off-the-shelf embeddings and embeddings trained on our corpus compared to the rubric baseline on all corpora. The numbers in parenthesis show the model numbers over which the current model performs significantly better. The best results in each row are in bold.

significance testing. The results shown in Table 9 support this hypothesis. The skip-gram word embedding model outperform all other models.

			Off-the-Shelf		On Our Corpus	
Train	**Test**	**Rubric(1)**	**Baroni(2)**	**Mikolov(3)**	**SG(4)**	**CBOW(5)**
MVP_L	MVP_H	0.582(3)	0.609 (1,3,5)	0.555	**0.615(1,2,3,5)**	0.596(1,3)
MVP_H	MVP_L	0.604	0.629(1,3,5)	0.620(1,5)	**0.644(1,2,3,5)**	0.605

Table 9: The performance (QWK) of the off-the-shelf embeddings and embeddings trained on our corpus compared to the rubric baseline. The numbers in parenthesis show the model numbers over which the current model performs significantly better. The best results in each row are in bold.

As we can see, the skip-gram embedding outperforms the continuous bag-of-words embedding in all experiments. One possible reason for this is that the skip-gram is better than the continuous bag-of-words for infrequent words [Mikolov et al., 2013b]. In the continuous bag-of-words, vectors from the context will be averaged before predicting the current word, while the skip-gram does not. Therefore, it remains a better representation for rare words. Most students tend to use words that appear directly from the article, and only a small portion of students introduce their own vocabularies into their essays. Therefore, the word embedding is good with infrequent words and tends to work well for our purposes.

In examining the performances of the two off-the-shelf word embeddings, Mikolov's embedding cannot help with our task, because it has less preprocessing of its training corpus. Therefore, the embedding is case sensitive and contains symbols and numbers. For example, it matches "2015" with "000". Furthermore, its training corpus comes from newspapers, which may contain more high-level English that students may not use, and professional writing has few to no spelling mistakes. Although Baroni's embedding also has no spelling mistakes, it was trained on a corpus containing more genres of writing and has more preprocessing. Thus, it is a better fit to our work compared to Mikolov's embedding.

In comparing the performance of the skip-gram embedding and Baroni's embedding, there are many differences. First, even though the skip-gram embedding partially solves the tense problem, Baroni's embedding solves it better because it has a larger training corpus. Second, the larger training corpus contains no or significantly fewer spelling mistakes, and therefore it cannot solve the spelling problem at all. On the other hand, the skip-gram embedding solves the spelling problem better, because it was trained on our own corpus. For instance, it can match "proverty" with "poverty", while Baroni's embedding cannot. Third, the skip-gram embedding cannot address a vocabulary problem as well as the Baroni's embedding because of the small training corpus. Baroni's embedding matches "power" with "electricity", while the skip-gram embedding does not. Nevertheless, the skip-gram embedding still partially addresses this problem, for example, it matches "mosquitoes" with "malaria" due to relatedness. Last, Baroni's embedding was trained on a corpus that is thousands of times larger than our corpus. However, it does not address our problems significantly better than the skip-gram embedding due to generalization. In contrast, our task-dependent word embedding is only trained on a small corpus while outperforming or at least performing equally well as Baroni's embedding.

Overall, the skip-gram embedding tends to find examples by implicit relations. For instance, "winning against poverty possible achievable lifetime" is an example from the article and in the meantime the prompt asks students "Did the author provide a convincing argument that winning the fight against poverty is achievable in our lifetime?". Consequently, students may mention this example by only answering "Yes, the author convinced me.". However, the skip-gram embedding can extract this implicit example.

3.7 Conclusion

We have presented several simple but promising uses of the word embedding method that improves evidence scoring in corpora of RTA written by upper elementary students. Although word embedding is pretty standard these days, other researches use word embedding for obtaining document representation directly. It is hard to interpret the meaning of the representation other than calculating semantic similarity between each other. The stand-alone representation cannot be explained. However, our method uses word embedding for feature extraction. The features themselves are interpretable, which makes them useful for other downstream applications, such as the AWE system. Our experiment results show that features extracted by our new method improve the performance of the learning model. Although pre-trained embedding models show better ability to resolve different vocabularies problem, they do not resolve the spelling problem. Furthermore, there is only a limited impact on scoring. In contrast, a task-dependent word embedding model trained on our small corpus was the most helpful in improving the baseline model.

From the other side, although this simple feature extraction method improves model performance, neural network models show a stronger ability for modeling essays. Therefore, in the next chapter, we present a neural network model that is modeling both essays and the source article directly. Besides, the extracted features were proved useful in a downstream application named eRevise [Zhang et al., 2019]. However, the expert effort is still needed to extract Topical Components. In Chapter 6 and Chapter 7, we present models that extracts Topical Components automatically.

4.0 Co-Attention Based Neural Network for Source-Dependent Essay Scoring

4.1 Introduction

Because neural network models show a stronger ability to model text than feature based models, researchers introduced neural network models into this area. However, other models focus on grading essays in a general and universal way, which means the model does not optimize for any single type of writing task. However, different types of writing tasks have their own characteristics. A one size fits all model always have a shortage. For example, the source-dependent essay scoring, the source article should be an essential external knowledge when grading the essay.

In this chapter, we present an investigation of using a co-attention based neural network for source-dependent essay scoring. We use a co-attention mechanism to help the model learn the importance of each part of the essay more accurately. Also, this work shows that the co-attention based neural network model provides reliable score prediction of source-dependent responses. We evaluate our model on two source-dependent response corpora. Results show that our model outperforms the baseline on both corpora. We also show that the attention of the model is similar to the expert opinions with examples. Besides, we use examples to show that our model can assign reasonable attention scores to different sentences in the essay. This work is illustrated in [Zhang and Litman, 2018].

4.2 Related Work

Previous research in AES including our approach from the prior chapter needed feature engineering. In very early work, Page [Page, 1968] developed an AES tool named Project Essay Grade (PEG) by only using linguistic surface features. A more recent well-known AES system is E-Rater [Burstein et al., 1998], which employs many more natural language processing (NLP) technologies. Later, Attali and Burstein [Attali and Burstein, 2004] released

E-Rater V2, where they created a new set of features to represent linguistic characteristic related to organization and development, lexical complexity, prompt-specific vocabulary usage, etc. Similarly to [Page, 1968], this system used regression equations for assessment of student essays. Such systems are also used to assess responses to English tests, for example, the Graduate Record Exam (GRE) and the Test of English as a Foreign Language(TOEFL). One limitation of all of the above models is that all need handcrafted features for training the model. In contrast, our model uses a neural network for the AES task and thus does not require feature engineering.

Recently, neural network models have been introduced into AES, making the development of handcrafted features unnecessary or at least optional. Alikaniotis et al. [Alikaniotis et al., 2016] and Taghipour and Ng [Taghipour and Ng, 2016] presented AES models that used Long Short Term Memory (LSTM) networks. Differently, Dong and Zhang [Dong and Zhang, 2016] used a Convolutional Neural Network (CNN) model for essay scoring by applying two CNN layers on both the word level and then sentence level. Later, Dong and Zhang [Dong et al., 2017] presented another work that uses attention pooling to replace the mean over time pooling after the convolutional layer in both word level and sentence levels. However, none of these neural network grading models consider the source article if it exists. In this chapter, we introduce a neural network model that takes the source article into account by using a co-attention mechanism instead of the self-attention mechanism of prior work.

Although some models reached better performance on the ASAP corpus [Tay et al., 2018, Nadeem et al., 2019, Liu et al., 2019], the same, they do not take the source article into account if it exists. Besides, those models use more complicated network structures, while our model is relatively simple.

Our work not only focuses on essay assessment using a holistic score, but also evaluates a particular dimension of argument-oriented writing skills, namely use of Evidence. Louis and Higgins [Louis and Higgins, 2010] analyze only the content of essays by detecting off-topic essays. Ong et al. [Ong et al., 2014] used argumentation mining techniques to evaluate if students use enough evidence to support their positions. However, these two prior studies are not suitable for our task because they did not measure the use of content or evidence from a source article. With respect to source-based dimensional essay analysis, Rahimi

et al. [Rahimi et al., 2014, Rahimi et al., 2017] developed a set of rubric-based features that compared a student's essay and a source article in terms of number of related words or paraphrases. Zhang and Litman [Zhang and Litman, 2017] improved their model by introducing word embedding into the feature extraction process to extract relationships previously missed due to lexical errors or use of different vocabulary. However, in both of these studies, human effort was still necessary for pre-processing the source article, for example, by having experts manually create a list of important words and phrases in the article which the system would compare with features extracted from the student's essay. In contrast, our work does not need any human effort to analyze the source article before essay grading. Although [Rahimi and Litman, 2016] investigated extracting example lists by using LDA [Blei et al., 2003] model, the data-driven model missed an example when there was no essay mentioning the example.

4.3 Model

Our network is inspired by the hierarchical neural network model [Dong et al., 2017]. In their model, they considered each essay as a sequence of sentences rather than a sequence of words. Their model has three parts. First, they used a convolutional layer and attention pooling layer to get sentence representation. Second, they used an LSTM layer and another attention pooling layer for document representation. Finally, they used a sigmoid layer for score prediction.

Differently from their model, our model replaces the attention pooling layer for document representation with a bi-directional attention flow layer and an additional modeling layer [Seo et al., 2017]. By doing so, our model considers students' essays associated with a source article and this attention mechanism captures the relationship between the essay and the source article. In particular, a higher attention score will be assigned to sentences that are mentioned in the article but less mentioned in other essays. Our model is a hierarchical neural network and consists of seven layers. Figure 5 shows the structure of our network. The layers in the dashed box were presented by [Dong et al., 2017]. The sentence level

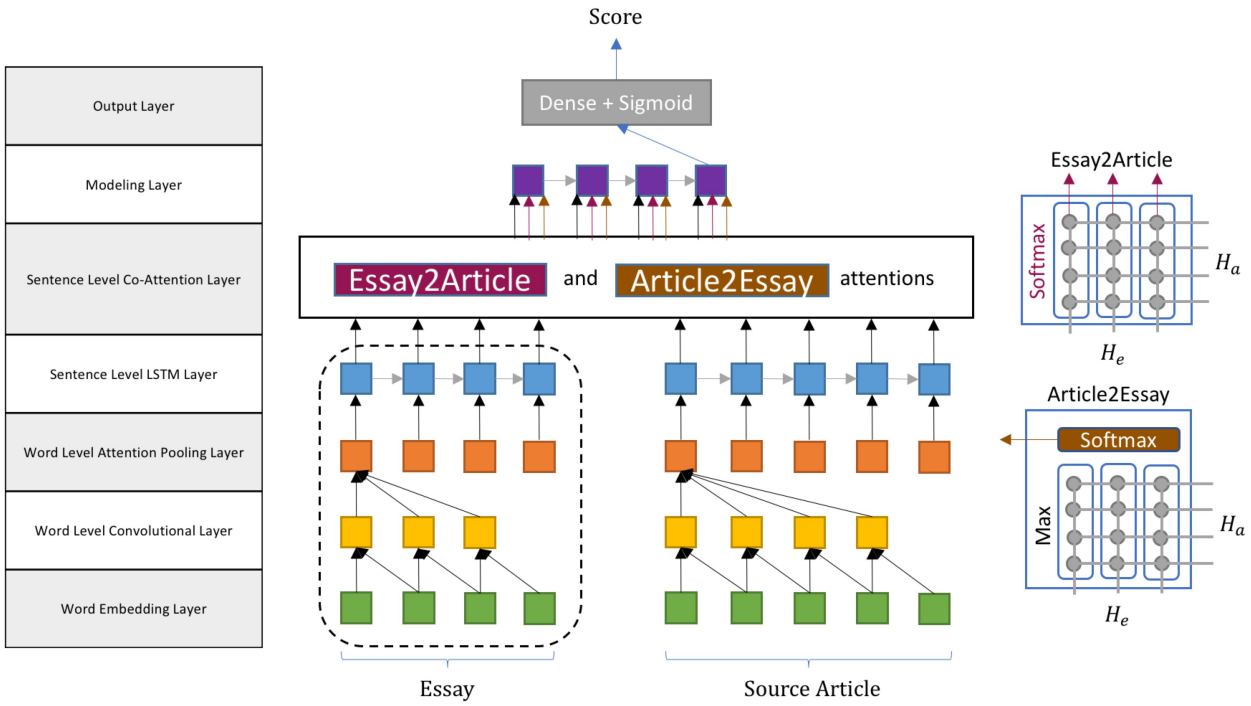


Figure 5: The co-attention based neural network structure.

co-attention layer was presented by [Seo et al., 2017].

4.3.1 Word Embedding Layer

This layer maps each word in sentences to a high dimension vector. We use the GloVe pre-trained word embeddings [Pennington et al., 2014] to obtain the word embedding vector for each word. It was trained on 6 billion words from Wikipedia 2014 and Gigaword 5. It has 400,000 uncased vocabulary items. The dimensionality of GloVe in our model is 50 dimensions. The outputs of this layer are two matrices, $L_E \in \mathbb{R}^{S_e \times W_e \times d_L}$ for the essay and $L_A \in \mathbb{R}^{S_a \times W_a \times d_L}$ for the article, where S_e, S_a, W_e, W_a, and d_L are number of sentences of the essay and the article, length of sentences of the essay and the article, and the embedding size, respectively. A dropout is applied after the word embedding layer [Dong et al., 2017].

4.3.2 Word Level Convolutional Layer

In this layer, we perform 1D convolution over the word representations of both L_E and L_A, so that we can get local representation of each sentence. For each word w_i in each sentence, we perform 1D convolution:

$$p_i = g([w_i : w_{i+k-1}] \cdot U_p + b_p) \tag{1}$$

where g is a nonlinear activation, k is the kernel size, U_p is the filter weight matrix, and b_p is the bias vector. The outputs of this layer are $C_e \in \mathbb{R}^{S_e \times P_e \times d_C}$ for the essay and $C_a \in \mathbb{R}^{S_a \times P_a \times d_C}$ for the article, where P_e and P_a are filtered lengths of sentences of the essay and the article, respectively. d_C is the number of filters of the 1D convolution layer.

4.3.3 Word Level Attention Pooling Layer

After the convolutional layer, a pooling layer is demanded to obtain the sentence representations. In this layer, we follow the same design presented by [Dong et al., 2017]. The attention pooling is defined as equations below:

$$m_i = tanh(U_m \cdot p_i + b_m) \tag{2}$$

$$v_i = \frac{e^{u_v \cdot m_i}}{\sum e^{u_v \cdot m_j}} \tag{3}$$

$$s = \sum v_i p_i \tag{4}$$

where U_m, u_v and b_m are weight matrix, vector, and bias vector, respectively. m_i and v_i are attention vector and attention weight for p_i. The outputs of this layer are $A_e \in \mathbb{R}^{S_e \times d_C}$ for the essay and $A_a \in \mathbb{R}^{S_a \times d_C}$ for the article.

4.3.4 Sentence Level LSTM Layer

In this layer, we use a Long Short-Term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997] over the sentence representations of the essay and the article to capture contextual evidence from previous sentences to refine the sentence representation.

The LSTM unit is a special kind of RNN unit which has long-term dependency learning ability. LSTMs use three gates to control information flow to avoid the long-term dependency problem by forgetting or remembering information in each LSTM unit. They are an input gate, a forget gate, and an output gate. The following equations define the LSTM unit:

$$f_t = \sigma(W_f \cdot [h_{t-1}, s_t] + b_f) \tag{5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, s_t] + b_i) \tag{6}$$

$$\tilde{c}_t = tanh(W_c \cdot [h_{t-1}, s_t] + b_c) \tag{7}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{8}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, s_t] + b_o) \tag{9}$$

$$h_t = o_t * tanh(c_t) \tag{10}$$

where s_t and h_t are the input sentence and the output state of time t, respectively. W_f, W_i, W_c, and W_o are weight matrices. b_f, b_i, b_c, and b_o are bias vectors. σ is the sigmoid function, and $*$ is element-wise multiplication. The output of this layer are $H_e \in \mathbb{R}^{S_e \times d_H}$ for the essay and $H_a \in \mathbb{R}^{S_a \times d_H}$ for the article, where d_H is the dimensionality of the output.

4.3.5 Sentence Level Co-Attention Layer

The concept of this layer is presented by [Seo et al., 2017] in the part of attention flow layer. This layer links information from H_e and H_a, and generates a collection of article aware features vector of essay sentences. The attention is computed in two directions, from essay to article, and vice versa. Both attention scores are figured from a similarity matrix by the following equation:

$$Sim = W_{sim}^T \cdot [he_t; ha_j; ha_t * ha_j^T] + b_{sim} \tag{11}$$

where W_{sim} is weight matrix, he_t and ha_j are t_{th} row vector of H_e and j_{th} row vector of H_a, b_{sim} is bias vector. $*$ is element-wise multiplication. $[;]$ is vector concatenation. After obtaining the similarity matrix $Sim \in \mathbb{R}^{S_e \times S_a}$, we compute the attention in two directions.

Essay to Article Attention measures which sentences in the article are similar to each sentence in students' essays. The following equations define the essay to article attention:

$$a_{ea} = softmax(Sim) \tag{12}$$

$$\tilde{H}_a = a_{ea} H_a \tag{13}$$

where $a_{ea} \in \mathbb{R}^{S_e \times S_a}$ represents the attention score of each sentence in the article associate with each sentence in the essay, $softmax$ is performed across each row. The output of this $\tilde{H}_a \in \mathbb{R}^{S_e \times d_H}$.

Article to Essay Attention measures which sentences in the essay have the closest meaning to one of the sentences in the article. The following equations define the article to essay attention:

$$a_{ae} = softmax(max_{col}(Sim)) \tag{14}$$

$$\tilde{h}_e = a_{ae}^T H_e \tag{15}$$

where $a_{ae} \in \mathbb{R}^{S_e}$, max_{col} is a maximum function performed across the column, and $\tilde{h}_e \in \mathbb{R}^{d_H}$. Because max_{col} will find out which sentence in the article has the closest meaning to each sentence in the essay, so $\tilde{h}_e$ represents the attention score of the most important sentence in the essay associated with the article. After tiling S_e times, the final output of this layer is $\tilde{H}_e \in \mathbb{R}^{S_e \times d_H}$.

The final output G is a concatenated matrix of H_e, $\tilde{H}_e$, and $\tilde{H}_a$ defined by:

$$G = [H_e; \tilde{H}_a; H_e * \tilde{H}_a; H_e * \tilde{H}_e] \tag{16}$$

where $*$ is element-wise multiplication, and $[;]$ is concatenation, H_e is the original representation of essay, $\tilde{H}_a$ is the essay to article attention, $H_e * \tilde{H}_a$ is the self-aware representation, and $H_e * \tilde{H}_e$ is article-aware representation. Therefore, the output of this layer is $G \in \mathbb{R}^{S_e \times 4d_H}$, the article-aware representation of each sentence in the essay.

4.3.6 Modeling Layer

G is the representation of each sentence, and we need the representation of the essay. Therefore, we introduce another LSTM layer for modeling the essay and only use the output of the final LSTM unit as the output of this layer $M \in \mathbb{R}^{d_M}$, where d_M is the dimensionality of the output of LSTM units.

4.3.7 Output Layer

After obtaining the essay representation M, a linear layer with sigmoid activation will predict the final output. The following equation defines the output layer:

$$y = sigmoid(W_o M + b_o) \tag{17}$$

where W_o is weight vector, and b_o is bias vector. y is the final predicted score of the essay.

4.4 Training

Loss. [Dong et al., 2017] used mean squared error (MSE) loss, thus we use the same loss function. MSE evaluates the average of squared error between the predicted score and the gold standard. Thus it is widely used in regression tasks. The following equation defines MSE:

$$mse(y, y') = \frac{1}{N} \sum_{i=1}^{N} (y_i - y'_i)^2 \tag{18}$$

where y_i is the predicted score, y'_i is the gold standard, N it the total number of samples.

Optimization. The optimizer we use is RMSprop [Dauphin et al., 2015]. The initial learning rate is 0.001, momentum is 0.9, and Dropout rate is 0.5 for preventing overfitting. These setting are the same as used by [Dong et al., 2017].

4.5 Experimental Setup

We configure experiments to test three hypotheses:

H1: the model we proposed (denoted by CO-ATTN) will outperform or at least perform equally well as the baseline (denoted by SELF-ATTN) [Dong et al., 2017] on four ASAP essay corpora in the holistic score prediction task.

H2: the model we proposed will outperform or at least perform equally well as the baseline on two RTA corpora in the Evidence score prediction task.

H3: the model we proposed will outperform or at least perform equally well as the non-neural network baselines on both corpora.

We use NLTK [Bird et al., 2009] for text preprocessing. The vocabulary size of the data is limited to 4000, and all scores are scaled to the range [0, 1], following [Taghipour and Ng, 2016] and [Dong et al., 2017]. In particular, the 4000 most frequent words are preserved, with all other words treated as unknowns. The assessment scores will be converted back to their original range during evaluation. We use Quadratic Weighted Kappa (QWK) to evaluate our model. QWK is not only the official criteria of ASAP corpus, but also adopted as evaluation

metric in [Rahimi et al., 2014, Taghipour and Ng, 2016, Dong et al., 2017, Rahimi et al., 2017, Zhang and Litman, 2017] for both ASAP and RTA corpora.

We use 5-fold cross-validation because both RTA and ASAP corpora have no released labeled test data. We split all corpora into 5 folds. For the ASAP corpus, the partition is the same as the setting presented by [Taghipour and Ng, 2016]. For the RTA corpus, since there is no prior work to split the corpus, we separate it into 5 folds randomly. 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.

To select the best model, we trained each model on 100 epochs and evaluated on the development set after each epoch. The best model is the model with the best QWK on the development set. This is done five times, once for each partition in the cross-validation. Then the average QWK score from these five evaluations on the test set is reported. Paired t-tests are used for significance tests with $p < 0.05$. Table 10 shows all hyper-parameters for training.

The code of SELF-ATTN are provided by [Dong et al., 2017], they used Keras [Chollet et al., 2015] 1.1.1 and Theano [Theano Development Team, 2016] 0.8.2 as the backend. Because we are using Keras 2.1.3 and TensorFlow [Abadi et al., 2015] 1.4.0 as the backend, we ran all experiments with our frameworks. Therefore, the numbers of SELF-ATTN have small differences to the numbers reported by the baseline model.

For non-neural network baselines, we introduce the SVR and BLRR baselines [Phandi et al., 2015] for the ASAP corpus, and SG baseline [Zhang and Litman, 2017] for the RTA corpus.

SVR and BLRR models use Enhanced AI Scoring Engine (EASE)[1] to extract four types of features, such as length, part of speech, prompt, and the bag of words. Then they use SVR and BLRR as the classifiers, respectively. We do not perform any significance test on both SVR and BLRR because we do not have detailed experiment data. Therefore, we only report the result presented in [Phandi et al., 2015].

SG model extracts evidence features based on hand-crafted topic and example lists, and uses random forest tree as the classifier. We follow the same data partition. However,

[1] https://github.com/edx/ease

we only use the training set for training and the testing set for testing while ignoring the development set so that we can perform the same paired t-tests in the experiments.

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 10: Hyper-parameters of training.

4.6 Results and Discussion

Results for H1. The results shown in Table 11 support this hypothesis. The CO-ATTN model yields higher performance than the SELF-ATTN model on all ASAP prompts. However, the CO-ATTN model only significantly outperforms the SELF-ATTN model on Prompt 3.

Results for H2. Again, the results shown in Table 11 support this hypothesis. The CO-ATTN model yields higher performance than the SELF-ATTN model, significantly on both of the RTA corpora.

Results for H3. The results shown in Table 11 still support this hypothesis. The CO-ATTN model yields higher performance than all non-neural network baselines.

Prompts	SVR	BLRR	SG	SELF-ATTN	CO-ATTN
MVP	NA	NA	0.653	0.681†	**0.697**∗†
$Space$	NA	NA	0.632	0.669†	**0.684**∗†
$ASAP_3$	0.630	0.621	NA	0.677	**0.697**∗
$ASAP_4$	0.749	0.784	NA	0.807	**0.809**
$ASAP_5$	0.782	0.784	NA	0.806	**0.815**
$ASAP_6$	0.771	0.775	NA	0.809	**0.812**

Table 11: The performance (QWK) of the baselines and our model. $*$ indicates that the model QWK is significantly better than the SELF-ATTN ($p < 0.05$). † indicates that the model QWK is significantly better than the SG ($p < 0.05$). The best results in each row are in bold.

The results show that in our tasks, the neural network approaches are better than non-neural network baselines. One possible reason is the final representation of the essay from neural network contains more information. However, some of the information might be ignored by hand-crafted features. For example, the importance of different evidence in RTA task is not considered in the SG model. It treats all evidence equally. However, the neural network models capture this information automatically.

Apparently, the CO-ATTN model performs better in the RTA tasks, because it always significantly outperforms the SELF-ATTN model. One possible reason is that the RTA task only considers the Evidence score. The CO-ATTN model is more suitable for the Evidence score prediction task because it can find pieces of evidence that appear in both students' essays and the source article better. In contrast, the SELF-ATTN model only considers students' essays associated with the scores. In this case, if a piece of evidence is not mentioned by students, this data-driven model cannot distinguish it. Consequently, some important pieces of evidence will be assigned to a lower weight. However, the CO-ATTN model considers not only the students' essays but also the source article. In other words,

if an important piece of evidence is not mentioned by too many students, but it is in the source article, the CO-ATTN model will assign this sentence higher attention.

In the ASAP holistic score prediction task, although we still see a benefit in using the CO-ATTN model, it is reduced. In this case, the benefit we saw in the Evidence dimension from the CO-ATTN model becomes less significant because the model also needs to consider more aspects of the essay, such as organization, grammar mistakes, and so on. Our results suggest that the co-attention mechanism of the CO-ATTN model cannot capture these aspects significantly better than the SELF-ATTN model. Therefore, the CO-ATTN model only significantly outperforms the SELF-ATTN model on Prompt 3.

In Table 12, we list 10 sentences from student MVP essays and their associated attention scores. Because we have a list of examples manually extracted by our experts as important evidence from the MVP source article, examining RTA data helps us understand the attention score assigned by our model. Bolded are examples extracted by the expert from the source article that the student includes in the essay. A lower attention score means this sentence is less important. Otherwise, the score is high. As we can see, sentences 1, 2, 3, and 4 are low attention sentences, sentences 5, 6, and 7 are mid attention sentences, and sentences 8, 9, and 10 are high attention sentences. The attention scores reflect the importance of these sentences accurately.

Sentence 1 is a short and general sentence related to the source article, but it has no specific evidence from it. Sentence 2 even has no content related to the source article. Sentence 3 has many details related to the source article. However, it still has no evidence directly from the source article. Sentence 4 mentions "*The author did convince me that winning the fight against poverty is achievable in our lifetime*" which comes from both the prompt and the source article, but this statement is so general that almost every student mentions this statement in the essay which makes this statement not distinguishable. For these reasons, these four sentences receive low attention scores.

Although sentence 5 is short, it mentions one piece of evidence. Sentence 6 talks about farming which is a topic from the source article. In the article, the things listed in this sentence are things the farmer needs to worry about. However, this sentence indicates "*the farmer don't have to worry*" because of the MVP project. Sentence 7 also mentions

conditions of hospitals nowadays. However, it mentions not only water but also electricity which is more than Sentence 5. For these reasons, these three sentences receive mid attention scores from low to high.

The last three sentences receive high attention scores because they all use more pieces of evidence directly from the source article. Sentence 8 talks about the school, and Sentence 9 talks about the hospital. Sentence 10 talks about farming. However, sentence 10 receives the highest attention score, because it mentions evidence from both before and after the MVP project.

From these sentences, we can also see that the attention score depends on neither the length of the sentence nor only the specificity of the sentence. It instead depends on how many important pieces of evidence there are in the sentence. For example, Sentence 3 is long and talks about some details of our modern life. Although it also talks about quality materials or better housing and clothing compared to people living in Kenya, it receives a low attention score because there is no specific evidence directly from the source article. In contrast, Sentence 9 is shorter than Sentence 3. However, it receives a higher attention score because it mentions many pieces of evidence from the source article.

Overall, the CO-ATTN model seems to capture the importance of sentences by assigning reasonable attention scores based on the relevance of the sentence to the source article.

4.7 Conclusion

In this work, we presented a co-attention based neural network model that outperforms a state-of-the-art attention based neural network model for essay scoring, not only for RTA Evidence assessment but also for holistic assessment of ASAP source-dependent responses. The advantages of our model are that it does not need any expert preprocessing of the source article; the input of this model is only the raw student essay and its source article. Moreover, our model somewhat captures the importance of different pieces of evidence, although it is not specifically designed for this purpose. However, quantitative experiments that can answer whether the attention scores are correlated to the importance of different

pieces of evidence need to be done. Also, this leads to an interesting future investigation, development of a neural network approach that both have an acceptable score prediction, and can simultaneously generate evidence lists from the source article. In Chapter 6, we are going to talk about a model that uses the intermediate output of the co-attention based neural network for extracting Topical Components. Besides, this model only works for source-dependent essay scoring. Although it reaches a better performance in this specific area, it cannot work for other situations that the source article does not exist. Therefore, in Chapter 5, we propose a hybrid model that combines a simpler neural network model and hand-crafted features so that to make the model works for more situations.

No.	Sentences	Attention
1	Life in Kenya is hard.	0.00173
2	In this essay I will give my top 3 reasons why.	0.00174
3	Because like I said, we have more advanced & better & more qualified materials than them, and these days kids & adults are spoiled, we have phones stores, houses & even shoes and clothes.	0.00243
4	The author did **convince me that winning the fight against poverty is achievable in our lifetime** because she showed me how many people in Sauri, Kenya need our help against poverty.	0.00229
5	**Water is connected to the hospitals.**	0.02936
6	So the farmer don't have to worry all the time that him or his **family won't have enough food** to eat and the farmer have to worry that their kids will get hungry and then sick.	0.05580
7	**The hospital aslo has water and electricity**.	0.07746
8	Also, there were **no school fees**, and **the school now serves lunch** for the students because they **didn't have any midday meals** to provide them with **energy they need** to help them with the rest of their days.	0.19483
9	In 2008 though, when they **checked for progress**, **the hospital had medicine**, **free of charge**, with **running water and electricty**.	0.20177
10	Also **farmers could not afford fertilizer and irrigation** but now **they placed irrigation** and have them **fertilizer for the crops**.	0.25855

Table 12: Example attention scores of essay sentences.

5.0 Attention Based Neural Network for Automated Essay Scoring with Hand-crafted Features

5.1 Introduction

In Chapter 4, we presented a co-attention neural network for grading source-based essays. Although it outperforms its baselines, the design of the neural network limited the usage scenario, because a source article is required for learning hand-crafted features from it.

In this chapter, we propose a hybrid model that builds on an attention-based neural network model for AES [Dong et al., 2017], in order to be able to combine hand-crafted features on the sentence and on the word level as well as on the essay level. While enabling the use of hand-crafted features as a side input, our approach offers the neural network the ability to model the hand-crafted features. We hypothesize that the strong modeling ability of neural networks will be able to learn useful knowledge from the hand-crafted features. Within-prompt experiments show that our proposed hybrid model outperforms a neural baseline model, supporting our hypothesis. We also conduct cross-prompt experiments to show the usefulness of our model in a more difficult scenario typical of classroom AES usage.

5.2 Related Work

Historically, most AES research has used feature-based models [Yannakoudakis and Briscoe, 2012, Farra et al., 2015, McNamara et al., 2015, Cummins et al., 2016, Amorim et al., 2018, Louis and Higgins, 2010, Persing and Ng, 2015, Ghosh et al., 2016, Nguyen and Litman, 2018, Persing et al., 2010], which typically require carefully designed hand-crafted features for essay representation and off-the-shelf learning algorithms for model training. Surprisingly, even though neural network models currently dominate most natural language processing research areas, feature-based models still have a role to play in the AES community. For example, the model of Cozma et al. [Cozma et al., 2018] combines bag-of-super-

word-embeddings [Butnaru and Ionescu, 2017] with a string kernel and outperforms most neural network models. Nevertheless, much AES research now uses neural network models because they generally demonstrate state-of-the-art performance compared to feature-based models [Taghipour and Ng, 2016, Alikaniotis et al., 2016, Dong and Zhang, 2016, Dong et al., 2017, Tay et al., 2018, Phandi et al., 2015, Jin et al., 2018, Zhang and Litman, 2018, Nadeem et al., 2019]. The most significant difference between either existing feature-based or neural models and our model is that we propose a hybrid model combining a neural network with hand-crafted features. In our hybrid, rather than playing the leading role in training, the hand-crafted features provide guidance for training the neural network model.

With respect to other hybrid approaches combining a neural network model with hand-crafted features, the model of Liu et al. [Liu et al., 2019], and Uto et al. [Uto et al., 2020] provide state-of-the-are performance on the Automated Student Assessment Prize (ASAP) corpus. However, all combined features are essay-level features (e.g., the vocabulary size of the essay), which means the model only concatenates all hand-crafted features with highly abstracted essay-level information. In contrast, our model provides the possibility to combine hand-crafted features from a smaller linguistic unit, such as sentence level or word level. For example, a sentence-level feature could be the topic distribution of each sentence, and a word-level feature could be the POS tag of each word. In addition, our model takes lower level input as a sequence and models the sequence further. Dasgupta et al. [Dasgupta et al., 2018] presented a hybrid model that uses LSTM Layer to model word-level feature sequence and combines pooling layer output on the essay level. This model is more similar to our model, but the same, our model provides flexibility to incorporate hand-crafted features from all linguistic units. Besides, we use a more complex way to combine hand-crafted features, which is an attention layer. This layer potentially provides the model with a stronger learning ability. Especially for low level hand-crafted features, the attention layer could distinguish the part of the sequence which is more important than others.

Beyond AES, Ko et al., [Ko et al., 2019] present a hybrid model for specificity prediction. However, their model still just concatenates hand-crafted features and a neural network without encoding the hand-crafted features. In contrast, Chen et al., [Chen et al., 2018] propose a hybrid model for automated speech scoring which has similarities with our model.

In particular, they use a pure neural network model to encode the lexical aspect and a hybrid model to model acoustic cues. However, while they use a neural network to encode hand-crafted features, they only concatenate the outputs of both neural networks. Consequently, the final prediction model treats features from both sides equally. In contrast, our hybrid model uses an attention layer to combine two models. Our model focuses on the pure neural network output and uses hand-crafted features as side input. Also, our model is a hierarchical model, which provides the possibility to combine hand-crafted features at different levels.

Recently, the natural language processing research community has been dominated by deep transformer architectures such as BERT [Devlin et al., 2018]. However, such a deep and complex model might not be suitable for AES tasks, because AES tasks typically have relatively small amounts of training data [Mayfield and Black, 2020]. Considering the cost-inefficiency of using such complex architectures, AES tasks often tend to prioritize simpler models. Therefore, in this work, we also start from a relatively simple neural model [Dong et al., 2017]. The performance of this attention-based, hierarchical neural network model on the ASAP corpus is 0.760, while the result of a more complex state-of-the-art neural network model is 0.773 [Liu et al., 2019]. Given the performance similarity, we build on the simpler neural model to demonstrate the utility of our hybrid approach.

5.3 Base Model

The main contribution of this work is to test a new hybrid neural network model, which can learn from both student essays and hand-crafted features. We mainly focus on exploring what hand-crafted features can be combined with the neural network, and how they can be combined. Therefore, we need to select a base neural network model and combine hand-crafted features with it. Since the state-of-the-art deep transformer architectures are not as helpful for the AES task as for other tasks in the NLP area [Mayfield and Black, 2020], we use a relatively simple model as our base model.

The base neural model [Dong et al., 2017] is a hierarchical neural network. In this model, each essay is considered to be a sequence of sentences rather than a sequence of

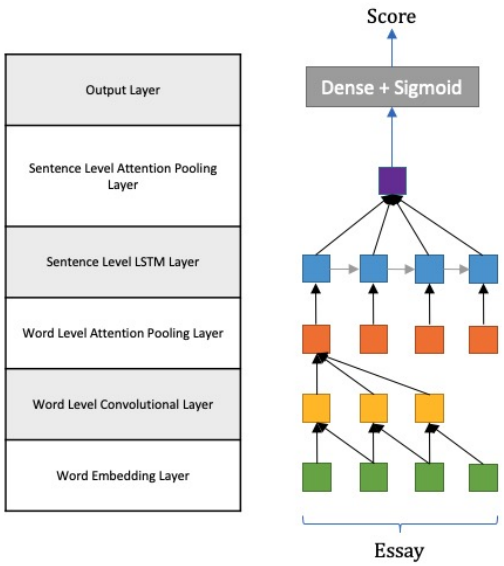


Figure 6: Architecture of the base model.

words. Coherence between words and sentences can thus be learned in two steps, rather than in only one mixture step. The model uses a CNN layer [LeCun et al., 1998] and a self-attention layer for word-level modeling to get sentence representation, and an LSTM layer [Hochreiter and Schmidhuber, 1997] and another self-attention layer for sentence-level modeling. Figure 6 shows the architecture of the base model.

Word Embedding Layer. This layer maps each word in sentences to a high dimension vector. Currently, we are using Glove pre-trained word embeddings [Pennington et al., 2014] to obtain the word embedding vector for each word. Following the same setting from previous work, the pre-trained embedding in the network was trained on 6 billion words from Wikipedia 2014 and Gigaword 5. It has 400,000 uncased vocabulary items, and the dimensionality of the GloVe model is 50 dimensions. As in [Dong et al., 2017], a dropout layer is applied after the word embedding layer to prevent the neural network from overfitting [Srivastava et al., 2014].

Word Level Convolutional Layer. This layer performs 1D convolution over the word representation. The output of this layer is the local representation of each sentence.

Word Level Attention Pooling Layer. This pooling layer is applied over the convolu-

tional layer and is designed to obtain the sentence representation by calculating the weighted sum of each sliding window. The output of this layer is the sentence representations of each essay.

Sentence Level LSTM Layer. We apply a Long Short-Term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997] over the sentence representations to capture contextual evidence from previous sentences to refine the sentence representation.

Sentence Level Attention Pooling Layer. Same as the Word Level Attention Pooling Layer, this pooling layer is applied over the LSTM layer and is designed to capture the essay representation by calculating the weighted sum of each sentence. The output of this layer is the essay representation, which will be passed to the final output layer.

Output Layer. After obtaining the essay representation, a linear layer with sigmoid activation predicts the final output. Note that the model treats AES as a regression problem. This setting provides the flexibility to grade essays with continuous or discrete scores, and with different score ranges.

Loss Function. Mean Squared Error (MSE) loss is the loss function. The MSE evaluates the average of squared error between the predicted score and the gold standard. Therefore, it is widely used in regression tasks.

Optimization. The optimizer of the base model is RMSprop [Dauphin et al., 2015]. Following Dong et al., [Dong et al., 2017], the initial learning rate is 0.001, momentum is 0.9. The dropout rate is 0.5.

5.4 Proposed Hybrid Model

We extend the base model from a pure neural network to a hybrid model that learns from both student essays and hand-crafted features. In this section, we introduce the hybrid model and the hand-crafted features to be tested.

5.4.1 Combination Models

Different hand-crafted features could be extracted from different linguistic levels, such as word-level, sentence-level, and essay-level. For example, a word-level feature could be the POS tag of each word, a sentence-level feature could be the length of each sentence, and an essay-level feature could be the topic distribution of the essay.

Depending on what hand-crafted features to combine, we might have to combine them on the same or a higher model level, such as sentence-level or even essay-level. In the combination model, we use an attention mechanism [Bahdanau et al., 2014] to learn the relation between essays and their hand-crafted features. The attention mechanism calculates a feature-aware essay representation.

Figure 7 and Figure 8 show the architecture of the word-level combination model with a word-level hand-crafted feature, and the architecture of the sentence-level combination model with a sentence-level hand-crafted feature, respectively. In Figure 7, the combination model combines essay-side representation and hand-crafted feature representation before the word level attention pooling layer, while the combination model in Figure 8 combines essay-side representation and hand-crafted feature representation before the sentence level attention pooling layer. By combining the hand-crafted features at different levels, we preserve the information from hand-crafted features from different abstract levels. However, a hand-crafted feature can only be combined on the same or a higher level. For example, an essay-level feature can be combined on the model essay level, but not the sentence level, because the architecture of the base model does not allow this unpacking operation. Figure 9, Figure 10, Figure 11, and Figure 12 show all other possible model architectures.

Since an individual feature may only provide limited extra information for learning, we might need to combine multiple features. Figure 13 shows an example model architecture that combines a word level feature on the model word level, and a sentence level feature on the model sentence level at the same time. One of the advantages of this model design is that the hand-crafted feature combination is modular. This provides flexibility to combine multiple features at one time. We only need to calculate attended representation for each feature, and eventually concatenate them together.

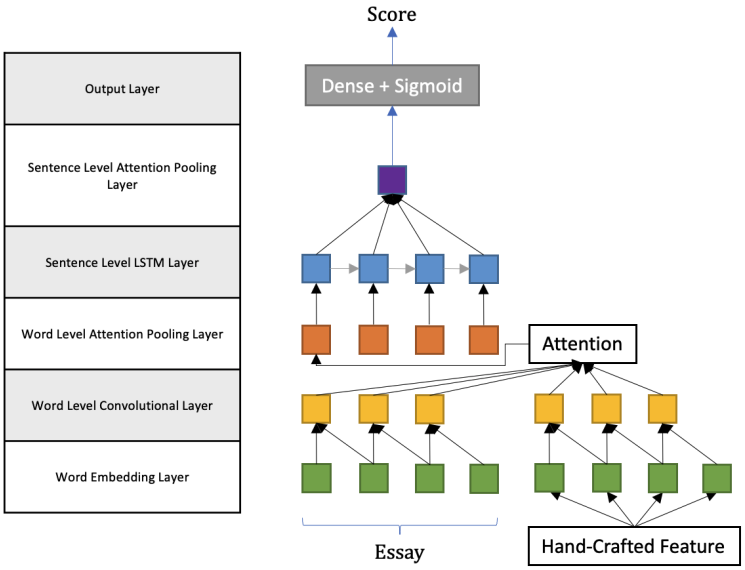


Figure 7: The architecture of the word level combination model with word level hand-crafted feature.

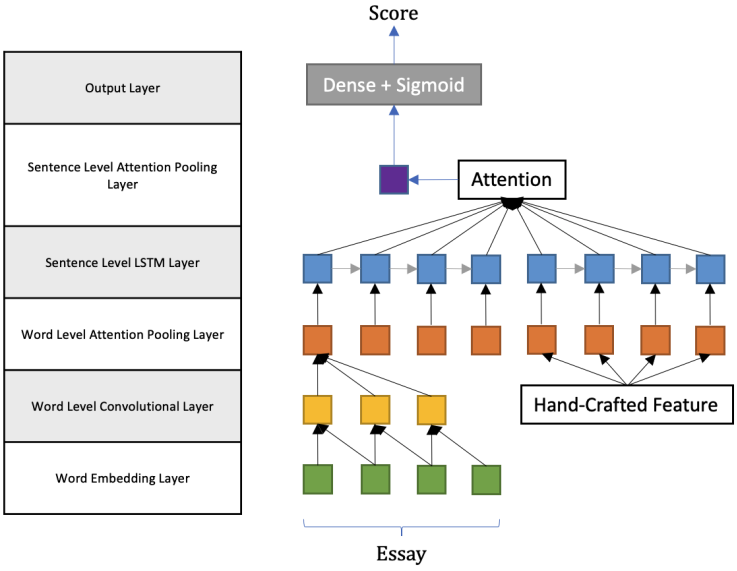


Figure 8: The architecture of the sentence level combination model with sentence level hand-crafted feature.

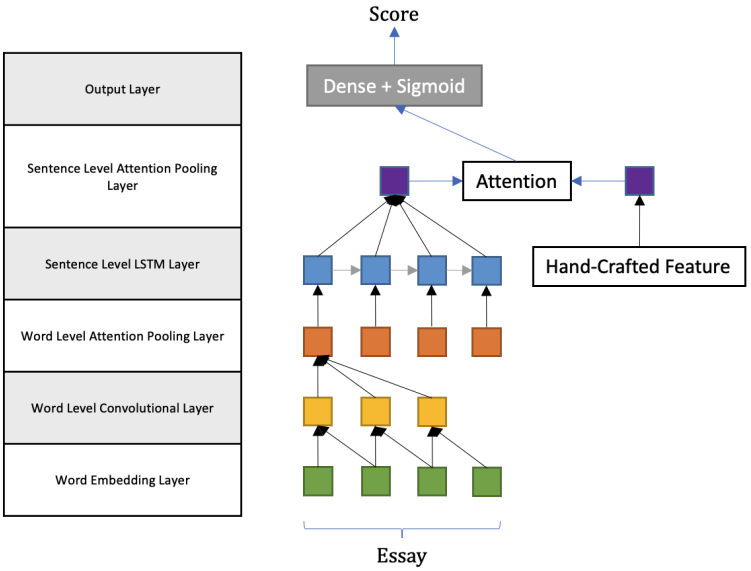


Figure 9: The architecture of the essay level combination model with essay level hand-crafted feature.

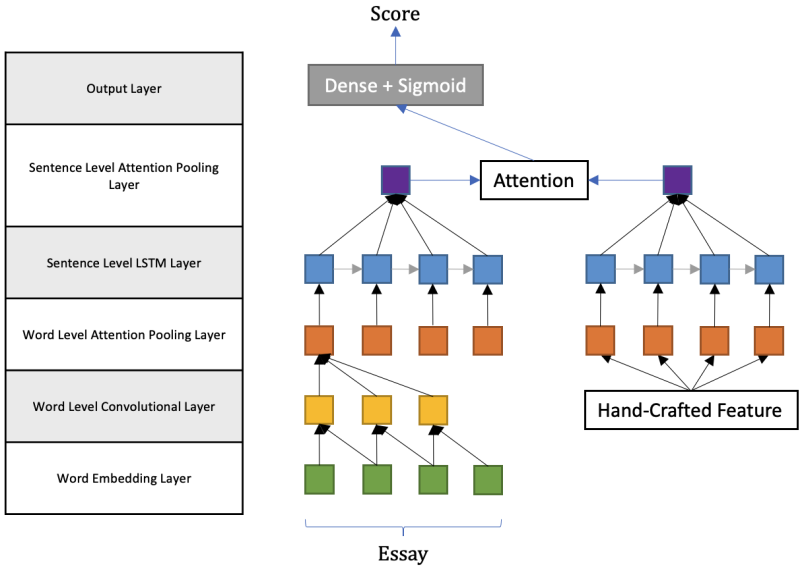


Figure 10: The architecture of the essay level combination model with sentence level hand-crafted feature.

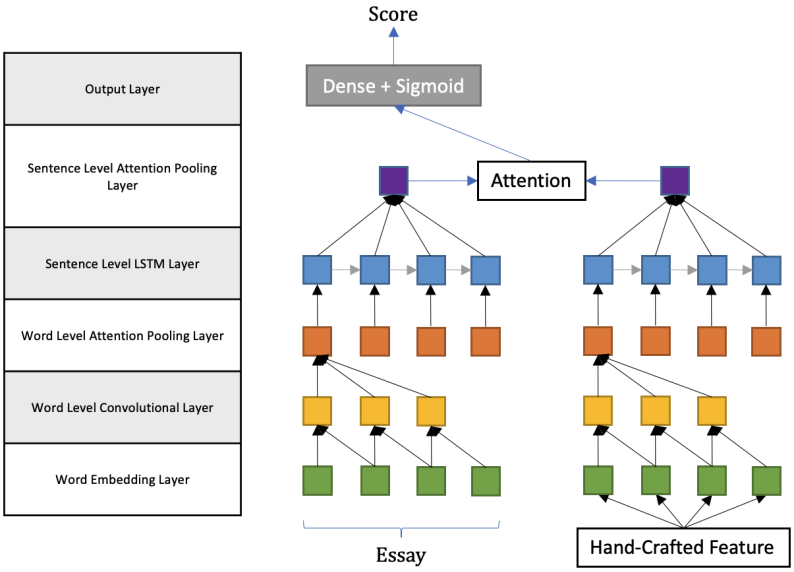


Figure 11: The architecture of the essay level combination model with word level hand-crafted feature.

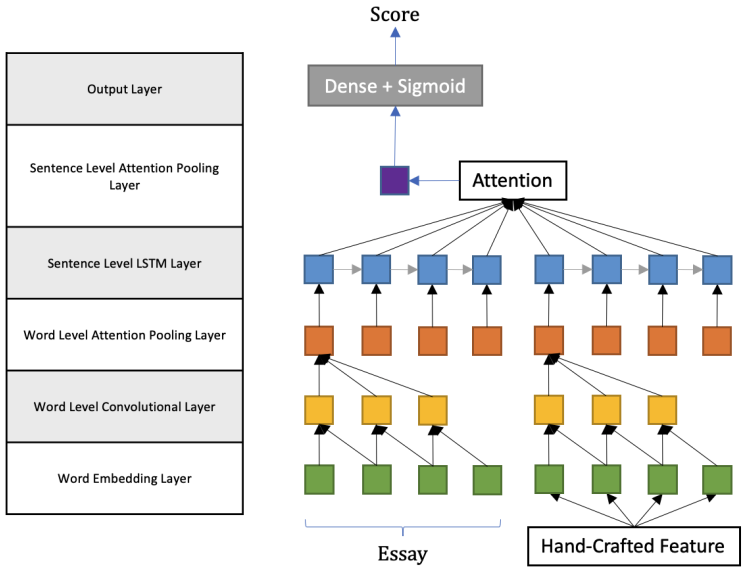


Figure 12: The architecture of the sentence level combination model with word level hand-crafted feature.

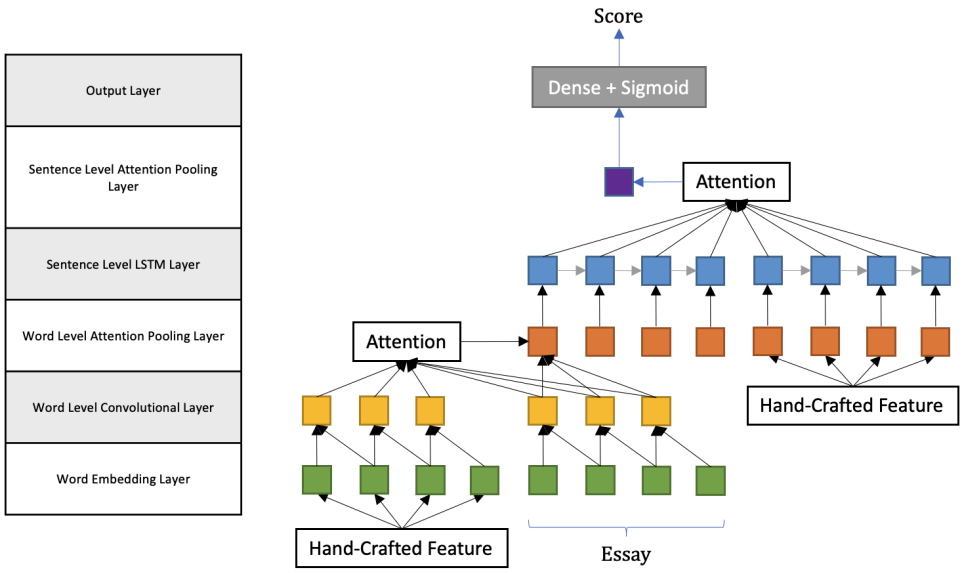


Figure 13: The architecture of a model combining a word level feature and a sentence level feature on word level and sentence level at the same time.

5.4.2 Hand-crafted Features

Hand-crafted features are widely used for AES, and each of them falls into a feature category [Ke and Ng, 2019], e.g., length-based features, lexical features, embeddings, word category features, prompt-relevant features, readability features, syntactic features, argument features, semantic features, and discourse features. We test a few features from each category in this work, except lexical features, embeddings, and semantic features. Lexical features are usually n-grams features. However, if we combine n-grams with a neural network model, the embedding layer is necessary to model each word. Therefore, in this work, we consider the lexical features are similar to embeddings, eventually. Although embeddings and semantic features are powerful for AES, we believe that the neural network model has modeled essays with embedding in a semantic way. Therefore, we consider it is redundant to combine similar features again, especially as the neural network shows a strong ability to model student essays.

Note that there is likely no one-size-fits-all feature that could be used in all writing

Category	Feature
Length-based	word_count_essay
	word_count_sent
Word Category	discourse_conn_word
	sentiment_word
	modal_word
Prompt-relevant	lda_essay
	lda_sent
	lda_word
Readability	readability_essay
	readability_sent
Syntactic	pos_word
	simple_pos_word
Argumentation	argument_word
Discourse	discourse_func_sent
	discourse_func_vec_sent

Table 13: All hand-crafted features to be tested in this work.

tasks. For example, a general prompt such as write about "a time that you failed and learned something useful" would likely have little related topics from an LDA perspective. However, in this section, we introduce all features that we could potentially combine into the base model. We present two feature selection strategies later. The category and feature columns of Table 13 show all features that will be tested in this work.

Length-based Features. These features are widely used for AES, since length is highly positively correlated with essay scores [Attali and Burstein, 2006, Chen and He, 2013, Östling et al., 2013, Phandi et al., 2015, Zesch et al., 2015b]. We include two features based on word count: essay length (denoted by ***word_count_essay***), and sentence length (***word_count_sent***).

Word Category Features. An essay should demonstrate the writer's ability for word usage. This feature could be computed based on an external wordlist or dictionary. However, the wordlist or the dictionary could contain a variant of categories of words such as lexical, syntactic, and semantic [Yannakoudakis and Briscoe, 2012, McNamara et al., 2015, Amorim

et al., 2018]. This feature is useful when the size of the training data is small because these features help generalize word n-gram features [Ke and Ng, 2019]. Therefore, this feature is potentially helpful for AES because essay corpora are often small. To be more specific, we select discourse connectives [Pitler and Nenkova, 2009] (denoted by ***discourse_conn_word***), sentiment words [Bird et al., 2009] (***sentiment_word***) and modals (***modal_word***). We implement this set of features as word-level features, with each word labeled as to whether it belongs to the word category.

Prompt-relevant Features. A good essay should be highly related to the prompt. Thus, this feature represents the relatedness between the essay and the prompt. A variety of similarity measures have often been used to compute this feature, such as word overlap, word topicality, and semantic similarity. In this category, we will use the LDA model [Blei et al., 2003] to compute this feature. Since this is a data-driven method, we assume most essays talk about the same topic, and that they are related to the prompt. Then the LDA model helps us find out if an essay is off topic. With the LDA model, we can know the topic distribution of the essay or a single sentence. Therefore, this feature could be either an essay-level feature (denoted by ***lda_essay***) or a sentence-level feature (***lda_sent***). Each essay or sentence will be represented by its topic distribution. Since we can also know the word-topic distribution from the LDA model, we can also use the distribution as word representation. Therefore, this feature could also be a word-level feature (***lda_word***).

Readability Features. A good essay should be easy to read by a specific group of people, which means the word choice should be neither too difficult nor too easy to read. A good writer should demonstrate vocabulary that matches their school level [Zesch et al., 2015b]. A widely used readability metric is the Flesch Reading Ease Test (FRE) [Kincaid et al., 1975]:

$$FRE = 206.835 - 1.015 * \frac{\#words}{\#sentences} - 84.6 * \frac{\#syllables}{\#words}$$

The score range of FRE is from $-\infty$ to 121.22, the lower the number, the harder to read. Table 14 shows the conversion table from FRE score to grade level. Although the FRE measures the essay level readability, we could also calculate FRE for independent sentences.

Therefore, this feature could be either an essay level (denoted by ***readability_essay***) or sentence level feature (***readability_sent***).

Score	School level
$\geq$ 90.0	5th grade
90-80.0	6th grade
80-70.0	7th grade
70-60.0	8th & 9th grade
60-50.0	10th to 12th grade
50-30.0	College
$\leq$ 30	College graduate

Table 14: FRE conversion table.

Syntactic Features. This feature encodes the syntactic information about the essay, and it demonstrates the writer's style [Zesch et al., 2015b]. This feature is a word-level feature, and we label each word with its part-of-speech tag. We use two tagsets. The first one is the most commonly used Penn Treebank Tagset [Taylor et al., 2003] (denoted by ***pos_word***). However, this tagset is too comprehensive and contains 36 different tags without special symbols. We doubt whether a comprehensive tagset is suitable for the AES task because the size of the training data is usually small. Thus, we also use a simple tagset that only contains "adjective", "noun", "adverb", and "verb" (***simple_pos_word***).

Argumentation Features. Using argumentative structures to score persuasive essays has drawn increasing attention [Persing and Ng, 2015, Ghosh et al., 2016, Nguyen and Litman, 2018]. We label each word with IOB-formatted labels for argument units (premises, claims) with TARGER [Chernodub et al., 2019]. Thus, this sequence tagging feature is a word-level feature (***argument_word***).

Discourse Features. Typically, there are four widely used discourse features: entity grids [Barzilay and Lapata, 2008], rhetorical structure theory trees [Mann and Thompson, 1988], lexical chains [Morris and Hirst, 1991], and discourse function labels [Persing et al., 2010]. In this work, we plan to test the discourse function label because it provides a label

for each sentence, which makes this feature a sentence level feature. Table 15 show the full possible discourse function labels and explanations for sentence. Possible labels are "Prompt", "Transition", "Thesis", "Main Idea", "Elaboration", "Support", "Conclusion", "Rebuttal", "Solution", and "Suggestion". A heuristic algorithm for labeling each sentence was developed [Persing et al., 2010][1]. For each sentence, this algorithm calculates a score for each label and assigns the label with the highest score to the sentence. Therefore, we created 2 forms of this feature. The first is the label for each sentence (denoted by ***discourse_func_sent***), while the second is the score vector (***discourse_func_vec_sent***).

Label	**Sentence Function**
Prompt	restates the prompt given to the author and contains no new material or opinions
Transition	shifts the focus to new topics but contains no meaningful information
Thesis	states the author's position on the topic for which he/she is arguing
Main Idea	asserts reasons and foundational arguments that support the thesis
Elaboration	further explains reasons and ideas but contains no evidence or examples
Support	provides evidence and examples to support the claims made in other statements
Conclusion	summarizes and concludes the entire argument or one of the main ideas
Rebuttal	considers counter-arguments that contrast with the thesis or main ideas
Solution	puts to rest the questions and problems brought up by counter-arguments
Suggestion	proposes solutions the problems brought up by the argument

Table 15: Descriptions of sentences function labels.

Besides, this is our initial exploration of this hybrid model. Therefore we listed widely used hand-crafted features as many as possible, and we only select one or two features from each category. We thought features were meant to be exhaustive over types but illustrative within each type. Thus, features like LDA features or discourse function label features might not be the most optimal feature over all AES tasks. However, we still observe that the LDA model is one of the best features at the essay level.

[1]For the complete list of sentence labeling heuristics, visit http://www.hlt.utdallas.edu/~persingq/ICLE/SentenceRules.txt

5.5 Experimental Setup

We use the Automated Student Assessment Prize (ASAP) corpus and the Response-to-Text Assessment (RTA) corpus to evaluate our hybrid model. We configure experiments to test four hypotheses:

- H1: The hand-crafted features work better when combined on the sentence-level or word-level of the neural model, compared to the essay-level.
- H2: The hybrid model that combines features extracted from the essay at the sentence-level or word-level works better, compared to essay-level features.
- H3: The hybrid model will outperform or at least perform as well as the base neural model when trained and tested on the same prompt.
- H4: The hybrid model will generalize better across different prompts because hand-crafted features generalize better over prompts.

For within-prompt experiments, we use 5-fold cross-validation as in prior work [Dong et al., 2017], to split the data for each prompt into 5 folds. In each fold, 60% of data are used for training, 20% are used for development, and 20% are used for testing. Note that we are using a different deep learning framework to implement the base model compared to that used in [Dong et al., 2017]. The original paper used Keras 1.1.1 and Theano 0.8.2, while we use TensorFlow 2.2.0. Thus, the base model AES results reported in this work have small differences compared to the numbers reported in the original paper.

For cross-prompt experiments, we extend the 4 single direction pairs of essay prompts used in prior work [Phandi et al., 2015] to 5 bi-direction pairs. More specifically, these 5 pairs of essay prompts were picked based on the similarity in their genres, score ranges, and median scores. The essay set pairs are $1 \leftrightarrow 2$, $3 \leftrightarrow 4$, $5 \leftrightarrow 6$, $7 \leftrightarrow 8$, and $M \leftrightarrow S$, where the pair $1 \leftrightarrow 2$ denotes using prompt 1 (or prompt 2) as the source prompt and prompt 2 (or prompt 1) as the target prompt. We use all essays from the source prompt for training. Target prompt data are randomly divided into 5 folds (same as the within-prompt experiment), where one fold is used as test data, and one fold is used as the development set. We do not include data from the target prompt for training in order to test the ability of

prompt adaptation of our model. We use the development set from the target prompt, but only to determine early stopping. All other hyper-parameters are not selected based on the development set. Consequently, our approach is not zero-shot but instead assumes a small amount of data from the target prompt.

Besides, we also include results of four other models reported in prior work [Phandi et al., 2015, Cozma et al., 2018, Liu et al., 2019, Cao et al., 2020] for pairs that were previously studied, denoted by ML-ρ, SKWE, TSLF, and HA, respectively. Note that although these results are numbers from the original papers, the size of the training and testing data are the same as in our experiments. Since we do not obtain model performance for each fold, we cannot perform significance tests between our model and these models.

Since some hand-crafted features used by our model are categorical, we need to change their representations to serve as the input of the neural network. We will test two forms, either one-hot representation or embedding representation.

Since our experiments combine multiple features at a time, we want to perform feature selection to select the best level of combination for each feature. The level of combination is the level that we combine the essay representation and feature representation, either word level, sentence level, or essay level. First, we combine one feature at a time, so that we know the best representation and combination level of each feature. Next, we need to figure out which features to combine. We adopt three strategies. First, we simply combine the best variant of each feature (denoted by FSA, where A stands for all features). Therefore, FSA shows the best representation and combination level of each feature (possible values described below). Second, we select one subset of features that works for all prompts. Specifically, we select features that improve the base model on the development set for at least 9 (out of 10) prompts (denoted by FS1). We also select features that **significantly** improve the base model on the development set for at least 6 prompts (denoted by FS2). The intuition is that we want to combine features that improve the base model on as many prompts as possible, while preserving a reasonable number of features. Third, we select a set of features separately for each prompt. We select a feature as long as using it in the hybrid model improves over the base model when evaluated on the development set for the prompt (denoted by FS3). Table 16 shows the selected features of FSA, FS1, and FS2. For each

feature set, the "Comb" column indicates the best combination level, and "Emb" indicates the best feature representation: "vec" means the feature is a vector, "one" means one-hot representation, and number means the number of dimensions of embedding representation. Table 17 shows the selected features of FS3. Note that all feature selection is made on the development set with an within-prompt experimental setting.

	FSA		**FS1**		**FS2**	
Feature	Comb	Emb	Comb	Emb	Comb	Emb
word_count_essay	essay	vec	NA	NA	essay	vec
word_count_sent	sent	vec	sent	vec	sent	vec
discourse_conn_word	sent	one	sent	one	sent	one
sentiment_word	sent	vec	NA	NA	NA	NA
modal_word	sent	one	NA	NA	NA	NA
lda_essay	essay	vec	essay	vec	NA	NA
lda_sent	essay	vec	NA	NA	NA	NA
lda_word	word	vec	NA	NA	NA	NA
readability_essay	essay	vec	NA	NA	essay	vec
redability_sent	sent	vec	sent	vec	NA	NA
pos_word	essay	one	essay	one	NA	NA
simple_pos_word	sent	one	NA	NA	NA	NA
argument_word	sent	one	sent	one	NA	NA
discourse_func_sent	sent	50	sent	50	essay	50
discourse_func_vec_sent	sent	vec	NA	NA	NA	NA

Table 16: Selected features of FSA, FS1, and FS2.

Following Dong et al., [Dong et al., 2017], the vocabulary size of the data is limited to 4000, all scores are scaled to the range [0, 1], all hyper-parameters for training shown in Table 18, and use Quadratic Weighted Kappa (QWK) for evaluation.

Prompt	$ASAP_1$		$ASAP_2$		$ASAP_3$		$ASAP_4$		$ASAP_5$		$ASAP_6$		$ASAP_7$		$ASAP_8$		MVP		$Space$	
Feature	Comb	Emb	Comb	Emb	Comb	Emb	Comb	Emb	Comb	Emb	Comb	Emb	Comb	Emb	Comb	Emb	Comb	Emb	Comb	Emb
word_count_essay	NA	NA	NA	NA	essay	vec	NA	NA	essay	vec	NA	NA	NA	NA	NA	NA	essay	vec	essay	vec
word_count_sent	sent	vec	sent	vec	NA	NA	sent	vec	NA	NA	sent	vec	sent	vec	NA	NA	NA	NA	NA	NA
discourse_conn_word	sent	50	sent	50	sent	one	sent	one	sent	one	essay	5	sent	50	NA	NA	word	50	sent	one
sentiment_word	NA	NA	sent	vec	sent	vec	sent	vec	word	vec	word	vec	sent	vec	NA	NA	sent	vec	sent	vec
modal_word	NA	NA	sent	50	NA	NA	sent	one	word	one	word	5	sent	one	NA	NA	NA	NA	NA	NA
lda_essay	NA	NA	NA	NA	essay	vec	essay	vec	NA	NA	NA	NA	NA	NA	essay	vec	NA	NA	NA	NA
lda_sent	NA	NA	sent	vec	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
lda_word	NA	NA	NA	NA	NA	NA	NA	NA	sent	vec	sent	vec	word	vec	NA	NA	word	vec	sent	vec
readability_essay	NA	NA	essay	vec	NA	NA	NA	NA	NA	NA	essay	vec	NA	NA	NA	NA	NA	NA	NA	NA
redability_sent	sent	vec	NA	NA	sent	vec	essay	vec	sent	vec	NA	NA	sent	vec	sent	vec	sent	vec	essay	vec
pos_word	essay	one	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	word	50	sent	5	NA	NA
simple_pos_word	NA	NA	NA	NA	word	one	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	sent	0
argument_word	essay	50	sent	50	essay	one	sent	50	word	one	sent	5	sent	one	word	50	sent	5	word	50
discourse_func_sent	sent	50	sent	one	essay	50	NA	NA	sent	one	NA	NA	sent	50	NA	NA	NA	NA	sent	5
discourse_func_vec_sent	NA	NA	NA	NA	NA	NA	essay	vec	NA	NA	essay	vec	NA	NA	NA	NA	sent	vec	NA	NA

Table 17: Selected features of FS3.

5.6 Results and Discussion

Results for H1. The feature list of FSA in Table 16 supports H1. We observe that the hand-crafted features work better when combined on sentence-level, compared to word-level and essay-level. Overall, we have 7 word-level features, 5 sentence-level features, and 3 essay-level features. Since essay-level features can only be combined on essay-level, we only focus on word-level features and sentence-level features. There are 5 (out of 7) word-level features that perform the best when combined on the sentence level. There are 4 (out of 5) sentence-level features that perform the best when combined on the sentence level. However, only 1 word-level feature and sentence-level feature perform the best when combined on the essay level.

One possible reason for the utility of sentence-level combination is that the word-level representation contains too much detailed information, including noise explicitly, while the sentence-level representation abstracts word-level representations and reduces explicit noise. Essay-level representation, which is even more abstracted, may in turn lose too much detailed information of the essay.

Results for H2. Table 19 (columns 3-5) shows the best feature of each linguistic

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	50
	Batch size	16
	Initial learning rate	0.001
	Momentum	0.9

Table 18: Hyper-parameters of training.

Pmt	**Base**	**LE**	**WCS**	**DCW**	**FSA**	**FS1**	**FS2**	**FS3**	**ML-ρ**	**SKWE**	**TSLF**	**HA**
$ASAP_1$	0.830	0.828	0.831	0.833	0.828	0.833	0.832	**0.838**	0.761	0.845	0.852	0.824
$ASAP_2$	0.672	0.675	**0.678**	0.673	0.656	0.675	0.671	0.671	0.606	0.729	0.736	0.699
$ASAP_3$	0.677	0.682	0.679	0.685	0.685	0.688	0.677	**0.690**	0.621	0.684	0.731	0.726
$ASAP_4$	0.807	0.814	0.812*	0.815	0.805	0.810	**0.817**	0.812	0.742	0.829	0.801	0.859
$ASAP_5$	0.806	0.803	0.806	0.793	0.809	0.809	**0.813**	0.808	0.784	0.833	0.823	0.822
$ASAP_6$	0.809	0.804	0.811	**0.813**	0.806	0.811	0.811	0.808	0.775	0.830	0.792	0.828
$ASAP_7$	0.797	0.801	0.805	**0.810**	0.787	0.806	0.801	**0.810**	0.730	0.804	0.762	0.840
$ASAP_8$	0.680	0.676	0.679	0.680	0.588	**0.685**	0.672	**0.685**	0.617	0.729	0.684	0.726
MVP	0.681	0.694	0.696	0.683	0.685	0.678	**0.692**	0.680	NA	NA	NA	NA
Space	0.669	0.670	0.678	0.677*	0.663	0.672	**0.680**	0.670	NA	NA	NA	NA
Avg ASAP	0.760	0.760	0.763*	0.764	0.745	**0.765***	0.762	**0.765***	0.705	0.785	0.773	0.791
Avg RTA	0.675	0.682	**0.687**	0.680	0.674	0.675	0.686*	0.675	NA	NA	NA	NA
Avg Overall	0.743	0.745	**0.747***	**0.747**	0.731	**0.747***	**0.747**	**0.747***	NA	NA	NA	NA

Table 19: The performance (QWK) of best single feature of each level, and of each feature selection set for within-prompt experiments. * indicates that the result is significantly better than the baseline ($p \leq 0.05$). The best results within the base model and proposed model of each row are in bold.

level. Obviously, the *word_count_sent* (WCS) feature and the *discourse_conn_word* (DCW) feature improve the base model more than the *lda_essay* (LE) feature. Besides, on average, all essay-level features improve 6.33 prompts, while sentence-level features and word-level features improve 6.60 and 6.14 prompts, respectively. All results support H2.

Results for H3. The results in Table 19 generally support H3. When using multiple features chosen via feature selection, FS1 and FS3 yield significantly higher performance than the base model on average. Although FS2 does not yield significant improvement on average, it significantly outperforms the base model on the RTA corpus. Even when the hybrid model uses only one feature, *lda_essay* (LE), *word_count_sent* (WCS) or *discourse_conn_word* (DCW) can also outperform the base model on average. In contrast, when using all hand-crafted features (FSA), the hybrid model performs worse than the base model. One possible reason is that size of the ASAP training set is relatively small. If we combine too many features, the hybrid model might become too complicated for the AES task.

	$1 \rightarrow 2$	$2 \rightarrow 1$	$3 \rightarrow 4$	$4 \rightarrow 3$	$5 \rightarrow 6$	$6 \rightarrow 5$	$7 \rightarrow 8$	$8 \rightarrow 7$	$M \rightarrow S$	$S \rightarrow M$
ML-ρ	0.434	-	0.522	-	0.187	-	0.171	-	-	-
SKWE	0.542	-	0.701	-	**0.728**	-	0.522	-	-	-
TSLF	-	-	-	-	-	-	-	-	-	-
HA	0.577	-	0.704	-	0.722	-	0.614	-	-	-
Base	0.502	0.426	0.692	0.630	0.438	0.095	0.552	0.491	0.498	0.491
FS1	0.595*	0.620*	**0.707***	0.666*	0.615*	0.580*	0.608*	**0.529***	0.517	**0.529***
FS2	**0.632***	**0.706***	0.688	**0.680***	0.613*	**0.672***	**0.638***	0.486	0.501	0.515
FS3	0.579*	**0.706***	0.697	**0.680***	0.630*	0.667*	0.621*	0.476	**0.509**	0.513

Table 20: The performance (QWK) of cross-prompt experiments. * indicates that the result is significantly better than the Base model ($p \leq 0.05$). The best results of each column are in bold.

Not surprisingly, FS1 and FS2 also include the *word_count_sent* and *discourse_conn_word* features. As expected, word count on sentence-level improves the model, because word count is highly predictive in isolation. As for the *discourse_conn_word feature*, it may explicitly introduce discourse information into the model, while the base model lacks such information.

Besides, we observe that the SKWE, the TSLF, and the HA models outperform our

model on average on ASAP, but not necessary on each individual prompt. However, we believe that this is because our model selected a relatively weak base model as an initial point. The results still show that our models significantly outperform the base model, which means the basic idea of combining hand-crafted features and the neural network model, still improves the neural network model in the within-prompt experimental setting. In addition, the TSLF model is also a hybrid model. However, it only combined essay-level features, while we investigated sentence-level features and word-level features other than essay-level features.

Results for H4. Since FSA does not perform well in within-prompt experiments, we exclude FSA from cross-prompt experiments. The results in Table 20 support H4. For each prompt pair, we can always find the best result from our proposed model, except prompt pair $5 \rightarrow 6$. Although our model does not outperform SKWE and HA for prompt pair $5 \rightarrow 6$, it still outperforms the base model, which indicates that our approach has a positive contribution to the base model.

Comparing to within-prompt experiments, the hybrid model shows a more considerable improvement over the base model. One possible reason is that the features we combined are general features that may be sharing the same behavior over prompts. For example, if the average lengths of source prompt and target prompt essays are similar, length-based features should still be highly predictive over prompts. A similar reason might hold for the readability feature, as long as the age groups are similar. The discourse-related features encode the discourse structure of an essay, and reflect the organization of an essay. Since the quality of the organization is content independent, thus, the discourse-related features generalize over prompts as well.

5.7 Conclusion

In this section, we presented an investigation of combining hand-crafted features into an attention-based neural network model. Rather than using essay-level hand-crafted features as a side input, we proposed combining sentence-level and word-level features so that

the neural network model could model hand-crafted features. Within-prompt experimental results demonstrated that with feature selection, our model could outperform the base model. A set of cross-prompt experiments also demonstrated that our hybrid model could outperform not only our base neural model, but also other baselines from the literature.

6.0 Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring

6.1 Introduction

While automated essay scoring (AES) can reliably grade essays at scale, automated writing evaluation (AWE) additionally provides formative feedback to guide essay revision. However, a neural AES typically does not provide useful feature representations for supporting AWE. Meanwhile, non-neural AES create feature representations more easily useable by AWE [Roscoe et al., 2014, Foltz and Rosenstein, 2015, Crossley and McNamara, 2016, Woods et al., 2017, Madnani et al., 2018, Zhang et al., 2019]. We believe that neural AES can also provide useful information for creating feature representations, e.g., by exploiting information in the intermediate layers.

In this chapter, we present an investigation of using the interpretable output of the attention layers of the Co-attention AES model with the goal of extracting Topical Components (TCs) needed for the eRevise [Zhang et al., 2019] system. For each source, the TCs consist of a comprehensive list of topics related to evidence which include: 1) important words indicating the set of evidence topics in the source, and 2) phrases representing specific examples for each topic that students need to find and use in their essays. Table 35 and Table 39 are topic words list and specific example phrases list of MVP article, respectively. We evaluate performance using a feature-based AES introduced in Chapter 3 requiring TCs. Results show that performance is comparable whether using automatically or manually constructed TCs for 1) representing essays as rubric-based features, 2) grading essays, 3) generating feedback. This work is illustrated in [Zhang and Litman, 2020].

6.2 Related Work

Three recent AWE systems have used non-neural AES to provide rubric-specific feedback. Woods et al. [Woods et al., 2017] developed an influence estimation process that used a logistic regresion AES to identify sentences needing feedback. Shibani et al. [Shibani et al., 2019] presented a web-based tool that provides formative feedback on rhetorical moves in writing. Zhang et al. [Zhang et al., 2019] used features created for a random forest AES to select feedback messages, although human effort was first needed to create TCs from a source text. We automatically extract TCs using neural AES, thereby eliminating this expert effort.

Others have also proposed methods for pre-processing source information external to an essay. Content importance models for AES predict the parts of a source text that students should include when writing a summary [Klebanov et al., 2014]. Methods for extracting important keywords or keyphrases also exist, both supervised (unlike our approach) [Meng et al., 2017, Mahata et al., 2018, Florescu and Jin, 2018] and unsupervised [Florescu and Caragea, 2017]. Rahimi and Litman [Rahimi and Litman, 2016] developed a TC extraction LDA model [Blei et al., 2003]. While the LDA model considers all words equally, our model takes essay scores into account by using attention to represent word importance. Both the unsupervised keyword and LDA models will serve as baselines in our experiments.

In the computer vision area, attention cropped images have been used for further image classification or object detection [Cao et al., 2015, Yuxin et al., 2018, Ebrahimpour et al., 2019]. In the NLP area, Lei et al. [Lei et al., 2016] proposed to use a generator to find candidate rationale and these are passed through the encoder for prediction. Our work is similar in spirit to this type of work.

6.3 Prior AES and AWE for the RTA

We have proposed two approaches to AES for the RTA: AES_{rubric} in Chapter 3 and AES_{neural} in Chapter 4. To support the needs of AWE (eRevise system[Zhang et al., 2019]), AES_{rubric} used a traditional supervised learning framework where rubric-motivated features

were extracted from every essay before model training The two aspects of TCs (*topic words*, *specific example phrases*) are italicized below to indicate TC usage during feature extraction.:

Number of Pieces of Evidence (NPE). An integer feature based on the list of *topic words* for each topic.

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 example phrases* (semantically) mentioned in the essay per topic.

Word Count (WOC). Number of words.

SPC_Total. Sum of all SPC features values.

SPC_Total_Merged. Number of unique *specific example phrases* from the SPC vector.

After feature-based AES, the eRevise system selected a level of feedback. Each level was associated with two (of four possible) detailed feedback messages on a scale of 1 to 3 (low to high) to guide student revision. The level was determined using an algorithm based on the AES feature analysis, enabling each student's feedback to be targeted to the needs of their particular essay, which extracted by expert provided TCs and are thus targeted to improving the quality of each student's particular essay.

Motivated by improving stand-alone AES performance (i.e., when an interpretable model was not needed for subsequent AWE), [Zhang and Litman, 2018] developed AES_{neural}, a hierarchical neural model with the co-attention mechanism in the sentence level to capture the relationship between the essay and the source. Neither feature engineering nor TC creation were needed before training.

6.4 Attention-Based TC Extraction

In this section we propose a method for extracting TCs based on the AES_{neural} attention level outputs. Since the self-attention and co-attention mechanisms were designed to capture sentence and phrase importance, we hypothesize that the attention scores can help determine if a sentence or phrase has important source-related information.

No.	Sentences	$attn_{sent}$	$attn_{phrase}$
1	People didn't **have the money to buy** the stuff in 2004.	0.00420	0.23372
2	The *hunger crisis has been* ***addressed with fertilizer and seeds***, as well as the *tools needed to maintain the food.*	0.08709	0.62848
3	***The school has no fees*** and *they serve lunch.*	0.10686	0.63369

Table 21: Example attention scores of essay sentences.

To provide intuition, Table 21 shows examples sentences from the student essay in Figure 3. Bolded are phrases with the highest self-attention score within the sentence. Italics are specific example phrases that refer to the manually constructed TCs for the source. $Attn_{sent}$ is the text to essay attention score that measures which essay sentences have the closest meaning to a source sentence. $Attn_{phrase}$ is the self-attention score of the bolded phrase that measures phrase importance. A sentence with a high attention score tends to include at least one specific example phrase, and vice versa. The phrase with the highest attention score tends to include at least one specific example phrase if the sentence has a high attention score.

Based on these observations, we first extract the output of two layers from the neural network: 1) the $attn_{sent}$ of each sentence, and 2) the output of the convolutional layer as the representation of the phrase with the highest $attn_{phrase}$ in each sentence (denoted by cnn_{phrase}). We also extract the plain text of the phrase with the highest $attn_{phrase}$ in each sentence (denoted by $text_{phrase}$). Then, our TC_{attn} method uses the extracted information in 3 main steps: 1) filtering out $text_{phrase}$ from sentences with low $attn_{sent}$, 2) clustering all remaining $text_{phrase}$ based on cnn_{phrase}, and 3) generating TCs from clusters.

The first filtering step keeps all $text_{phrase}$ where the original sentences have $attn_{sent}$ higher than a threshold. The intuition is that lower $attn_{sent}$ indicates less source-related information.

The second step clusters these $text_{phrase}$ based on their corresponding representations cnn_{phrase}. We use k-medoids to cluster $text_{phrase}$ into M clusters, where M is the number of topics in the source text. Then, for $text_{phrase}$ in each topic cluster, we use k-medoids to cluster them into N clusters, where N is the number of the specific example phrases we want

to extract from each topic. The outputs of this step are $M * N$ clusters.

The third step uses the topic and example clustering to extract TCs. As noted earlier, TCs include two parts: topic words, and specific example phrases. Since our method is data-driven and students introduce their vocabulary into the corpus, essay text is noisy. To make the TC output cleaner, we filter out words that are not in the source text. To obtain topic words, we combine all $text_{phrase}$ from each topic cluster to calculate the word frequency per topic. To make topics unique, we assign each word to the topic cluster in which it has the highest normalized word frequency. We then include the top K_{topic} words based on their frequency in each topic cluster. To obtain example phrases, we combine all $text_{phrase}$ from each example cluster to calculate the word frequency per example, then include the top $K_{example}$ words based on their frequency in each example cluster.

6.5 Experimental Setup

Figure 14 shows an overview of four TC extraction methods to be evaluated. TC_{manual} (upper bound) uses a human expert to extract TCs from a source text. TC_{attn} is our proposed method and automatically extracts TCs using *both* a source text and student essays. TC_{lda} [Rahimi and Litman, 2016] (baseline) builds on LDA to extract TCs from student essays only, while TC_{pr} (baseline) builds on PositionRank [Florescu and Caragea, 2017] to instead extract TCs from only the source text.

Since PositionRank is not designed for TC extraction, we needed to further process its output to create TC_{pr}. To extract topic words, we extract all keywords from the output. Next, we map each word to a higher dimension with word embedding. Lastly, we cluster all keywords using k-medoids into PR_{topic} topics. To extract example phrases, we put them into only one topic and remove all redundant example phrases if they are subsets of other example phrases.

We configure experiments to test three hypotheses:

H1: The AES_{rubric} model for scoring Evidence will perform comparably when extracting features using either TC_{attn} or TC_{manual}, and will perform worse when using TC_{lda}

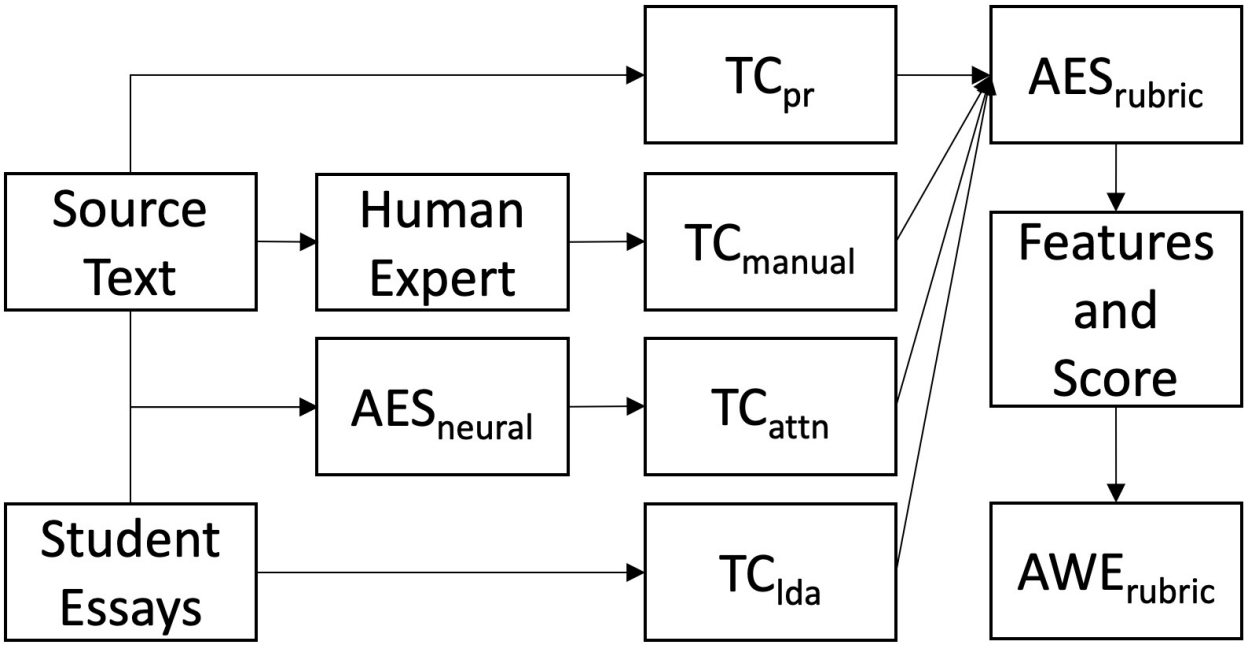


Figure 14: An overview of four TC extraction systems.

or TC_{pr}.

H2: The correlation between the human Evidence score and the feature values (NPE and sum of SPC features) will be comparable when extracted using TC_{attn} and TC_{manual}, and will be stronger than when using TC_{lda} and TC_{pr}.

H3: The eRevise model will assign similar feedback to student essays when extracting features using either TC_{attn} or TC_{manual}, and will assign less similar feedback to student essays when using TC_{lda} or TC_{pr}.

We design our experiment to test hypotheses both extrinsically and intrinsically. Extrinsically, the experiments for H1 and H3 test the impact of using our proposed TC extraction method on the downstream AES_{rubric} task and eRevise AWE system, while the H2 experiment examines the impact on the essay representation itself.

Following Chapter 3, we stratify essay corpora: 40% for training word embeddings and extracting TCs, 20% for selecting the best embedding and parameters, and 40% for testing. We use the hyper-parameters from Chapter 4 for neural training. Table 22 shows all other

parameters selected using the development set.

6.6 Results and Discussion

Results for H1. H1 is supported by the results in Table 23, which compares the Quadratic Weighted Kappa (QWK) between human and AES_{rubric} Evidence scores (values 1-4) when AES_{rubric} uses TC_{manual} versus each of the automatic methods. TC_{attn} always yields better performance, and even significantly better than TC_{manual}.

Results for H2. The results in Table 24 support H2. TC_{attn} outperforms the two automated baselines, and for NPE even yields stronger correlations than the manual TC method.

Results for H3. The results in Table 25 partially support H3. TC_{attn} outperforms the two automated baselines on RTA_{Space}, while performs worse than TC_{lda} on RTA_{MVP}. One possible reason why all models perform well on RTA_{Space} is that essay scores are required for selecting feedback. Since all models perform not bad on essay scoring, it may affect the feedback selection process. However, feedback selection for RTA_{MVP} only relies on absolute feature value. Since automated TCs have different numbers of topics and examples, we have to scale each feature's number to the range that the manual list has. Besides, feedback selection for RTA_{MVP} requires manually defined important topics, which all automated methods cannot do. Therefore, automated methods perform worse on RTA_{MVP} than RTA_{Space}.

Qualitative Analysis. The manually-created topic words for RTA_{MVP} represent 4 topics, which are "hospital", "malaria", "farming" and "school"[1]. Although Table 22 shows that the automated list has more topics for topic words and might have broken one topic into separate topics, a good automated list should have more topics related to the 4 topics above. We manually assign a topic for each of the topic words from the different automated methods. TC_{lda} has 4 related topics out of 9 (44.44%), TC_{pr} has 6 related topics out of 19 (31.58%), and TC_{attn} has 10 related topics out of 16 (62.50%). Obviously, TC_{attn} preserves more related topics than our baselines.

[1]All Topic Words generated by different models can be found in the Appendix C and Appendix D

Prompt	Component	Parameter	TC_{lda}	TC_{pr}	TC_{attn}
RTA_{MVP}	Topic Words	Number of Topics	9	19	16
		Number of Words	30	20	25
	Example Phrases	Number of Topics	20	1	18
		Number of Phrases	15	20	15
RTA_{Space}	Topic Words	Number of Topics	15	20	10
		Number of Words	10	10	20
	Example Phrases	Number of Topics	10	1	9
		Number of Phrases	20	50	20

Table 22: Parameters for different models.

Moving to the second aspect of TCs (specific example phrases), Table 26 shows the first 10 specific example phrases for a manually-created category that introduces the changes made by the MVP project[2]. This category is a mixture of different topics because it talks about the "hospital", "malaria", "school", and "farming" at the same time. TC_{attn} has overlap with TC_{manual} on different topics. However, TC_{lda} mainly talks about "hospital", because the nature of the LDA model doesn't allow mixing specific example phrases about different topics in one category. Unfortunately, TC_{pr} does not include any overlapped specific phrase in the first 10 items; they all refer to some general example phrases from the beginning of the source article. Although there are some related specific example phrases in the full list, they are mainly about school. This is because the PositionRank algorithm tends to assign higher scores to words that appear early in the text.

[2]All Specific Example Phrases generated by different models can be found in the Appendix C and Appendix D

Prompt	TC_{manual} (1)	TC_{lda} (2)	TC_{pr} (3)	TC_{attn} (4)
MVP	0.643 (2,3)	0.614 (3)	0.525	**0.648 (1,2,3)**
Space	0.609 (3)	0.615 (3)	0.559	**0.622 (1,3)**

Table 23: The performance (QWK) of AES_{rubric} using different TC 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.

6.7 Conclusion

This work proposes TC_{attn}, a method for using the attention scores in a neural AES model taking essay scores into account to capture the importance of essays, sentences, and words when to extract the Topical Components of a source text automatically. Evaluations show the potential of TC_{attn} for eliminating expert effort without degrading AES_{rubric} performance or the feature representations themselves. TC_{attn} outperforms baselines and generates comparable or even better results than a manual approach.

Although TC_{attn} outperforms all baselines and requires no human effort on TC extraction, annotation of essay evidence scores is still needed. In Chapter 7, we investigate to train the AES_{neural} using the gold standard that can be extracted automatically.

Prompt	Feature	TC_{manual}	TC_{lda}	TC_{pr}	TC_{attn}
MVP	NPE	0.542	0.482	0.587	**0.639**
	SPC (sum)	0.689	0.585	0.365	**0.679**
Space	NPE	0.484	0.513	0.494	**0.625**
	SPC (sum)	0.601	0.574	0.533	**0.598**

Table 24: Pearson's r comparing feature values computed using each TC extraction method with human (gold-standard) Evidence essay scores. All correlation values are significant ($p \leq 0.05$). The best results between automated methods in each row are in bold.

Prompt	TC_{lda} (2)	TC_{pr} (3)	TC_{attn} (4)
MVP	**0.332**	0.008	**0.170**
Space	0.601	0.614	**0.644**

Table 25: The QWK of feedback level selection comparing each automated TCs to TC_{manual}. The best results between automated methods in each row are in bold.

TC_{manual}	TC_{lda}	TC_{pr}	TC_{attn}
progress just four years	running water electricity	brighter future hannah	electricity running water irrigation set
medicine most common diseases	water connected hospital generator electricity	millennium villages project	poor showed treatment school supplies
water connected hospital	patients afford	unpaved dirt road	farmers could crops afford bed
hospital generator electricity	rooms packed patients probably	bar sauri primary school	electricity hospital
bed nets used every sleeping site	share beds	future hannah	better fertilizer medicine enough also
hunger crisis addressed fertilizer seeds	recieve treatment	sauri primary school	rooms packed patients
tools needed maintain food supply	doctor clinical officer running hospital	villages project	food fertilizer crops get supply
no school fees	doctors clinical	millennium development goals	five net costs 5
school attendance rate way up	water fertilizer knowledge	village leaders	nets net bed free
kids go school now	receive treatment	dirt road	running water supplies schools almost
...	...	...	...

Table 26: Specific example phrases for the RTA_{MVP} progress topic.

7.0 Essay Quality Signals as Weak Supervision for Automated Topical Component Extraction

7.1 Introduction

In the last chapter, we presented a method for automatically using the attention scores in the co-attention neural network model to extract the Topical Components (TCs) of a source text in order to eliminate human effort. However, to eliminate one human effort, this system requires another human effort, which is the human grading of essays for training the co-attention neural network training. Unfortunately, collecting human grading for a corpus of essays is too expensive. We believe that grading student essays costs more than creating TCs.

In this chapter, 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 co-attention neural network AES model. 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.

7.2 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. [Zesch et al., 2015a] presented a semi-supervised method to reduce the size of the required human-labeled corpus. Meanwhile, Ramachandran et al. [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 grading 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. We 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 human 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.

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

7.3 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 27 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 28 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, we developed a method for TC extraction using AES_{neural} in Chapter 6. The algorithm was based on the observations that the co-attention layer on the sentence level assigned higher attention scores to important

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

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 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.

7.4 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.

7.4.1 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

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

Table 29: Pearson's r comparing different essay quality signals with evidence score.

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 (Table 2), 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 29 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.

7.4.2 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).

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 rubirc 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 29 shows that the correlations between topic distribution similarity and evidence scores are 0.359 and 0.253 for RTA_{MVP} and RTA_{Space}, respectively.

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 30: Hyper-parameters for neural training.

7.5 Experimental Setup

Figure 15 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 15:

H1: While weakly supervised training might not yield state-of-the-art, AES_{neural} per-

Prompt	Component	Parameter	$TCes$	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 31: Selected parameters for different models.

formance 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.

H2: weakly supervised training can nonetheless yield versions of AES_{neural} that are still useful for automated TC extraction.

$\mathbf{AES_{neural}}$ 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 Chapter 4, we use 5-fold cross-validation in this experiment. All hyperparameters for the AES_{neural} training are shown in Table 30.

Extracted TCs (H2). We configure experiments to evaluate the four TC extraction methods in Figure 15 both extrinsically and intrinsically. We thus break H2 into two subhypotheses: 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

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 32: 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.

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 Chapter 4: 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 Chapter 4 for the co-attention neural network training as shown in Table 30. Table 31 show all other parameters selected using the development sets for all models.

7.6 Results and Discussion

Results for H1. Table 32, 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. Unsur-

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 33: 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.

prisingly, the models trained on human scores significantly outperform our proposed weaker essay quality signals on both prompts. 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 29). 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.

Results for H2a. Table 33, 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} 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

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 34: 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}.

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.

Results for H2b. H2b is partially supported by the results in Table 34. For NPE feature, TC_{wc} always yields better performance than TC_{manual}. TC_{tds} yield 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 com-

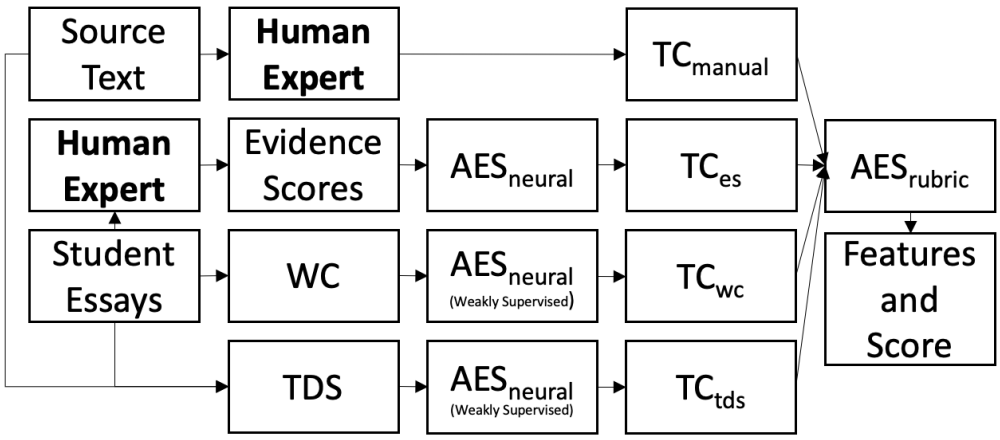


Figure 15: An overview of four TC extraction systems.

paring to TC_{manual}. Although correlations for SPC feature are worse than TC_{es}, considering word count is the most intuitive essay quality 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 33). 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.

7.7 Conclusion

This work 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.

8.0 Summary

In this thesis, we proposed three models that assess student essays for the evidence dimension of the RTA corpus and the holistic score of the ASAP corpus. Besides, we proposed a TC extraction model that makes the eRevise system - an AWE system for the RTA writing tasks fully automated.

The prior AES model for RTA extracts features only by considering the lexical form of each word. Therefore, the model could not match words with spelling mistakes or the use of different vocabularies. The first model we proposed in Chapter 3 uses the word embedding model for feature extraction. Word embedding is used to evaluate word similarity, with two words considered as similar after thresholding, thus enabling both lexical and semantic matching. The experiments show that our proposed model outperforms the existing model. The new feature extraction method addressed the existing model's inability to find topic words or specific examples mentioned in the essay if the student makes spelling mistakes or uses different vocabulary.

The second proposed model introduced the co-attention mechanism into the neural network model (Chapter 4). As the neural network models show their stronger ability for text modeling, more and more neural network models provide state-of-the-art results. The co-attention model takes information from the source article into account. Therefore, this model is optimized for assessing source-based writing tasks. The experiments show that our proposed model outperforms the state-of-the-art models for both RTA tasks and source-based ASAP tasks. We also showed that the co-attention model could capture the relation between student essays and the source article. Important phrases and sentences in the student essay earn higher attention scores.

The third model combines hand-crafted features into an attention-based neural network model (Chapter 5). The two proposed models above only focus on a source-based writing task, while this model could be used for a broader scope. Besides, rather than using essay-level hand-crafted features as a side input, we proposed combining sentence-level and word-level features so that the neural network model could model hand-crafted features.

Experimental results show that our model improves the baseline, especially in cross-prompt experiments.

Although neural network models exhibit more reliable performance than the feature-based model, hand-crafted features are still essential for the AWE system like eRevise. However, the eRevise still need hand-crafted Topical Components (TCs) in order to select feedback. Therefore, we proposed a model in Chapter 6 to address this problem by using the attention output of the Co-Attention neural network model to extract Topical Components. Experiments show that the proposed model outperforms the existing TC extraction model and general topic words or example phrases.

In Chapter 7, we 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 co-attention neural network model. However, they still help generate TCs, which is required by the feature-based model. We observe that even a simple signal like word count does not hurt the state-of-the-art baseline. Since there is no need for human effort, we believe that our work brings AES technology closer to being useful in real classroom scenarios.

The first three proposed models are AES models that assess student essays automatically and more accurately than our baselines. They show the potential to be deployed in the real classroom scenario to reduce human efforts. Besides, we introduced multiple NLP technologies into the AES research area, such as the co-attention mechanism and different ways to use word embedding and hand-crafted features in order to improve the AES model. However, these models still have limitations.

The feature-based AES model can only be used for assessing the RTA task, because the interpretable features are specifically designed for this task and are hard to be generalized to other tasks. The co-attention model only works for source-based writing tasks due to the design of its architecture. These two models mainly focus on grading the evidence dimension score of RTA corpus, because these two model designs heavily rely on the relation between the essay and the source article. According to our observation, although the co-

attention model improves the performance of grading holistic scores on the ASAP corpus, the improvement is not as significant as the improvement on the RTA corpus. The hybrid model addressed some of the limitations above, but it only shows marginal improvement over the baseline model. Although it shows better performance in the cross-prompt experiments, it still cannot dominate all baselines. However, to date, we have only explored 15 categories of features. There are definitely more hand-crafted features that can be combined, and that should be explored in future work. Also, the proposed hybrid model combines hand-crafted features using a relatively simple base model. Another possible future direction is to combine hand-crafted features using neural models with a more complex structure design.

In general, the automated essay scoring systems provide reliable scores for essays. However, the system might also be tricked easily. For example, the systems were not designed to detect plagiarism. If an essay is copied from the source article, we guess at this point is that the model would assign a high evidence score to the essay. Such problem is more eminent in the feature-based model, because the essay covers all topics and examples mentioned by the source article, and the feature values will be high enough to make the essay to receive a prominent evidence score. This leads to an interesting direction of future investigation, which is to detect plagiarism in student essays. Besides, one of the criteria of the evidence dimension rubric is the elaboration of evidence. Based on this criterion, a good summary of the source article should not receive a high evidence score because the student needs to elaborate upon evidence. To address this criterion, the feature-based model uses the concentration feature to determine if a student mentions one of the topics in at least three sentences. However, we think a good summary may trick this feature because it merely summarizes the source article. Therefore, another feature that measures the specificity of the part of the essay that without any specific example should address this problem. The co-attention model and the hybrid model may implicitly address the problem, but more studies should be conducted. Our belief that the neural network model is harder to be tricked is not fully reinforced by the model's robustness. Instead, given the fact that the neural network model is basically a black box, and we do not know how the model works internally, we are in a state of uncertainty while still believing the model. This raises another problem, where the teacher cannot know why an essay is assigned a specific score by the model, and no further

feedback can be given. Therefore, more research needs to be done on how we can understand the behavior of the neural network model, as well as when and how to interfere with human effort. This is not sound intuitive because the target of this research is to exempt human effort. However, we believe this is still a necessary step, which is to make teachers understand the automated system better. Otherwise, the automated scores are not trustworthy because the system can be tricked, and such system cannot be deployed in real classrooms.

Jump into the last two proposed models. In a narrow sense, these models focus on reducing or eliminating human effort for the feature-based model by extracting Topical Components automatically. Broadly speaking, we explored a way of using the intermediate output of the neural network model other than its final predictions to extract keywords and key phrases from an article, as well as essays related to the article. We showed that the intermediate output of the neural network model, which was previously considered not interpretable, can be used directly to generate meaningful output. Besides, the eRevise AWE system that cooperates with the feature-based AES model only works for two specific prompts of the RTA currently because of the lack of manually extracted Topical Components. Meanwhile, the TCs extraction models show their potential to help the classroom deploy the eRevise AWE system or the feature-based AES model with more customized prompts supported.

However, limitations still exist in these two systems. For instance, the feature-based AES model could achieve comparable results when using the proposed methods. However, it still does not outperform neural network models. Besides, although the proposed models outperform the manual TCs in some cases, we do not have an understanding of these results. Our best guess is that human experts add topic words or examples, which are not good essay score indicators. For instance, some "important" specific examples, which are mentioned by almost all student essays, the grader will consider this example as not contributing to the essay score very much. However, we believe that whether we need to put this kind of examples on the list is debatable because these important examples still contribute to essay quality. Besides, a comprehensive list is also necessary for providing feedback. Therefore, one possible future research direction is adding topic words or specific examples that do not contribute to the essay score while human experts believe they are important. On the

other side, these two systems assume human experts are upper bound of system performance. Therefore, these two systems are only relative systems in separating individuals but are short of an external criterion to standard. It will be hard to evaluate the system without a clear criterion because human experts are sometimes subjective, and the "gold standard" we are operating against is also subjective. The problem reveals itself when we evaluate the performance of automated methods, especially the performance of the feedback selection of the eRevise AWE system. We can only compare the relative performance between automated methods and the manual TCs. However, the relative performance introduces bias into the system. This problem raises the same concerns: whether we need human effort to interfere with the automated system; When, where, and why to put human effort into the process, if the answer is yes. We would like to consider a monitoring mechanism during the automated process in the future. Therefore, we can have a better understanding of the system, as well as its final output. Besides, a more evident criterion is also required, and they can be better represented by mathematical fashion for quantitative evaluation. Step back to the system performance of the feedback selection of the eRevise AWE system. Unfortunately, the automated method does not perform well, especially on the RTA_{MVP} corpus. We think this is because the feedback selection of the RTA_{Space} corpus relies on predicted essay scores while the algorithm of the RTA_{MVP} corpus does not. Since the automated methods perform well on the AES task, they also perform well on feedback selection. However, a deeper study is necessary to understand the reason. Otherwise, this becomes a barrier between the automated system and the real classroom application. One possible method is removing similar or duplicated items in TCs. Since we use a clustering algorithm for the final TCs extraction process, this unsupervised algorithm introduces similar or duplicated specific examples into the list. This affects extracted feature values. Furthermore, this also affects selected feedback because feedback selection algorithms rely on extracted NPE and SPC values. Also, we only investigate two different essay quality signals for the weakly supervised method. Therefore, one interesting direction for future investigation is exploring more possible quality signals. Besides, the specific examples are generated from clustering results, so words in a specific example are not in readable orders. This leads to another interesting future investigation: make all examples in the specific examples list more human-understandable, even though it

does not affect the system performance due to the nature of the feature-based AES model.

To sum up, this research is only a small step toward applying automated systems to the real classroom scenario. Our systems show positive improvement in reducing human effort. However, more questions remain open. One important factor is the necessity of human effort. The automated systems are designed to eliminate human effort, though we think the human effort is still essential in this topic because of insufficient understanding of automated systems' internal processes. This could be a potential hazard when we deploy such automated systems into the educational area. Therefore, we still believe that automated systems positively contribute to the classroom, while human effort is also required for monitoring and interfering with the system when necessary. The time and reason for interfering are matters that are also worth discussing.

Appendix A Source Articles of $ASAP_3$ to $ASAP_6$

A.1 Source Article of $ASAP_3$

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 fine.

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.

Kurmaskie.

A.2 Source Article of $ASAP_4$

Winter Hibiscus by Minfong Ho

Saeng, a teenage girl, and her family have moved to the United States from Vietnam. As Saeng walks home after failing her driver's test, she sees a familiar plant. Later, she goes to a florist shop to see if the plant can be purchased.

It was like walking into another world. A hot, moist world exploding with greenery. Huge flat leaves, delicate wisps of tendrils, ferns and fronds and vines of all shades and shapes grew in seemingly random profusion.

"Over there, in the corner, the hibiscus. Is that what you mean?" The florist pointed at a leafy potted plant by the corner.

There, in a shaft of the wan afternoon sunlight, was a single blood-red blossom, its five petals splayed back to reveal a long stamen tipped with yellow pollen. Saeng felt a shock of recognition so intense, it was almost visceral.1

"Saebba," Saeng whispered.

A saebba hedge, tall and lush, had surrounded their garden, its lush green leaves dotted with vermilion flowers. And sometimes after a monsoon rain, a blossom or two would have blown into the well, so that when she drew the well water, she would find a red blossom floating in the bucket.

Slowly, Saeng walked down the narrow aisle toward the hibiscus. Orchids, lanna bushes, oleanders, elephant ear begonias, and bougainvillea vines surrounded her. Plants that she had not even realized she had known but had forgotten drew her back into her childhood world.

When she got to the hibiscus, she reached out and touched a petal gently. It felt smooth and cool, with a hint of velvet toward the center—just as she had known it would feel.

And beside it was yet another old friend, a small shrub with waxy leaves and dainty flowers with purplish petals and white centers. "Madagascar periwinkle," its tag announced.

How strange to see it in a pot, Saeng thought. Back home it just grew wild, jutting out from the cracks in brick walls or between tiled roofs.

And that rich, sweet scent—that was familiar, too. Saeng scanned the greenery around her and found a tall, gangly plant with exquisite little white blossoms on it. "Dok Malik," she said, savoring the feel of the word on her tongue, even as she silently noted the English name on its tag, "jasmine."

One of the blossoms had fallen off, and carefully Saeng picked it up and smelled it. She closed her eyes and breathed in, deeply. The familiar fragrance filled her lungs, and Saeng could almost feel the light strands of her grandmother's long gray hair, freshly washed, as she combed it out with the fine-toothed buffalo-horn comb. And when the sun had dried it, Saeng would help the gnarled old fingers knot the hair into a bun, then slip a dok Malik bud into it.

Saeng looked at the white bud in her hand now, small and fragile. Gently, she closed her palm around it and held it tight. That, at least, she could hold on to. But where was the fine-toothed comb? The hibiscus hedge? The well? Her gentle grandmother?

A wave of loss so deep and strong that it stung Saeng's eyes now swept over her. A blink, a channel switch, a boat ride into the night, and it was all gone. Irretrievably, irrevocably gone.

And in the warm moist shelter of the greenhouse, Saeng broke down and wept.

It was already dusk when Saeng reached home. The wind was blowing harder, tearing off the last remnants of green in the chicory weeds that were growing out of the cracks in the sidewalk. As if oblivious to the cold, her mother was still out in the vegetable garden, digging up the last of the onions with a rusty trowel. She did not see Saeng until the girl had quietly knelt down next to her.

Her smile of welcome warmed Saeng. "Ghup ma laio le? You're back?" she said cheerfully. "Goodness, it's past five. What took you so long? How did it go? Did you—?" Then she noticed the potted plant that Saeng was holding, its leaves quivering in the wind.

Mrs. Panouvong uttered a small cry of surprise and delight. "Dok faeng-noi!" she said. "Where did you get it?"

"I bought it," Saeng answered, dreading her mother's next question.

"How much?"

For answer Saeng handed her mother some coins.

"That's all?" Mrs. Panouvong said, appalled, "Oh, but I forgot! You and the Lambert boy ate Bee-Maags . . . ."

"No, we didn't, Mother," Saeng said.

"Then what else—?"

"Nothing else. I paid over nineteen dollars for it."

"You what?" Her mother stared at her incredulously. "But how could you? All the seeds for this vegetable garden didn't cost that much! You know how much we—" She paused, as she noticed the tearstains on her daughter's cheeks and her puffy eyes.

"What happened?" she asked, more gently.

"I—I failed the test," Saeng said.

For a long moment Mrs. Panouvong said nothing. Saeng did not dare look her mother in the eye. Instead, she stared at the hibiscus plant and nervously tore off a leaf, shredding it to bits.

Her mother reached out and brushed the fragments of green off Saeng's hands. "It's a beautiful plant, this dok faeng-noi," she finally said. "I'm glad you got it."

"It's—it's not a real one," Saeng mumbled.

"I mean, not like the kind we had at—at—" She found that she was still too shaky to say the words at home, lest she burst into tears again. "Not like the kind we had before," she said.

"I know," her mother said quietly. "I've seen this kind blooming along the lake. Its flowers aren't as pretty, but it's strong enough to make it through the cold months here, this winter hibiscus. That's what matters."

She tipped the pot and deftly eased the ball of soil out, balancing the rest of the plant in her other hand. "Look how root-bound it is, poor thing," she said. "Let's plant it, right now."

She went over to the corner of the vegetable patch and started to dig a hole in the ground. The soil was cold and hard, and she had trouble thrusting the shovel into it. Wisps of her gray hair trailed out in the breeze, and her slight frown deepened the wrinkles around her

eyes. There was a frail, wiry beauty to her that touched Saeng deeply.

"Here, let me help, Mother," she offered, getting up and taking the shovel away from her.

Mrs. Panouvong made no resistance. "I'll bring in the hot peppers and bitter melons, then, and start dinner. How would you like an omelet with slices of the bitter melon?"

"I'd love it," Saeng said.

Left alone in the garden, Saeng dug out a hole and carefully lowered the "winter hibiscus" into it. She could hear the sounds of cooking from the kitchen now, the beating of eggs against a bowl, the sizzle of hot oil in the pan. The pungent smell of bitter melon wafted out, and Saeng's mouth watered. It was a cultivated taste, she had discovered—none of her classmates or friends, not even Mrs. Lambert, liked it—this sharp, bitter melon that left a golden aftertaste on the tongue. But she had grown up eating it and, she admitted to herself, much preferred it to a Big Mac.

The "winter hibiscus" was in the ground now, and Saeng tamped down the soil around it. Overhead, a flock of Canada geese flew by, their faint honks clear and—yes—familiar to Saeng now. Almost reluctantly, she realized that many of the things that she had thought of as strange before had become, through the quiet repetition of season upon season, almost familiar to her now. Like the geese. She lifted her head and watched as their distinctive V was etched against the evening sky, slowly fading into the distance.

When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again.

A.3 Source Article of $ASAP_5$

Narciso Rodriguez

from Home: The Blueprints of Our Lives

My parents, originally from Cuba, arrived in the United States in 1956. After liv-

ing for a year in a furnished one-room apartment, twenty-one-year-old Rawedia Maria and twenty-seven-year-old Narciso Rodriguez, Sr., could afford to move into a modest, three-room apartment I would soon call home.

In 1961, I was born into this simple house, situated in a two-family, blond-brick building in the Ironbound section of Newark, New Jersey. Within its walls, my young parents created our traditional Cuban home, the very heart of which was the kitchen. My parents both shared cooking duties and unwittingly passed on to me their rich culinary skills and a love of cooking that is still with me today (and for which I am eternally grateful). Passionate Cuban music (which I adore to this day) filled the air, mixing with the aromas of the kitchen. Here, the innocence of childhood, the congregation of family and friends, and endless celebrations that encompassed both, formed the backdrop to life in our warm home.

Growing up in this environment instilled in me a great sense that "family" had nothing to do with being a blood relative. Quite the contrary, our neighborhood was made up of mostly Spanish, Cuban, and Italian immigrants at a time when overt racism was the norm and segregation prevailed in the United States. In our neighborhood, despite customs elsewhere, all of these cultures came together in great solidarity and friendship. It was a close-knit community of honest, hardworking immigrants who extended a hand to people who, while not necessarily their own kind, were clearly in need.

Our landlord and his daughter, Alegria (my babysitter and first friend), lived above us, and Alegria graced our kitchen table for meals more often than not. Also at the table were Sergio and Edelmira, my surrogate grandparents who lived in the basement apartment. (I would not know my "real" grandparents, Narciso the Elder and Consuelo, until 1970 when they were allowed to leave Cuba.) My aunts Bertha and Juanita and my cousins Arnold, Maria, and Rosemary also all lived nearby and regularly joined us at our table. Countless extended family members came and went — and there was often someone staying with us temporarily until they were able to get back on their feet. My parents always kept their arms and their door open to the many people we considered family, knowing that they would do the same for us.

My mother and father had come to this country with such courage, without any knowledge of the language or the culture. They came selflessly, as many immigrants do, to give

their children a better life, even though it meant leaving behind their families, friends, and careers in the country they loved. They struggled both personally and financially, braving the harsh northern winters while yearning for their native tropics and facing cultural hardships. The barriers to work were strong and high, and my parents both had to accept that they might not be able to find the kind of jobs they deserved. In Cuba, Narciso, Sr., had worked in a laboratory and Rawedia Maria had studied chemical engineering. In the United States, they had to start their lives over entirely, taking whatever work they could find. The faith that this struggle would lead them and their children to better times drove them to endure these hard times.

I will always be grateful to my parents for their love and sacrifice. I've often told them that what they did was a much more courageous thing than I could have ever done. I've often told them of my admiration for their strength and perseverance, and I've thanked them repeatedly. But, in reality, there is no way to express my gratitude for the spirit of generosity impressed upon me at such an early age and the demonstration of how important family and friends are. These are two lessons that my parents did not just tell me. They showed me with their lives, and these teachings have been the basis of my life.

It was in this simple house that my parents welcomed other refugees to celebrate their arrival to this country and where I celebrated my first birthdays. It was in the warmth of the kitchen in this humble house where a Cuban feast (albeit a frugal Cuban feast) always filled the air with not just scent and music but life and love. It was here where I learned the real definition of "family." And for this, I will never forget that house or its gracious neighborhood or the many things I learned there about how to love. I will never forget how my parents turned this simple house into a home.

— Narciso Rodriguez, Fashion designer

Hometown: Newark, New Jersey

A.4 Source Article of $ASAP_6$

The Mooring Mast

by Marcia Amidon Lüsted

When the Empire State Building was conceived, it was planned as the world's tallest building, taller even than the new Chrysler Building that was being constructed at Forty-second Street and Lexington Avenue in New York. At seventy-seven stories, it was the tallest building before the Empire State began construction, and Al Smith was determined to outstrip it in height.

The architect building the Chrysler Building, however, had a trick up his sleeve. He secretly constructed a 185-foot spire inside the building, and then shocked the public and the media by hoisting it up to the top of the Chrysler Building, bringing it to a height of 1,046 feet, 46 feet taller than the originally announced height of the Empire State Building.

Al Smith realized that he was close to losing the title of world's tallest building, and on December 11, 1929, he announced that the Empire State would now reach the height of 1,250 feet. He would add a top or a hat to the building that would be even more distinctive than any other building in the city. John Tauranac describes the plan:

[The top of the Empire State Building] would be more than ornamental, more than a spire or dome or a pyramid put there to add a desired few feet to the height of the building or to mask something as mundane as a water tank. Their top, they said, would serve a higher calling. The Empire State Building would be equipped for an age of transportation that was then only the dream of aviation pioneers.

This dream of the aviation pioneers was travel by dirigible, or zeppelin, and the Empire State Building was going to have a mooring mast at its top for docking these new airships, which would accommodate passengers on already existing transatlantic routes and new routes that were yet to come.

The Age of Dirigibles

By the 1920s, dirigibles were being hailed as the transportation of the future. Also known today as blimps, dirigibles were actually enormous steel-framed balloons, with envelopes of cotton fabric filled with hydrogen and helium to make them lighter than air. Unlike a balloon,

a dirigible could be maneuvered by the use of propellers and rudders, and passengers could ride in the gondola, or enclosed compartment, under the balloon.

Dirigibles had a top speed of eighty miles per hour, and they could cruise at seventy miles per hour for thousands of miles without needing refueling. Some were as long as one thousand feet, the same length as four blocks in New York City. The one obstacle to their expanded use in New York City was the lack of a suitable landing area. Al Smith saw an opportunity for his Empire State Building: A mooring mast added to the top of the building would allow dirigibles to anchor there for several hours for refueling or service, and to let passengers off and on. Dirigibles were docked by means of an electric winch, which hauled in a line from the front of the ship and then tied it to a mast. The body of the dirigible could swing in the breeze, and yet passengers could safely get on and off the dirigible by walking down a gangplank to an open observation platform.

The architects and engineers of the Empire State Building consulted with experts, taking tours of the equipment and mooring operations at the U.S. Naval Air Station in Lakehurst, New Jersey. The navy was the leader in the research and development of dirigibles in the United States. The navy even offered its dirigible, the Los Angeles, to be used in testing the mast. The architects also met with the president of a recently formed airship transport company that planned to offer dirigible service across the Pacific Ocean.

When asked about the mooring mast, Al Smith commented:

[It's] on the level, all right. No kidding. We're working on the thing now. One set of engineers here in New York is trying to dope out a practical, workable arrangement and the Government people in Washington are figuring on some safe way of mooring airships to this mast.

Designing the Mast

The architects could not simply drop a mooring mast on top of the Empire State Building's flat roof. A thousand-foot dirigible moored at the top of the building, held by a single cable tether, would add stress to the building's frame. The stress of the dirigible's load and the wind pressure would have to be transmitted all the way to the building's foundation, which was nearly eleven hundred feet below. The steel frame of the Empire State Building would have to be modified and strengthened to accommodate this new situation. Over sixty

thousand dollars' worth of modifications had to be made to the building's framework.

Rather than building a utilitarian mast without any ornamentation, the architects designed a shiny glass and chrome-nickel stainless steel tower that would be illuminated from inside, with a stepped-back design that imitated the overall shape of the building itself. The rocket-shaped mast would have four wings at its corners, of shiny aluminum, and would rise to a conical roof that would house the mooring arm. The winches and control machinery for the dirigible mooring would be housed in the base of the shaft itself, which also housed elevators and stairs to bring passengers down to the eighty-sixth floor, where baggage and ticket areas would be located.

The building would now be 102 floors, with a glassed-in observation area on the 101st floor and an open observation platform on the 102nd floor. This observation area was to double as the boarding area for dirigible passengers.

Once the architects had designed the mooring mast and made changes to the existing plans for the building's skeleton, construction proceeded as planned. When the building had been framed to the 85th floor, the roof had to be completed before the framing for the mooring mast could take place. The mast also had a skeleton of steel and was clad in stainless steel with glass windows. Two months after the workers celebrated framing the entire building, they were back to raise an American flag again—this time at the top of the frame for the mooring mast.

The Fate of the Mast

The mooring mast of the Empire State Building was destined to never fulfill its purpose, for reasons that should have been apparent before it was ever constructed. The greatest reason was one of safety: Most dirigibles from outside of the United States used hydrogen rather than helium, and hydrogen is highly flammable. When the German dirigible Hindenburg was destroyed by fire in Lakehurst, New Jersey, on May 6, 1937, the owners of the Empire State Building realized how much worse that accident could have been if it had taken place above a densely populated area such as downtown New York.

The greatest obstacle to the successful use of the mooring mast was nature itself. The winds on top of the building were constantly shifting due to violent air currents. Even if the dirigible were tethered to the mooring mast, the back of the ship would swivel around

and around the mooring mast. Dirigibles moored in open landing fields could be weighted down in the back with lead weights, but using these at the Empire State Building, where they would be dangling high above pedestrians on the street, was neither practical nor safe.

The other practical reason why dirigibles could not moor at the Empire State Building was an existing law against airships flying too low over urban areas. This law would make it illegal for a ship to ever tie up to the building or even approach the area, although two dirigibles did attempt to reach the building before the entire idea was dropped. In December 1930, the U.S. Navy dirigible Los Angeles approached the mooring mast but could not get close enough to tie up because of forceful winds. Fearing that the wind would blow the dirigible onto the sharp spires of other buildings in the area, which would puncture the dirigible's shell, the captain could not even take his hands off the control levers.

Two weeks later, another dirigible, the Goodyear blimp Columbia, attempted a publicity stunt where it would tie up and deliver a bundle of newspapers to the Empire State Building. Because the complete dirigible mooring equipment had never been installed, a worker atop the mooring mast would have to catch the bundle of papers on a rope dangling from the blimp. The papers were delivered in this fashion, but after this stunt the idea of using the mooring mast was shelved. In February 1931, Irving Clavan of the building's architectural office said, "The as yet unsolved problems of mooring air ships to a fixed mast at such a height made it desirable to postpone to a later date the final installation of the landing gear."

By the late 1930s, the idea of using the mooring mast for dirigibles and their passengers had quietly disappeared. Dirigibles, instead of becoming the transportation of the future, had given way to airplanes. The rooms in the Empire State Building that had been set aside for the ticketing and baggage of dirigible passengers were made over into the world's highest soda fountain and tea garden for use by the sightseers who flocked to the observation decks. The highest open observation deck, intended for disembarking passengers, has never been open to the public.

Appendix B Grading Rubrics of ASAP

B.1 Grading Rubric of $ASAP_1$

Score Point 1. An undeveloped response that may take a position but offers no more than very minimal support. Typical elements:

- Contains few or vague details.
- Is awkward and fragmented.
- May be difficult to read and understand.
- May show no awareness of audience.

Score Point 2. An under-developed response that may or may not take a position. Typical elements:

- Contains only general reasons with unelaborated and/or list-like details.
- Shows little or no evidence of organization.
- May be awkward and confused or simplistic.
- May show little awareness of audience.

Score Point 3. A minimally-developed response that may take a position, but with inadequate support and details. Typical elements:

- Has reasons with minimal elaboration and more general than specific details.
- Shows some organization.
- May be awkward in parts with few transitions.
- Shows some awareness of audience.

Score Point 4. A somewhat-developed response that takes a position and provides adequate support. Typical elements:

- Has adequately elaborated reasons with a mix of general and specific details.
- Shows satisfactory organization.
- May be somewhat fluent with some transitional language.

- Shows adequate awareness of audience.

Score Point 5. A developed response that takes a clear position and provides reasonably persuasive support. Typical elements:

- Has moderately well elaborated reasons with mostly specific details.
- Exhibits generally strong organization.
- May be moderately fluent with transitional language throughout.
- May show a consistent awareness of audience.

Score Point 6. A well-developed response that takes a clear and thoughtful position and provides persuasive support. Typical elements:

- Has fully elaborated reasons with specific details.
- Exhibits strong organization.
- Is fluent and uses sophisticated transitional language.
- May show a heightened awareness of audience.

B.2 Grading Rubric of $ASAP_2$

B.2.1 Domain 1: Writing Applications

Score Point 6. A Score Point 6 paper is rare. It fully accomplishes the task in a thorough and insightful manner and has a distinctive quality that sets it apart as an outstanding performance.

Ideas and Content

Does the writing sample fully accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Does it

- present a unifying theme or main idea without going off on tangents?
- stay completely focused on topic and task?

Does the writing sample include thorough, relevant, and complete ideas? Does it

- include in-depth information and exceptional supporting details that are fully developed?
- fully explore many facets of the topic?

Organization

Are the ideas in the writing sample organized logically? Does the writing

- present a meaningful, cohesive whole with a beginning, a middle, and an end (i.e., include an inviting introduction and a strong conclusion)?
- progress in an order that enhances meaning?
- include smooth transitions between ideas, sentences, and paragraphs to enhance meaning of text (i.e., have a clear connection of ideas and use topic sentences)?

Style Does the writing sample exhibit exceptional word usage? Does it

- include vocabulary to make explanations detailed and precise, descriptions rich, and actions clear and vivid (e.g., varied word choices, action words, appropriate modifiers, sensory details)?
- demonstrate control of a challenging vocabulary?

Does the writing sample demonstrate exceptional writing technique?

- Is the writing exceptionally fluent?
- Does it include varied sentence patterns, including complex sentences?
- Does it demonstrate use of writer's techniques (e.g., literary conventions such as imagery and dialogue and/or literary genres such as humor and suspense)?

Voice

Does the writing sample demonstrate effective adjustment of language and tone to task and reader? Does it

- exhibit appropriate register (e.g., formal, personal, or dialect) to suit task?
- demonstrate a strong sense of audience?
- exhibit an original perspective (e.g., authoritative, lively, and/or exciting)?

Score Point 5. A Score Point 5 paper represents a solid performance. It fully accomplishes the task, but lacks the overall level of sophistication and consistency of a Score Point 6 paper.

Ideas and Content

Does the writing sample fully accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Does it

- present a unifying theme or main idea without going off on tangents?
- stay focused on topic and task?

Does the writing sample include many relevant ideas? Does it

- provide in-depth information and more than adequate supporting details that are developed?
- explore many facets of the topic?

Organization

Are the ideas in the writing sample organized logically? Does the writing

- present a meaningful, cohesive whole with a beginning, a middle, and an end (i.e., include a solid introduction and conclusion)?
- progress in an order that enhances meaning of text?
- include smooth transitions (e.g., use topic sentences) between sentences and paragraphs to enhance meaning of text? (Writing may have an occasional lapse.)

Style

Does the writing sample exhibit very good word usage? Does it

- include vocabulary to make explanations detailed and precise, descriptions rich, and actions clear and vivid?
- demonstrate control of vocabulary?

Does the writing sample demonstrate very good writing technique?

- Is the writing very fluent?
- Does it include varied sentence patterns, including complex sentences?

- Does it demonstrate use of writer's techniques (e.g., literary conventions such as imagery and dialogue and/or literary genres such as humor and suspense)?

Voice

Does the writing sample demonstrate effective adjustment of language and tone to task and reader? Does it

- exhibit appropriate register (e.g., formal, personal, or dialect) to suit task?
- demonstrate a sense of audience?
- exhibit an original perspective (e.g., authoritative, lively, and/or exciting)?

Score Point 4. A Score Point 4 paper represents a good performance. It accomplishes the task, but generally needs to exhibit more development, better organization, or a more sophisticated writing style to receive a higher score.

Ideas and Content

Does the writing sample accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Does it

- present a unifying theme or main idea? (Writing may include minor tangents.)
- stay mostly focused on topic and task?

Does the writing sample include relevant ideas? Does it

- include sufficient information and supporting details? (Details may not be fully developed; ideas may be listed.)
- explore some facets of the topic?

Organization

Are the ideas in the writing sample organized logically? Does the writing

- present a meaningful whole with a beginning, a middle, and an end despite an occasional lapse (e.g., a weak introduction or conclusion)?
- generally progress in an order that enhances meaning of text?
- include transitions between sentences and paragraphs to enhance meaning of text? (Transitions may be rough, although some topic sentences are included.)

Style

Does the writing sample exhibit good word usage? Does it

- include vocabulary that is appropriately chosen, with words that clearly convey the writer's meaning?
- demonstrate control of basic vocabulary?

Does the writing sample demonstrate good writing technique?

- Is the writing fluent?
- Does it exhibit some varied sentence patterns, including some complex sentences?
- Does it demonstrate an attempt to use writer's techniques (e.g., literary conventions such as imagery and dialogue and/or literary genres such as humor and suspense)?

Voice

Does the writing sample demonstrate an attempt to adjust language and tone to task and reader? Does it

- generally exhibit appropriate register (e.g., formal, personal, or dialect) to suit task? (The writing may occasionally slip out of register.)
- demonstrate some sense of audience?
- attempt an original perspective?

Score Point 3. A Score Point 3 paper represents a performance that minimally accomplishes the task. Some elements of development, organization, and writing style are weak.

Ideas and Content

Does the writing sample minimally accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Does it

- attempt a unifying theme or main idea?
- stay somewhat focused on topic and task?

Does the writing sample include some relevant ideas? Does it

- include some information with only a few details, or list ideas without supporting details?
- explore some facets of the topic?

Organization

Is there an attempt to logically organize ideas in the writing sample? Does the writing

- have a beginning, a middle, or an end that may be weak or absent?
- demonstrate an attempt to progress in an order that enhances meaning? (Progression of text may sometimes be unclear or out of order.)
- demonstrate an attempt to include transitions? (Are some topic sentences used? Are transitions between sentences and paragraphs weak or absent?)

Style

Does the writing sample exhibit ordinary word usage? Does it

- contain basic vocabulary, with words that are predictable and common?
- demonstrate some control of vocabulary?

Does the writing sample demonstrate average writing technique?

- Is the writing generally fluent?
- Does it contain mostly simple sentences (although there may be an attempt at more varied sentence patterns)?
- Is it generally ordinary and predictable?

Voice

Does the writing sample demonstrate an attempt to adjust language and tone to task and reader? Does it

- demonstrate a difficulty in establishing a register (e.g., formal, personal, or dialect)?
- demonstrate little sense of audience?
- generally lack an original perspective?

Score Point 2. A Score Point 2 paper represents a performance that only partially accomplishes the task. Some responses may exhibit difficulty maintaining a focus. Others

may be too brief to provide sufficient development of the topic or evidence of adequate organizational or writing style.

Ideas and Content

Does the writing sample only partially accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Does it

- attempt a main idea?
- sometimes lose focus or ineffectively display focus?

Does the writing sample include few relevant ideas? Does it

- include little information and few or no details?
- explore only one or two facets of the topic?

Organization

Is there a minimal attempt to logically organize ideas in the writing sample?

- Does the writing have only one or two of the three elements: beginning, middle, and end?
- Is the writing sometimes difficult to follow? (Progression of text may be confusing or unclear.)
- Are transitions weak or absent (e.g., few or no topic sentences)?

Style

Does the writing sample exhibit minimal word usage? Does it

- contain limited vocabulary? (Some words may be used incorrectly.)
- demonstrate minimal control of vocabulary?

Does the writing sample demonstrate minimal writing technique?

- Does the writing exhibit some fluency?
- Does it rely mostly on simple sentences?
- Is it often repetitive, predictable, or dull?

Voice

Does the writing sample demonstrate language and tone that may be inappropriate to task and reader? Does it

- demonstrate use of a register inappropriate to the task (e.g., slang or dialect in a formal setting)?
- demonstrate little or no sense of audience?
- lack an original perspective?

Score Point 1. A Score Point 1 paper represents a performance that fails to accomplish the task. It exhibits considerable difficulty in areas of development, organization, and writing style. The writing is generally either very brief or rambling and repetitive, sometimes resulting in a response that may be difficult to read or comprehend.

Ideas and Content

Does the writing sample fail to accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Is it

- difficult for the reader to discern the main idea?
- too brief or too repetitive to establish or maintain a focus?

Does the writing sample include very few relevant ideas?

- Does it include little information with few or no details or unrelated details?
- Is it unsuccessful in attempts to explore any facets of the prompt?

Organization

Are the ideas in the writing sample organized illogically?

- Does it have only one or two of the three elements: beginning, middle, or end?
- Is it difficult to follow, with the order possibly difficult to discern?
- Are transitions weak or absent (e.g., without topic sentences)?

Style

Does the writing sample exhibit less than minimal word usage? Does it

- contain limited vocabulary, with many words used incorrectly?
- demonstrate minimal or less than minimal control of vocabulary?

Does the writing sample demonstrate less than minimal writing technique? Does it

- lack fluency?
- demonstrate problems with sentence patterns?

- consist of writing that is flat and lifeless?

Voice

Does the writing sample demonstrate language and tone that may be inappropriate to task and reader? Does it

- demonstrate difficulty in choosing an appropriate register?
- demonstrate a lack of a sense of audience?
- lack an original perspective?

B.2.2 Domain 2: Language Conventions

Score 4. Does the writing sample exhibit a superior command of language skills?

A Score Point 4 paper exhibits a superior command of written English language conventions. The paper provides evidence that the student has a thorough control of the concepts outlined in the Indiana Academic Standards associated with the student's grade level. In a Score Point 4 paper, there are no errors that impair the flow of communication. Errors are generally of the first-draft variety or occur when the student attempts sophisticated sentence construction.

- Does the writing sample demonstrate a superior command of capitalization conventions?
- Does the writing sample demonstrate a superior command of the mechanics of punctuation?
- Does the writing sample demonstrate a superior command of grade-level-appropriate spelling?
- Does the writing sample demonstrate a superior command of grammar and Standard English usage?
- Does the writing sample demonstrate a superior command of paragraphing?
- Does the writing sample demonstrate a superior command of sentence structure by not using run-on sentences or sentence fragments?

Score 3. Does the writing sample exhibit a good control of language skills?

In a Score Point 3 paper, errors are occasional and are often of the first-draft variety; they have a minor impact on the flow of communication.

- Does the writing sample demonstrate a good control of capitalization conventions?
- Does the writing sample demonstrate a good control of the mechanics of punctuation?
- Does the writing sample demonstrate a good control of grade-level-appropriate spelling?
- Does the writing sample demonstrate a good control of grammar and Standard English usage?
- Does the writing sample demonstrate a good control of paragraphing?
- Does the writing sample demonstrate a good control of sentence structure by only occasionally using run-on sentences or sentence fragments?

Score 2. Does the writing sample exhibit a fair control of language skills?

In a Score Point 2 paper, errors are typically frequent and may occasionally impede the flow of communication.

- Does the writing sample demonstrate a fair control of capitalization conventions?
- Does the writing sample demonstrate a fair control of the mechanics of punctuation?
- Does the writing sample demonstrate a fair control of grade-level-appropriate spelling?
- Does the writing sample demonstrate a fair control of grammar and Standard English usage?
- Does the writing sample demonstrate a fair control of paragraphing?
- Does the writing sample demonstrate a fair control of sentence structure by frequently using run-on sentences or sentence fragments?

Score 1. Does the writing sample exhibit a minimal or less than minimal control of language skills?

In a Score Point 1 paper, errors are serious and numerous. The reader may need to stop and reread part of the sample and may struggle to discern the writer's meaning.

- Does the writing sample demonstrate a minimal control of capitalization conventions?
- Does the writing sample demonstrate a minimal control of the mechanics of punctuation?
- Does the writing sample demonstrate a minimal control of grade-level-appropriate spelling?
- Does the writing sample demonstrate a minimal control of grammar and Standard English usage?
- Does the writing sample demonstrate a minimal control of paragraphing?
- Does the writing sample demonstrate a minimal control of sentence structure by using many run-on sentences or sentence fragments?

NOTE. The elements of this rubric are applied holistically; no element is intended to supersede any other element. The variety and proportion of errors in relation to the length of the writing sample are considered. A very brief paper consisting of two or three sentences may receive no more than 2 score points.

B.3 Grading Rubric of $ASAP_3$

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.

B.4 Grading Rubric of $ASAP_4$

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.

B.5 Grading Rubric of $ASAP_5$

Score Point 4. The response is a clear, complete, and accurate description of the mood created by the author. The response includes relevant and specific information from the memoir.

Score Point 3. The response is a mostly clear, complete, and accurate description of the mood created by the author. The response includes relevant but often general information from the memoir.

Score Point 2. The response is a partial description of the mood created by the author. The response includes limited information from the memoir and may include misinterpretations.

Score Point 1. The response is a minimal description of the mood created by the author. The response includes little or no information from the memoir and may include misinterpretations.

OR

The response relates minimally to the task.

Score Point 0. The response is incorrect or irrelevant or contains insufficient information to demonstrate comprehension.

B.6 Grading Rubric of $ASAP_6$

Score Point 4. The response is a clear, complete, and accurate description of the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. The response includes relevant and specific information from the excerpt.

Score Point 3. The response is a mostly clear, complete, and accurate description of the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. The response includes relevant but often general information from the excerpt.

Score Point 2. The response is a partial description of the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. The response

includes limited information from the excerpt and may include misinterpretations.

Score Point 1. The response is a minimal description of the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. The response includes little or no information from the excerpt and may include misinterpretations.
OR
The response relates minimally to the task.

Score Point 0. The response is totally incorrect or irrelevant, or contains insufficient evidence to demonstrate comprehension.

B.7 Grading Rubric of $ASAP_7$

A rating of 0-3 on the following four traits:

Ideas (points doubled)

Score 3. Tells a story with ideas that are clearly focused on the topic and are thoroughly developed with specific, relevant details. **Score 2.** Tells a story with ideas that are somewhat focused on the topic and are developed with a mix of specific and/or general details. **Score 1.** Tells a story with ideas that are minimally focused on the topic and developed with limited and/or general details. **Score 0.** Ideas are not focused on the task and/or are undeveloped.

Organization

Score 3. Organization and connections between ideas and/or events are clear and logically sequenced. **Score 2.** Organization and connections between ideas and/or events are logically sequenced. **Score 1.** Organization and connections between ideas and/or events are weak. **Score 0.** No organization evident.

Style

Score 3. Command of language, including effective and compelling word choice and varied sentence structure, clearly supports the writer's purpose and audience. **Score 2.** Adequate command of language, including effective word choice and clear sentences, supports the writer's purpose and audience. **Score 1.** Limited use of language, including lack of vari-

ety in word choice and sentences, may hinder support for the writer's purpose and audience. **Score 0.** Ineffective use of language for the writer's purpose and audience.

Conventions

Score 3. Consistent, appropriate use of conventions of Standard English for grammar, usage, spelling, capitalization, and punctuation for the grade level. **Score 2.** Adequate use of conventions of Standard English for grammar, usage, spelling, capitalization, and punctuation for the grade level. **Score 1.** Limited use of conventions of Standard English for grammar, usage, spelling, capitalization, and punctuation for the grade level. **Score 0.** Ineffective use of conventions of Standard English for grammar, usage, spelling, capitalization, and punctuation.

B.8 Grading Rubric of $ASAP_8$

A rating of 1-6 on the following six traits:

Ideas and Content

Score 6. The writing is exceptionally clear, focused, and interesting. It holds the reader's attention throughout. Main ideas stand out and are developed by strong support and rich details suitable to audience and purpose. The writing is characterized by

- clarity, focus, and control.
- main idea(s) that stand out.
- supporting, relevant, carefully selected details; when appropriate, use of resources provides strong, accurate, credible support.
- a thorough, balanced, in-depth explanation / exploration of the topic; the writing makes connections and shares insights.
- content and selected details that are well-suited to audience and purpose.

Score 5. The writing is clear, focused and interesting. It holds the reader's attention. Main ideas stand out and are developed by supporting details suitable to audience and purpose. The writing is characterized by

- clarity, focus, and control.
- main idea(s) that stand out.
- supporting, relevant, carefully selected details; when appropriate, use of resources provides strong, accurate, credible support.
- a thorough, balanced explanation / exploration of the topic; the writing makes connections and shares insights.
- content and selected details that are well-suited to audience and purpose.

Score 4. The writing is clear and focused. The reader can easily understand the main ideas. Support is present, although it may be limited or rather general. The writing is characterized by

- an easily identifiable purpose.
- clear main idea(s).
- supporting details that are relevant, but may be overly general or limited in places; when appropriate, resources are used to provide accurate support.
- a topic that is explored / explained, although developmental details may occasionally be out of balance with the main idea(s); some connections and insights may be present.
- content and selected details that are relevant, but perhaps not consistently well-chosen for audience and purpose.

Score 3. The reader can understand the main ideas, although they may be overly broad or simplistic, and the results may not be effective. Supporting detail is often limited, insubstantial, overly general, or occasionally slightly off-topic. The writing is characterized by

- an easily identifiable purpose and main idea(s).
- predictable or overly-obvious main ideas; or points that echo observations heard elsewhere; or a close retelling of another work.
- support that is attempted, but developmental details are often limited, uneven, somewhat off-topic, predictable, or too general (e.g., a list of underdeveloped points).

- details that may not be well-grounded in credible resources; they may be based on clichés, stereotypes or questionable sources of information.
- difficulties when moving from general observations to specifics.

Score 2. Main ideas and purpose are somewhat unclear or development is attempted but minimal. The writing is characterized by

- a purpose and main idea(s) that may require extensive inferences by the reader.
- minimal development; insufficient details.
- irrelevant details that clutter the text.
- extensive repetition of detail.

Score 1. The writing lacks a central idea or purpose. The writing is characterized by

- ideas that are extremely limited or simply unclear.
- attempts at development that are minimal or nonexistent; the paper is too short to demonstrate the development of an idea.

Organization

Score 6. The organization enhances the central idea(s) and its development. The order and structure are compelling and move the reader through the text easily. The writing is characterized by

- effective, perhaps creative, sequencing and paragraph breaks; the organizational structure fits the topic, and the writing is easy to follow.
- a strong, inviting beginning that draws the reader in and a strong, satisfying sense of resolution or closure.
- smooth, effective transitions among all elements (sentences, paragraphs, ideas).
- details that fit where placed.

Score 5. The organization enhances the central idea(s) and its development. The order and structure are strong and move the reader through the text. The writing is characterized by

- effective sequencing and paragraph breaks; the organizational structure fits the topic, and the writing is easy to follow.

- an inviting beginning that draws the reader in and a satisfying sense of resolution or closure.
- smooth, effective transitions among all elements (sentences, paragraphs, ideas).
- details that fit where placed.

Score 4. Organization is clear and coherent. Order and structure are present, but may seem formulaic. The writing is characterized by

- clear sequencing and paragraph breaks.
- an organization that may be predictable.
- a recognizable, developed beginning that may not be particularly inviting; a developed conclusion that may lack subtlety.
- a body that is easy to follow with details that fit where placed.
- transitions that may be stilted or formulaic.
- organization which helps the reader, despite some weaknesses.

Score 3. An attempt has been made to organize the writing; however, the overall structure is inconsistent or skeletal. The writing is characterized by

- attempts at sequencing and paragraph breaks, but the order or the relationship among ideas may occasionally be unclear.
- a beginning and an ending which, although present, are either undeveloped or too obvious (e.g., "My topic is..."; "These are all the reasons that...").
- transitions that sometimes work. The same few transitional devices (e.g., coordinating conjunctions, numbering, etc.) may be overused.
- a structure that is skeletal or too rigid.
- placement of details that may not always be effective.
- organization which lapses in some places, but helps the reader in others.

Score 2. The writing lacks a clear organizational structure. An occasional organizational device is discernible; however, the writing is either difficult to follow and the reader has to reread substantial portions, or the piece is simply too short to demonstrate organizational skills. The writing is characterized by

- some attempts at sequencing, but the order or the relationship among ideas is frequently unclear; a lack of paragraph breaks.
- a missing or extremely undeveloped beginning, body, and/or ending.
- a lack of transitions, or when present, ineffective or overused.
- a lack of an effective organizational structure.
- details that seem to be randomly placed, leaving the reader frequently confused.

Score 1. The writing lacks coherence; organization seems haphazard and disjointed. Even after rereading, the reader remains confused. The writing is characterized by

- a lack of effective sequencing and paragraph breaks.
- a failure to provide an identifiable beginning, body and/or ending.
- a lack of transitions.
- pacing that is consistently awkward; the reader feels either mired down in trivia or rushed along too rapidly.
- a lack of organization which ultimately obscures or distorts the main point.

Voice

Score 6. The writer has chosen a voice appropriate for the topic, purpose, and audience. The writer demonstrates deep commitment to the topic, and there is an exceptional sense of "writing to be read." The writing is expressive, engaging, or sincere. The writing is characterized by

- an effective level of closeness to or distance from the audience (e.g., a narrative should have a strong personal voice, while an expository piece may require extensive use of outside resources and a more academic voice; nevertheless, both should be engaging, lively, or interesting. Technical writing may require greater distance.).
- an exceptionally strong sense of audience; the writer seems to be aware of the reader and of how to communicate the message most effectively. The reader may discern the writer behind the words and feel a sense of interaction.
- a sense that the topic has come to life; when appropriate, the writing may show originality, liveliness, honesty, conviction, excitement, humor, or suspense.

Score 5. The writer has chosen a voice appropriate for the topic, purpose, and audience. The writer demonstrates commitment to the topic, and there is a sense of "writing to be read." The writing is expressive, engaging, or sincere. The writing is characterized by an appropriate level of closeness to or distance from the audience (e.g., a narrative should have a strong personal voice, while an expository piece may require extensive use of outside resources and a more academic voice; nevertheless, both should be engaging, lively, or interesting. Technical writing may require greater distance.).

- a strong sense of audience; the writer seems to be aware of the reader and of how to communicate the message most effectively. The reader may discern the writer behind the words and feel a sense of interaction.
- a sense that the topic has come to life; when appropriate, the writing may show originality, liveliness, honesty, conviction, excitement, humor, or suspense.

Score 4. A voice is present. The writer seems committed to the topic, and there may be a sense of "writing to be read." In places, the writing is expressive, engaging, or sincere. The writing is characterized by

- a suitable level of closeness to or distance from the audience.
- a sense of audience; the writer seems to be aware of the reader but has not consistently employed an appropriate voice. The reader may glimpse the writer behind the words and feel a sense of interaction in places.
- liveliness, sincerity, or humor when appropriate; however, at times the writing may be either inappropriately casual or personal, or inappropriately formal and stiff.

Score 3. The writer's commitment to the topic seems inconsistent. A sense of the writer may emerge at times; however, the voice is either inappropriately personal or inappropriately impersonal. The writing is characterized by

- a limited sense of audience; the writer's awareness of the reader is unclear.
- an occasional sense of the writer behind the words; however, the voice may shift or disappear a line or two later and the writing become somewhat mechanical.
- a limited ability to shift to a more objective voice when necessary.
- text that is too short to demonstrate a consistent and appropriate voice.

Score 2. The writing provides little sense of involvement or commitment. There is no evidence that the writer has chosen a suitable voice. The writing is characterized by

- little engagement of the writer; the writing tends to be largely flat, lifeless, stiff, or mechanical.
- a voice that is likely to be overly informal and personal.
- a lack of audience awareness; there is little sense of "writing to be read."
- little or no hint of the writer behind the words. There is rarely a sense of interaction between reader and writer.

Score 1. The writing seems to lack a sense of involvement or commitment. The writing is characterized by

- no engagement of the writer; the writing is flat and lifeless.
- a lack of audience awareness; there is no sense of "writing to be read."
- no hint of the writer behind the words. There is no sense of interaction between writer and reader; the writing does not involve or engage the reader.

Word Choice

Score 6. Words convey the intended message in an exceptionally interesting, precise, and natural way appropriate to audience and purpose. The writer employs a rich, broad range of words which have been carefully chosen and thoughtfully placed for impact. The writing is characterized by

- accurate, strong, specific words; powerful words energize the writing.
- fresh, original expression; slang, if used, seems purposeful and is effective.
- vocabulary that is striking and varied, but that is natural and not overdone.
- ordinary words used in an unusual way.
- words that evoke strong images; figurative language may be used.

Score 5. Words convey the intended message in an interesting, precise, and natural way appropriate to audience and purpose. The writer employs a broad range of words which have been carefully chosen and thoughtfully placed for impact. The writing is characterized by

- accurate, specific words; word choices energize the writing.
- fresh, vivid expression; slang, if used, seems purposeful and is effective.
- vocabulary that may be striking and varied, but that is natural and not overdone.
- ordinary words used in an unusual way.
- words that evoke clear images; figurative language may be used.

Score 4. Words effectively convey the intended message. The writer employs a variety of words that are functional and appropriate to audience and purpose. The writing is characterized by

- words that work but do not particularly energize the writing.
- expression that is functional; however, slang, if used, does not seem purposeful and is not particularly effective.
- attempts at colorful language that may occasionally seem overdone.
- occasional overuse of technical language or jargon.
- rare experiments with language; however, the writing may have some fine moments and generally avoids clichés.

Score 3. Language lacks precision and variety, or may be inappropriate to audience and purpose in places. The writer does not employ a variety of words, producing a sort of "generic" paper filled with familiar words and phrases. The writing is characterized by

- words that work, but that rarely capture the reader's interest.
- expression that seems mundane and general; slang, if used, does not seem purposeful and is not effective.
- attempts at colorful language that seem overdone or forced.
- words that are accurate for the most part, although misused words may occasionally appear; technical language or jargon may be overused or inappropriately used.
- reliance on clichés and overused expressions.
- text that is too short to demonstrate variety.

Score 2. Language is monotonous and/or misused, detracting from the meaning and impact. The writing is characterized by

- words that are colorless, flat or imprecise.
- monotonous repetition or overwhelming reliance on worn expressions that repeatedly detract from the message.
- images that are fuzzy or absent altogether.

Score 1. The writing shows an extremely limited vocabulary or is so filled with misuses of words that the meaning is obscured. Only the most general kind of message is communicated because of vague or imprecise language. The writing is characterized by

- general, vague words that fail to communicate.
- an extremely limited range of words.
- words that simply do not fit the text; they seem imprecise, inadequate, or just plain wrong.

Sentence Fluency

Score 6. The writing has an effective flow and rhythm. Sentences show a high degree of craftsmanship, with consistently strong and varied structure that makes expressive oral reading easy and enjoyable. The writing is characterized by

- a natural, fluent sound; it glides along with one sentence flowing effortlessly into the next.
- extensive variation in sentence structure, length, and beginnings that add interest to the text.
- sentence structure that enhances meaning by drawing attention to key ideas or reinforcing relationships among ideas.
- varied sentence patterns that create an effective combination of power and grace.
- strong control over sentence structure; fragments, if used at all, work well.
- stylistic control; dialogue, if used, sounds natural.

Score 5. The writing has an easy flow and rhythm. Sentences are carefully crafted, with strong and varied structure that makes expressive oral reading easy and enjoyable. The writing is characterized by

- a natural, fluent sound; it glides along with one sentence flowing into the next.

- variation in sentence structure, length, and beginnings that add interest to the text.
- sentence structure that enhances meaning.
- control over sentence structure; fragments, if used at all, work well.
- stylistic control; dialogue, if used, sounds natural.

Score 4. The writing flows; however, connections between phrases or sentences may be less than fluid. Sentence patterns are somewhat varied, contributing to ease in oral reading. The writing is characterized by

- a natural sound; the reader can move easily through the piece, although it may lack a certain rhythm and grace.
- some repeated patterns of sentence structure, length, and beginnings that may detract somewhat from overall impact.
- strong control over simple sentence structures, but variable control over more complex sentences; fragments, if present, are usually effective.
- occasional lapses in stylistic control; dialogue, if used, sounds natural for the most part, but may at times sound stilted or unnatural.

Score 3. The writing tends to be mechanical rather than fluid. Occasional awkward constructions may force the reader to slow down or reread. The writing is characterized by

- some passages that invite fluid oral reading; however, others do not.
- some variety in sentence structure, length, and beginnings, although the writer falls into repetitive sentence patterns.
- good control over simple sentence structures, but little control over more complex sentences; fragments, if present, may not be effective.
- sentences which, although functional, lack energy.
- lapses in stylistic control; dialogue, if used, may sound stilted or unnatural.
- text that is too short to demonstrate variety and control.

Score 2. The writing tends to be either choppy or rambling. Awkward constructions often force the reader to slow down or reread. The writing is characterized by

- significant portions of the text that are difficult to follow or read aloud.
- sentence patterns that are monotonous (e.g., subject-verb or subject-verb-object).
- a significant number of awkward, choppy, or rambling constructions.

Score 1. The writing is difficult to follow or to read aloud. Sentences tend to be incomplete, rambling, or very awkward. The writing is characterized by

- text that does not invite—and may not even permit—smooth oral reading.
- confusing word order that is often jarring and irregular.
- sentence structure that frequently obscures meaning.
- sentences that are disjointed, confusing, or rambling.

Conventions

Score 6. The writing demonstrates exceptionally strong control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar and usage) and uses them effectively to enhance communication. Errors are so few and so minor that the reader can easily skim right over them unless specifically searching for them. The writing is characterized by

- strong control of conventions; manipulation of conventions may occur for stylistic effect.
- strong, effective use of punctuation that guides the reader through the text.
- correct spelling, even of more difficult words.
- correct grammar and usage that contribute to clarity and style.
- skill in using a wide range of conventions in a sufficiently long and complex piece.
- little or no need for editing.

Score 5. The writing demonstrates strong control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar and usage) and uses them effectively to enhance communication. Errors are few and minor. Conventions support readability. The writing is characterized by

- strong control of conventions.
- effective use of punctuation that guides the reader through the text.

- correct spelling, even of more difficult words.
- correct capitalization; errors, if any, are minor.
- correct grammar and usage that contribute to clarity and style.
- skill in using a wide range of conventions in a sufficiently long and complex piece.
- little need for editing.

Score 4. The writing demonstrates control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar and usage). Significant errors do not occur frequently. Minor errors, while perhaps noticeable, do not impede readability. The writing is characterized by

- control over conventions used, although a wide range is not demonstrated.
- correct end-of-sentence punctuation; internal punctuation may sometimes be incorrect.
- spelling that is usually correct, especially on common words.
- correct capitalization; errors, if any, are minor.
- occasional lapses in correct grammar and usage; problems are not severe enough to distort meaning or confuse the reader.
- moderate need for editing.

Score 3. The writing demonstrates limited control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar and usage). Errors begin to impede readability. The writing is characterized by

- some control over basic conventions; the text may be too simple or too short to reveal mastery.
- end-of-sentence punctuation that is usually correct; however, internal punctuation contains frequent errors.
- spelling errors that distract the reader; misspelling of common words occurs.
- capitalization errors.
- errors in grammar and usage that do not block meaning but do distract the reader.
- significant need for editing.

Score 2. The writing demonstrates little control of standard writing conventions. Frequent, significant errors impede readability. The writing is characterized by

- little control over basic conventions.
- many end-of-sentence punctuation errors; internal punctuation contains frequent errors.
- spelling errors that frequently distract the reader; misspelling of common words often occurs.
- capitalization that is inconsistent or often incorrect.
- errors in grammar and usage that interfere with readability and meaning.
- substantial need for editing.

Score 1. Numerous errors in usage, spelling, capitalization, and punctuation repeatedly distract the reader and make the text difficult to read. In fact, the severity and frequency of errors are so overwhelming that the reader finds it difficult to focus on the message and must reread for meaning. The writing is characterized by

- very limited skill in using conventions.
- basic punctuation (including end-of-sentence punctuation) that tends to be omitted, haphazard, or incorrect.
- frequent spelling errors that significantly impair readability.
- capitalization that appears to be random.
- a need for extensive editing.

Appendix C Topical Components for MVP Corpus

C.1 Topic Words Results

Table 35 shows all topic words for the RTA_{MVP} from TC_{manual}. Table 36 shows all topic words for the RTA_{MVP} from TC_{lda}. Table 37 shows all topic words for the RTA_{MVP} from TC_{pr}. Table 38 shows all topic words for the RTA_{MVP} from TC_{attn}.

C.2 Specific Example Phrases Results

Table 39 shows all specific example phrases for the RTA_{MVP} from TC_{manual}. Table 40 shows all specific example phrases for the RTA_{MVP} from TC_{lda}. Table 41 shows all specific example phrases for the RTA_{MVP} from TC_{pr}. Table 42 shows all specific example phrases for the RTA_{MVP} from TC_{attnl}.

Topic 1	Topic 2	Topic 3	Topic 4
care	bed	farmer	school
health	net	fertilizer	supplies
hospital	malaria	irrigation	fee
treatment	infect	dying	student
doctor	bednet	crop	midday
electricity	mosquito	seed	meal
disease	bug	water	lunch
water	sleeping	harvest	supply
sick	die	hungry	book
medicine	cheap	feed	paper
generator	infect	food	pencil
no	biting	irrigation	energy
die			free
kid			children
bed			kid
patient			go
clinical			attend
officer			
running			

Table 35: Topic words of TC_{manual}.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
help	kenya	poverty	food	money	school	people	hospital	years
poor	like	think	fertilizer	need	kids	sauri	medicine	africa
world	better	author	crops	nets	supplies	malaria	hospitals	project
good	know	lifetime	water	thing	children	sick	water	villages
things	life	article	farmers	afford	schools	2008	free	sauri
time	help	possible	needed	donate	lunch	disease	electricity	village
work	think	convinced	grow	right	education	2004	diseases	helped
hard	sauri	fight	dying	dollar	afford	nets	medicines	change
going	live	proverty	problem	treatment	energy	mosquitoes	doctors	lives
alot	clothes	said	family	survive	learn	getting	2008	goals
reason	states	achievable	families	needs	students	says	gave	improved
happen	place	time	stop	stuff	went	years	doctor	2015
helping	health	convince	lack	person	adults	progress	examples	help
goal	important	believe	hunger	cause	fees	died	2004	changed
believe	feel	hannah	tools	patients	parents	text	shape	year
problems	happy	shows	seeds	provide	2004	away	cure	changes
countries	tell	reasons	plants	cost	lunches	mosquitos	running	started
difference	care	convincing	fertilizers	beds	books	prevent	treat	great
places	shoes	fighting	farming	means	home	treated	support	millennium
change	story	wrote	able	dont	wanted	dieing	common	progress
little	america	story	solved	dollars	chores	said	beds	came
improve	ways	agree	supply	medical	meal	come	patients	girl
country	wants	saying	irrigation	jobs	wood	night	said	2025
achieve	makes	opinion	wont	everyday	materials	bite	generator	place
hope	clothing	winning	afford	gone	learning	death	clean	program
helps	community	sachs	hungry	doctors	able	sleep	electricty	tells
everybody	economy	progress	plant	lots	suplies	impoverished	giving	small
start	history	conclusion	look	sickness	meals	living	drink	millenium
easy	paragraph	says	farms	live	paper	amazing	cures	read
making	thats	future	feed	fact	attendance	easily	evidence	happened

Table 36: Topic words of TC_{lda}.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
end	adults	village	millennium	thing	fight	work	people
					lifetime	world	kids
Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
paper	sleeping	diseases	midday	development	irrigation	plenty	doctor
supplies	bed	medicine	school	villages	fertilizer	access	hospital
chores	net	malaria	fees	project	farmers	care	shape
books	nets	disease	students	goals	crops	medicines	patients
pencils	site	mosquitoes	meal	plan	plant	schools	treatment
		charge	energy	economy	seeds	today	officer
			lunch	quality	outcome	supply	water
				supporters	lack	areas	electricity
					tools	kind	generator
Topic 17	Topic 18	Topic 19					
backs	joyful	road					
women	dirt	brighter					
ground	jump	future					
bananas	bar	hannah					
cloth	music	car					
mothers	singing	sauri					
feet	everyone	market					
clothing	dancing	year					
day	help	time					
rooms	health	place					
family	advice	years					
	items	poverty					
	targets	life					
	death	communities					
	night	leaders					
	costs	glimpse					
	die	africa					
	knowledge	chemicals					
	food	solutions					
	parents	millions					

Table 37: Topic words of TC_{pr}.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
poverty	hospital	school	lunch	free	electricity	goals	supply
fight	2004	schools	serves	medicine	water	problems	maintain
winning	disease	fees	parents	crops	generator	day	diseases
	yala	students	attend	charge	also	cloth	hunger
			passed	farmers	running	three	lives
				medicines	energy	made	adults
					connected	books	life
						2015	dying
						knowledge	death
						learn	away
						one	treated
Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
fertilizer	years	project	many	bed	supplies	afford	way
seeds	four	world	people	nets	food	lifetime	would
addressed	villages	millennium	kenya	used	net	could	rate
irrigation	80	village	sauri	every	rooms	achievable	attendance
necessary	progress	across	pencils	sleeping	packed	together	help
tools	last	work	africa	site	patients	malaria	kids
lack	occurred	end	yet	midday	needed	take	enough
plenty	year	worry	sachs	meal	5	future	better
plant	changes	supporters	though	dramatic	keep	worked	go
common	outcome	time	feed	change	poor	care	get
become	today	2025	two	clinical	five	family	place
	first	history	health	officer	like	hard	solutions
	along	selling	set	tattered	come	good	really
			crisis	clothing	little	doctor	targets
			areas	chemicals	treatment	either	see
			items	malarial	minimal	whole	die
				preventable	almost	save	hungry
				treatable	harvest	millions	dancing
				costs	showed	easy	walked
					cheap	met	bare
						ever	feet
						around	hannah
						mosquitoes	impoverished
						easily	encouraging
							probably

Table 38: Topic words of TC_{attn}.

Category 1	Category 2	Category 3
unpaved roads	united nations intervention	yala sub district hospital
tattered clothing	safer healthier better life	three kids bed two adults rooms packed patients
bare feet	out poverty stabilize economy quality life communities	not medicine treatment could afford
less than 1 dollar day	africa kenya sauri	no doctor only clinical officer running hospital
	goals met 2015 2025	no running water electricity
	80 villages across sub-sahara africa	sad people dying near death preventable
Category 4	**Category 5**	**Category 6**
malaria common disease preventable treatable	crops dying	kids not attend go school
mosquitoes carry malaria infect people biting	not afford fertilizer irrigation	not afford school fees
kids die malaria adults sick 20 000 day	outcome poor crops	kids help chores fetching water wood
bed nets mosquitoes away people save millions lives	lack fertilizer water	schools minimal supplies books paper pencils
bed nets cost 5 dollar	enough food crops harvest feed whole family hungry sick	concentrate not energy
cheap medicines treat malaria		no midday meal lunch
Category 7	**Category 8**	
progress just four years	progress encouraging supporters	
yala sub district hospital has medicine	solutions problems keep people impoverished	
medicine free charge	change poverty stricken areas good	
medicine most common diseases	poverty history not easy task hard	
water connected hospital	winning against poverty possible achievable lifetime	
hospital generator electricity		
bed nets used every sleeping site		
hunger crisis addressed fertilizer seeds		
tools needed maintain food supply		
kids go school now		
no school fees		
now serves lunch students		
school attendance rate way up		

Table 39: Specific example phrases of TC_{manual}.

Category 1	Category 2	Category 3
nets sleeping site sauri	years later	easy task
afford nets	took years	lived dollar
	started 2004	thing history
		stuff need
		earn money
Category 4	Category 5	Category 6
kids adults	achieve goal	donate money
2015 2025	reach goal	tattered clothes
hungry sick	going school	tattered clothing
cheap medicines	story says	bare feet
goals supposed	achieve goals	donating money
		save millions lives
Category 7	Category 8	Category 9
plan people poverty	clean water	yala subdistrict hospital medicine free charge common diseases
stabilize economy quality life communities	water wood	free lunch
assure access health care help	fresh water	yala district
people people	needs help	preventable treatable
near death	medicines free charge	common africa
poor crops lack	chores fetching	diseases like
homeless people	fetching water	common disease africa
		hospital good shape
		district hospital
Category 10	Category 11	Category 12
life time	children adults	stop poverty
united nations	mosquitoes carry malaria	long time
united states	disease called malaria	world work change
life communities	come night	beat poverty
like books paper pencils	malarial mosquitoes	ending poverty
learn life kenya	easily adults sick	want learn
important kids	solutions problems people impoverished	places like
thinks important	mosquitoes away	shows winning fight poverty achievable lifetime
wants know	infect people biting	want kind poverty
	away sleeping	poverty assure access
Category 13	Category 14	Category 15
amazing progress years	good shape	grow crops
text says	good education	feed family
text said	went school	needed help
year girl	areas good	farmers worry
year 2004	trying help	crops dying afford necessary fertilizer irrigation
paragraph says	worked hard	fertilizer knowledge
progress shows winning fight poverty achievable	second reason	hunger crisis addressed fertilizer seeds tools needed maintain food supply
treated chemicals	second example	feed families
paragraph states	girl went	hunger crisis adressed
progress encouraging supporters millennium villages	hannah sachs convinced winning	family plant seeds outcome poor
	went kenya	farmers worried
Category 16	Category 17	Category 18
running water electricity	millennium village project	attendance rate
water connected hospital generator electricity	millenium village project	midday meal
patients afford	millennium villages project helped	serves lunch students
rooms packed patients probably	change dramatically	midday meals
share beds	dramatic changes occured villages subsaharan africa	served lunch
recieve treatment	place live	students wanted learn
doctor clinical officer running hospital	happened years	books pencils
doctors clinical	dramatic changes occurred villages	kids attend school
water fertilizer knowledge	millennium development goals	schools minimal
receive treatment	change povertystricken areas good	schools hospitals
running bare	coming years	school school fees
afford treatment	encouraging supporters millennium villages project	practical items
	occurred villages subsaharan	kids sauri attend school parents afford school fees
		attendence rate
		parents money
Category 19	Category 20	
author convince winning fight poverty achievable lifetime	work hard	
author convinced winning fight poverty achievable lifetime	better place	
author wants	better health	
author convince winning fight proverty	brighter future	
winning fight proverty achievable lifetime	things like	
winning fight poverty achievable life time	things need	
article brighter future	fighting poverty	
wining fight poverty achievable	work change	
article states	hard work	
winning fight poverty acheivable	agree author	
author provided	working hard	
author thinks	better life 2008	
based article author convince	better life2008	
convinced poverty	reading article	
poverty acheivable lifetime	things changed	

Table 40: Specific example phrases of TC_{lda}.

Category 1
brighter future hannah
millennium villages project
unpaved dirt road
bar sauri primary school
future hannah
sauri primary school
villages project
millennium development goals
village leaders
dirt road
car jump
little kids
preventable diseases people
many kids
diseases people
kids die
school supplies
primary school
school fees
infect people

Table 41: Specific example phrases of TC_{pr}.

Category 1	Category 2	Category 3	Category 4
winning fight	could feed bed net afford	four years progress lifetime year	fees students school supplies schools
poverty winning world villages	people school work hard books	villages occurred 80 across along	school fees supplies afford fertilizer
winning fight poverty	also every diseases kids health	net 5	tools crops school fees seeds
winning poverty	preventable family people care	years many villages sauri project	farmers rooms patients crops people
fight poverty	afford school fees bed nets	outcome poor crops	school lunch meal midday supplies
poverty fight winning	also would energy learn help	progress years kenya africa today	lunch students serves midday
fight poverty winning	people fees school farmers could	rate people	medicine 2004 5 years keep
	lunch could work electricity medicine	villages kenya 80 farmers many	school lunch schools also fees
	could afford fertilizer	four years lifetime poverty year	years school showed hospital water
	school supplies little afford enough	years four last five day	school parents attend
	food also	farmers two many poverty	school medicine fertilizer hospital bed
	also tools	years changes fertilizer addressed	school schools fees free two
	supply maintain food also tattered	years villages kenya project attendance	school fees schools lunch free
		energy poverty hunger electricity	lunch school crops food farmers
			water fertilizer energy school medicines
Category 5	Category 6	Category 7	Category 8
sauri knowledge	supplies medicines	electricity running water irrigation set	bed showed diseases
afford school fees	better medicine water energy	poor showed treatment school supplies	lunch meal energy
bed nets help keep	hospital electricity connected	farmers could crops afford bed	dramatic change bed nets
food attendance rooms end many	bed nets 5 also	electricity hospital	poverty better lives made many
problems also people energy many	water electricity hospital fertilizer	better fertilizer medicine enough also	achievable lifetime sauri
food supply maintain electricity supplies	electricity water energy	rooms packed patients	malaria good bed net used
school fees	bed showed	food fertilizer crops get supply	bed net
2004 also year rate school	bed nets used	five net costs 5	common diseases
farmers needed food supply villages	generator energy	nets net bed free	work together poverty
	bed nets free	running water supplies schools almost	hospital go school could afford
	water electricity also fertilizer supplies	bed supplies knowledge medicines afford	project progress made food good
	electricity water running also generator	supplies food supply farmers water	also hospital doctor clinical showed
	generator electricity	supplies midday school food hunger	years made malaria take changes
	fertilizer bed net water	many food	could better future people lunch
	fertilizer addressed school supplies crisis		
Category 9	Category 10	Category 11	Category 12
help students supplies people schools	years four free schools medicine	medicine electricity tools fertilizer medicines	schools also school students attendance
people years four three though	school schools free supplies fees	water electricity connected schools running	free charge school maintain supply
villages years 80 poverty many	crops fertilizer farmers tools plant	students lunch serves school 2004	crops farmers 2004 first food
worked together end	water electricity supplies school energy	medicine crops free hospital also	lack fertilizer school bed nets
pencils students supplies yet	medicine school supplies years hunger	school supplies farmers attendance crops	bed nets years hospital
villages many kenya sauri 80	fertilizer crops lack farmers water	water supplies schools free hospital	hospital disease four years 2004
years food supply hunger crisis	fertilizer irrigation crops medicine water	schools crops supplies free charge	every sleeping site
sauri net	medicines school medicine fertilizer free	school schools lunch also free	school bed also occurred 80
net 5	free charge medicine school medicines	school fees schools	years four schools last students
school supplies items	seeds plant crops fertilizer	school fees lunch	school supplies schools also 2004
sachs many	free schools lunch school charge	lunch schools school seeds food	crops farmers schools project also
	bed nets water fertilizer medicines	school fees schools free lunch	hospital years medicine school water
	free charge medicine school fertilizer	schools supplies electricity farmers fertilizer	free charge schools years meal
	crops farmers fertilizer electricity knowledge	students lunch	medicine hospital made
	school fees schools free medicines	schools school farmers crops bed	free charge school years hunger
Category 13	Category 14	Category 15	Category 16
bed nets	villages africa millennium 80 across	supply books	seeds fertilizer addressed food medicine
water running medicine medicines supplies	80 villages across	electricity water	seeds supply fertilizer crops plenty
bed nets medicine crops electricity	poverty fight people kenya end	poverty many lives hunger every	fertilizer seeds crops
sauri free bed nets	world 2015	diseases lack water day every	tools fertilizer
crops fertilizer plant food irrigation	poor village sauri	adults one bed two last	crops farmers also water could
bed nets every water medicine	well project villages poor end	people food work many energy	crops seeds water needed
fertilizer crops water keep tools	achievable kenya	villages village school people many	addressed fertilizer seeds
kenya bed nets	many villages people problems kenya	school food schools hospital people	seeds fertilizer food also water
bed nets also adults	project villages kenya village people	years changes four free occurred	seeds fertilizer water
sauri bed nets	goals four years met needed	water every work school fees	fertilizer food
every bed nets	poverty village fight africa sauri	years hospital villages charge connected	fertilizer irrigation necessary farmers tools
diseases medicine medicines common preventable	attendance rate way selling come	food maintain supply electricity supplies	fertilizer seeds irrigation farmers lack
nets bed water sauri years	work world help last together	2015 2025 dying hunger death	fertilizer lack crops become sauri
crops fertilizer enough farmers	poverty many 2015 millennium progress	diseases malaria	
		site sauri	
Category 17	Category 18		
enough would work hard better	water connected hospital		
people world sauri kids poverty	nets bed used crops afford		
many people poverty could take	midday meal		
kenya would better walked bare	midday meal lunch		
poverty problems crisis though many	bed nets used		
people kenya targets 80 villages	bed every sleeping site net		
almost kids die people	bed nets every used school		
rate way progress better africa	hospital water running clinical officer		
attendance rate way	water hospital bed nets		
see world	bed nets could keep		
go hungry get people could	bed nets used every sleeping		
get food work would probably	hospital charge bed nets preventable		
world winning fight way place			
people easily sauri history way			
help people poverty place many			

Table 42: Specific example phrases of TC_{attn}.

Appendix D Topical Components for Space Corpus

D.1 Topic Words Results

Table 43 shows all topic words for the RTA_{Space} from TC_{manual}. Table 44 shows all topic words for the RTA_{Space} from TC_{lda}. Table 45 shows all topic words for the RTA_{Space} from TC_{pr}. Table 46 shows all topic words for the RTA_{Space} from TC_{attn}.

D.2 Specific Example Phrases Results

Table 47 shows all specific example phrases for the RTA_{Space} from TC_{manual}. Table 48 shows all specific example phrases for the RTA_{Space} from TC_{lda}. Table 49 shows all specific example phrases for the RTA_{Space} from TC_{pr}. Table 50 shows all specific example phrases for the RTA_{Space} from TC_{attnl}.

Topic 1	Topic 2	Topic 3	Topic 4
spending	earth	medicine	challenge
money	suffering	monitor	motivation
improve	pollution	astronauts	creative
life	fuel	scientists	knowledge
people	air	health	goals
hunger	oceans	stress	inspire
poverty	clean	safety	innovative
pay	energy	instruments	progress
housing	investment	doctors	problems
food	heal	body	competition
medicine	spending	exercise	explore
dying	dollars	machines	advancement
water	budget	airplanes	race
disease		weather	
malaria		technologies	
cost		innovations	
afford		inventions	
$5			
dollars			
budget			
problem			
satellite			
land			
condition			
crop			
soil			
rainfall			
drought			

Table 43: Topic words of TC_{manual}.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
space	space	earth	food	need
reason	stop	think	lives	thing
problems	helped	space	improve	cost
hunger	moon	helping	medicine	benefit
benefits	race	funding	says	great
problem	states	idea	point	worth
lead	explore	needs	evidence	country
rocket	rockets	travel	information	cause
exploration	wars	focus	text	sick
solve	science	issues	paragraph	disagree
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
things	world	money	education	help
space	good	alot	spent	technology
important	reasons	poor	government	helps
believe	author	spend	billion	life
article	stuff	save	budget	future
continue	said	spending	spend	work
fund	agree	opinion	dollars	knowledge
conclusion	needed	exploring	little	making
society	convinced	homeless	compared	research
best	discover	schools	uses	makes
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
nasa	better	people	poverty	like
satellites	pollution	africa	planet	know
ways	weather	malaria	live	going
scientists	airplanes	suffering	time	planets
stress	machines	dying	living	able
astronauts	fuel	diseases	humans	learn
crops	cars	countries	place	right
medical	created	afford	america	dont
monitor	ocean	nets	proverty	different
scientist	finding	disease	trying	happen

Table 44: Topic words of TC_{lda}.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
disease	arguments	safety	belief	stem
theft	question	health	space	decades
access	president	conditions	exploration	people
solutions	viewpoint	instruments	money	rocket
spread	point	doctors	spent	citizens
diseases	favor	body	program	poverty
malaria	challenges	reaction	year	countries
mosquito	innovations	satellites	government	care
bites	lots	land	dollars	help
africa	challenge	condition	budget	compare
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
example	medicine	half	suffering	life
	food	americans	hunger	lives
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
benefits	difficulty	power	homes	factories
area	housing	forms	water	
society		energy		
		cars		
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
gasoline	math	priority	machines	missions
pollution	investment	needs	exercise	scientists
fuels	education	earth	airplanes	astronauts
oil	progress	argue	weather	ways
air	science	problems	forecasting	planets
oceans		cost	technologies	
			engineers	

Table 45: Topic words of TC_{pr}.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
15	factories	hunger	better	crops	billion	medical	nations	track	people
americans	homes	example	machines	soil	dollars	instruments	competition	exploring	help
use	food	africa	exercise	rainfall	19	math	beneficial	affected	poverty
cost	us	suffering	airplanes	conditions	national	developed	states	scientists	countries
5	produce	instead	technologies	malaria	total	science	among	innovations	problems
million	distribute	space	include	fuels	670	education	motivate	life	also
budget	race	cold	forecasting	fossil	70	exploration	defense	deal	many
makes	cars	area	weather	drought	spends	knowledge	united	measure	already
need	human	way	disease	pollution	spent	improved	war	doctors	remain
consider	reaction	air	scientific	diseases	used	new	gasoline	clean	society
	spread	improve	satellites	burning	year	advancements	russia	significant	justify
	cleaner	saying	develop	condition	renewable	medicine	competed	however	live
	much	led	nasa	land		resulted	spaceships	lives	important
	power	tangible	harming	stress		inventions	progress	lead	solve
	rising	learned	called	mosquito		especially	needs	addition	believe
	meet	suffered	engineers			rocket	large	first	paying
		lots	information			providing		fired	could
		dying	oil			care		program	like
		oceans	really			american			bitten
		find	helps			spirit			nets

Table 46: Topic words of TC_{attn}.

Category 1	Category 2	Category 3
money spent differently used	dying because no access clean water	Earth suffering
improve people's lives	no medical care	pollution harming Earth
opposed space program	no disease prevention	pollution from burning fossil fuels
rocket fired theft	malaria spread mosquito bites	burning gas oil
suffer hunger poverty cause	malaria kills many people Africa	harming air oceans
46.2 million Americans	people dying Africa	need new cleaner forms energy
15% live in poverty	lower spread malaria with nets	program develop clean energy worthy investment
nearly half Americans	protect people from mosquitos	19 billion dollars could help Earth
cannot pay housing food medicine	nets cost $5	
19 billion dollars could help Americans	people cannot afford nets	
Category 4	**Category 5**	**Category 6**
19 billion dollars not too much	tangible benefits like medicine	scientists developed innovations improved lives
only 1.2% national budget	scientists monitored astronaut health	better exercise machines
670 billion spent national defense 26.3%	astronauts stressful conditions	better airplanes
70 billion spent education 4.8%	medical instruments developed	better weather forecasting
6.3 billion spent renewable energy	doctors learned about reaction to stress	
Category 7	**Category 8**	
hunger poverty tackled solved	important challenge provides motivation	
satellites monitor land	brings out best	
satellites track measure crops soil rainfall drought	remain creative society	
improves food production distribution	strive better technology	
solves serious problems	more scientific knowledge	
human suffering avoided	make progress	
compete with spaceships instead bomb-dropping airplanes	challenging goals innovative work	
	motivate beneficial competition among nations	
	Cold War	
	United States Russia competed	
	first land moon	
	visit other planets	
	space race	
	investment progress education math science	

Table 47: Specific example phrases of TC_{manual}.

Category 1	Category 2	Category 3
fund space exploration	space program	explore space
solve problems earth	space explorations	funding space exploration
second reason	land moon	needs earth
solve problem	united states russia competed	earth suffering
president eisenhower	especially math science	think fund space exploration
rocket fired theft citizens	spaceships instead	outer space
final reason	stop funding space exploration	helping people
exploration space	human suffering avoided	earth monitor lots land
problems world	bomb dropping	good idea
reason think	suffering pollution	space travel possible
investment space exploration	want learn	good thing
problems hunger poverty tackled space exploration	science math	needs help
problem hunger	significant investment progress american education	planet live
worth cost	suffering hunger poverty	long term
tangible benefits space exploration	race explore	think earth
suffered hunger poverty	instead bombdropping	long time
hunger proverty	russia competed prove greatness	helping solve problems earth
theft citizens suffered hunger	greatness race	
exploration good		
exploration lead		
Category 4	**Category 5**	**Category 6**
million americans live poverty	nets cost	things like
living poverty	africa year	article states
improve lives	afford nets	conclusion think
area medicine	malaria disease spread mosquito bites	space exploration funded
track measure condition crops soil rainfall drought	kills people	author convince space exploration desirable
trouble paying	disease called malaria	agree author
million dollars	dying access clean water medical care simple solutions	article says
food water	nets beds	author said
satellites circle earth monitor lots land	people people	article importance space exploration
grow food	bitten sleep	question consider
information improve produce distribute food	affected malaria	motivation bring
nearly half americans difficulty paying housing food medicine point lives	hanging large	believe pollution burning fossil fuels gasoline harming oceans
satellites monitor	nets protect	reading article believe
track measure conditions crops soil rainfall drought	suffer hunger	conclusion believe
example satellites	cure diseases	author convinced space exploration desirable needs earth
satellites track measure condition	afford food	author gave
improve life earth	millions people	challenge provides
rainfall droughts	people disease	space exploration helps remain creative society
live proverty	dying hunger	space exploration important motivate beneficial competition nations
monitor land	water food	bring best
Category 7	**Category 8**	**Category 9**
global warming	money spent	total national budget
life earth	waste money	renewable energy
scientific knowledge	exploring space	national defence
like africa	spending money	billions dollars
cleaner energy	spend money space exploration	favor space exploration
cleaner forms energy power cars homes factories	wasting money	argue billion dollars
world hunger	help solve problems	billion dollars spent education
like said	save lives	renewable clean energy
discover things	opinion think	billion dollar
stuff like	help money	spending billion
like malaria	heal people earth	government spends billion dollars year space exploration
place live	need cleaner forms energy power cars homes factories	education especially math science
world problems	instead spending	national defense 263
strive better technologies	money money	
innovative work	help space	
like cold	waste time	
finding ways	poor countries	
	people need	
Category 10		
better technology		
better place		
fossil fuel		
ways monitor		
learned human body reaction stress		
human problems		
improved life		
exercising machines		
learn human		
nasa scientists developed innovations improved lives		
helped doctors		
technologies scientific knowledge		
monitor health stressful conditions		
better exercise machines better airplanes better weather forecasting		
medical instruments developed doctors		
nasa allowed astronauts missions scientists ways		
doctors learn		
better exercise machines airplanes		
gasoline harming oceans		
astronauts health stressful conditions		

Table 48: Specific example phrases of TC_{lda}.

Category 1
space exploration people
space exploration stem
space exploration argue
space exploration
exploration people
space program
space travel
exploration stem
exploration argue
many people
many serious problems
much human suffering
much needs
serious problems
many tangible benefits
human problems
serious consideration
total national budget
many scientists
national budget
difficult problems
american education
several decades
disease spread
clean energy
money spent
national defense
tangible benefits
nasa scientists
worthy investment
significant investment
clean water
human suffering
large nets
challenging goals
human body
long-term benefits
stressful conditions
harsh conditions
simple solutions
medical care
innovative work
medical instruments
bomb-dropping airplanes
mosquito bites
creative society
fossil fuels
power cars
weather forecasting
exercise machines

Table 49: Specific example phrases of TC_{pr}.

Category 1	Category 2	Category 3
billion dollars 19 use 70	poverty africa people us 15	example fund defense saying reaction
billion also spent 670 nets	homes factories poverty condition africa	disease africa
national budget education poverty us	malaria spread disease people diseases	people earth also diseases poverty
dollars billion education spent 19	poverty people live countries paying	weather improved us russia competed
19 billion dollars need million	poverty people	exploration also education competition among
exploration 19 billion dollars	human reaction	education math especially science better
cost 5 people dollars	hunger poverty problems suffering help	problems people many us help
19 dollars billion use makes	malaria disease called diseases poverty	information human air
education million dollars	people americans poverty malaria million	medical education medicine improve help
use 19 billion dollars	gasoline oil	example astronauts knowledge
dollars national defense budget education	malaria people affected many africa	space exploration better also scientific
total national	human problems solve life exploration	race explore space exploration
americans 15	nations competed exploring	many inventions include doctors learned
also consider 15	rainfall national better weather forecasting	war cold africa like hunger
education 15 use used total	crops diseases also food suffering	area example cold war society
billion dollars 5 19 cost	cars homes factories	hunger poverty helping suffered suffering
19 billion dollars national budget	crops produce improve also pollution	led many tangible
	food distribute produce us fuels	food produce crops distribute pollution
	homes factories	exercise machines
Category 4	**Category 5**	**Category 6**
better technologies scientific satellites forecasting	fossil fuels burning pollution	billion dollars 19 70 cost
exercise machines better include technologies	condition crops soil solve malaria	billion dollars 670 19
exercise airplanes machines improve	crops rainfall soil drought land	19 billion
better exercise machines nations human	soil rainfall drought malaria disease	billion dollars 19 earth 15
exercise machines airplanes	crops soil conditions rainfall malaria	billion dollars spent 70 19
fossil fuels	crops soil rainfall	billion dollars used 19
us information solve better technologies	human reaction stress advancements	billion dollars us national
better forecasting weather airplanes us	pollution burning fuels fossil	total national 19 billion
malaria disease food called fuels	fossil fuels crops gasoline soil	billion dollars national exploration
medical advancements technologies scientific knowledge	malaria diseases disease like called	19 billion dollars
machines airplanes	conditions crops soil	19 billion dollars spends
medical better medicine air help	crops rainfall drought soil	billion dollars 19 70 education
machines better airplanes	condition crops soil	5 people
better exercise machines airplanes weather	soil crops rainfall	billion dollars 19 670 70
solve exploration us help helping		19 billion dollars used
machines airplanes improved exercise hunger		billion dollars 19 670
satellites technologies nasa engineers		70 billion dollars
fossil fuels harming gasoline like		billion dollars 19 exploration 670
		billion dollars 19 670 spends
		billion dollars national defense fund
Category 7	**Category 8**	**Category 9**
medical instruments developed suffering diseases	many year nations 5 people	program saying providing us better
exploration cold war inventions advancements	beneficial nations motivate competition among	race led instead significant war
math science	crops malaria rainfall factories cars	also help us remain exploration
medical instruments machines	national defense billion 70 exploration	help food produce pollution problems
medicine diseases malaria	us race war improved education	airplanes race instead human measure
education knowledge better	national us nations way distribute	countries crops million factories clean
education math science society human	better forecasting states satellites united	poverty cars homes people malaria
medical instruments developed knowledge new	fuels fossil gasoline poverty crops	instruments russia explore
math science improved	technologies scientific race exploring space	medical instruments machines
science math education exploring	competition among motivate nations national	exploration war track better society
exploration nations spaceships knowledge technologies	africa malaria crops hunger poverty	us airplanes exploration scientists satellites
education states american spirit math	exploring spaceships airplanes weather forecasting	15 resulted
medical advancements example knowledge society	national total 15 american people	5 afford live us
also improved developed example use	satellites us information earth exploration	problems earth solve space human
scientific new better knowledge medicine	technologies important national defense problems	states nations million many united
math science especially education improved	nations beneficial competition among national	rocket fired first earth
factories united states competed	states united war russia nations	many innovations machines led us
education medical national defense medicine	us airplanes crops land help	society justify countries nets cost
medical instruments developed	nations united states competed russia	
	advancements technologies new medical improved	

Table 50: Specific example phrases of TC_{attn}.

Bibliography

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Alikaniotis et al., 2016] Alikaniotis, D., Yannakoudakis, H., and Rei, M. (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.

[Amorim et al., 2018] Amorim, E., Cançado, M., and Veloso, A. (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.

[Attali and Burstein, 2004] Attali, Y. and Burstein, J. (2004). Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2).

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

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Baroni et al., 2014] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

[Barzilay and Lapata, 2008] Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[Blanchard et al., 2013] Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

[Blei and Lafferty, 2009] Blei, D. M. and Lafferty, J. D. (2009). Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Burstein et al., 2004] Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *Ai magazine*, 25(3):27–27.

[Burstein et al., 2001] Burstein, J., Kukich, K., Wolff, S., Lu, C., and Chodorow, M. (2001). Enriching automated essay scoring using discourse marking.

[Burstein et al., 1998] Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 206–210. Association for Computational Linguistics.

[Burstein et al., 2010] Burstein, J., Tetreault, J., and Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 681–684.

[Butnaru and Ionescu, 2017] Butnaru, A. M. and Ionescu, R. T. (2017). From image to text classification: A novel approach based on clustering word embeddings. *Procedia computer science*, 112:1783–1792.

[Cao et al., 2015] Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al. (2015). Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964.

[Cao et al., 2020] Cao, Y., Jin, H., Wan, X., and Yu, Z. (2020). Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1011–1020.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[Chen and He, 2013] Chen, H. and He, B. (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.

[Chen et al., 2018] Chen, L., Tao, J., Ghaffarzadegan, S., and Qian, Y. (2018). End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6234–6238. IEEE.

[Chen et al., 2010] Chen, Y.-Y., Liu, C.-L., Lee, C.-H., Chang, T.-H., et al. (2010). An unsupervised automated essay-scoring system. *IEEE Intelligent systems*, 25(5):61–67.

[Chernodub et al., 2019] Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., and Panchenko, A. (2019). Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras. https://keras.io.

[Correnti et al., 2013] Correnti, R., Matsumura, L. C., Hamilton, L., and Wang, E. (2013). Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.

[Cozma et al., 2018] Cozma, M., Butnaru, A., and Ionescu, R. T. (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.

[Crossley and McNamara, 2016] Crossley, S. A. and McNamara, D. S. (2016). *Adaptive educational technologies for literacy instruction*. Routledge.

[Cummins et al., 2016] Cummins, R., Zhang, M., and Briscoe, E. J. (2016). Constrained multi-task learning for automated essay scoring. Association for Computational Linguistics.

[Dasgupta et al., 2018] Dasgupta, T., Naskar, A., Dey, L., and Saha, R. (2018). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102.

[Dauphin et al., 2015] Dauphin, Y., de Vries, H., and Bengio, Y. (2015). Equilibrated adaptive learning rates for non-convex optimization. In *Advances in neural information processing systems*, pages 1504–1512.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Dong and Zhang, 2016] Dong, F. and Zhang, Y. (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.

[Dong et al., 2017] Dong, F., Zhang, Y., and Yang, J. (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.

[Ebrahimpour et al., 2019] Ebrahimpour, M. K., Li, J., Yu, Y.-Y., Reesee, J., Moghtaderi, A., Yang, M.-H., and Noelle, D. C. (2019). Ventral-dorsal neural networks: Object detection via selective attention. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 986–994. IEEE.

[Farag et al., 2018] Farag, Y., Yannakoudakis, H., and Briscoe, T. (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.

[Farra et al., 2015] Farra, N., Somasundaran, S., and Burstein, J. (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.

[Florescu and Caragea, 2017] Florescu, C. and Caragea, C. (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.

[Florescu and Jin, 2018] Florescu, C. and Jin, W. (2018). Learning feature representations for keyphrase extraction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Foltz and Rosenstein, 2015] Foltz, P. W. and Rosenstein, M. (2015). Analysis of a large-scale formative writing assessment system with automated feedback. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 339–342. ACM.

[Ghosh et al., 2016] Ghosh, D., Khanam, A., Han, Y., and Muresan, S. (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.

[Gierl et al., 2014] Gierl, M. J., Latifi, S., Lai, H., Boulais, A.-P., and De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical education*, 48(10):950–962.

[Granger et al., 2009] Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). International corpus oflearner english. version 2. *Louvain-la-Neuve: Presses universitaires de Louvain*.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Jin et al., 2018] Jin, C., He, B., Hui, K., and Sun, L. (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.

[Ke and Ng, 2019] Ke, Z. and Ng, V. (2019). Automated essay scoring: a survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308. AAAI Press.

[Kenter and de Rijke, 2015] Kenter, T. and de Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1411–1420. ACM.

[Kiela et al., 2015] Kiela, D., Hill, F., and Clark, S. (2015). Specializing word embeddings for similarity or relatedness. In *EMNLP*, pages 2044–2048.

[Kincaid et al., 1975] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

[Klebanov and Flor, 2013] Klebanov, B. B. and Flor, M. (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.

[Klebanov et al., 2016] Klebanov, B. B., Flor, M., and Gyawali, B. (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.

[Klebanov et al., 2014] Klebanov, B. B., Madnani, N., Burstein, J., and Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 247–252.

[Ko et al., 2019] Ko, W.-J., Durrett, G., and Li, J. J. (2019). Domain agnostic real-valued specificity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6610–6617.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Lei et al., 2016] Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

[Li et al., 2018] Li, X., Chen, M., Nie, J., Liu, Z., Feng, Z., and Cai, Y. (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.

[Liu et al., 2019] Liu, J., Xu, Y., and Zhao, L. (2019). Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.

[Louis and Higgins, 2010] Louis, A. and Higgins, D. (2010). Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on In-*

novative Use of NLP for Building Educational Applications, pages 92–95. Association for Computational Linguistics.

[Louis and Nenkova, 2013] Louis, A. and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

[Madnani et al., 2018] Madnani, N., Burstein, J., Elliot, N., Klebanov, B. B., Napolitano, D., Andreyev, S., and Schwartz, M. (2018). Writing mentor: Self-regulated writing feedback for struggling writers. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 113–117.

[Mahata et al., 2018] Mahata, D., Kuriakose, J., Shah, R. R., and Zimmermann, R. (2018). Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 634–639.

[Mann and Thompson, 1988] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

[Mayfield and Black, 2020] Mayfield, E. and Black, A. W. (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.

[McNamara et al., 2015] McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

[Meng et al., 2017] Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., and Chi, Y. (2017). Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Morris and Hirst, 1991] Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

[Nadeem et al., 2019] Nadeem, F., Nguyen, H., Liu, Y., and Ostendorf, M. (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.

[Nguyen and Litman, 2018] Nguyen, H. V. and Litman, D. J. (2018). Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Ong et al., 2014] Ong, N., Litman, D., and Brusilovsky, A. (2014). Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28.

[Östling et al., 2013] Östling, R., Smolentzov, A., Hinnerich, B. T., and Höglin, E. (2013). Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.

[Page, 1968] Page, E. B. (1968). The use of the computer in analyzing student essays. *International review of education*, 14(2):210–225.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[Persing et al., 2010] Persing, I., Davis, A., and Ng, V. (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.

[Persing and Ng, 2013] Persing, I. and Ng, V. (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.

[Persing and Ng, 2014] Persing, I. and Ng, V. (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1534–1543.

[Persing and Ng, 2015] Persing, I. and Ng, V. (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.

[Phandi et al., 2015] Phandi, P., Chai, K. M. A., and Ng, H. T. (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.

[Pitler and Nenkova, 2009] Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16.

[Rahimi and Litman, 2016] Rahimi, Z. and Litman, D. (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.

[Rahimi et al., 2017] Rahimi, Z., Litman, D., Correnti, R., Wang, E., and Matsumura, L. C. (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*, pages 1–35.

[Rahimi et al., 2014] Rahimi, Z., Litman, D. J., Correnti, R., Matsumura, L. C., Wang, E., and Kisa, Z. (2014). Automatic scoring of an analytical response-to-text assessment. In *International Conference on Intelligent Tutoring Systems*, pages 601–610. Springer.

[Ramachandran et al., 2015] Ramachandran, L., Cheng, J., and Foltz, P. (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.

[Rei and Cummins, 2016] Rei, M. and Cummins, R. (2016). Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. *arXiv preprint arXiv:1606.03144*.

[Röder et al., 2015] Röder, M., Both, A., and Hinneburg, A. (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.

[Roscoe et al., 2014] Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., and McNamara, D. S. (2014). The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, 34:39–59.

[Rudner and Liang, 2002] Rudner, L. M. and Liang, T. (2002). Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

[Seo et al., 2017] Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[Shermis and Burstein, 2003] Shermis, M. D. and Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

[Shibani et al., 2019] Shibani, A., Knight, S., and Shum, S. B. (2019). Contextualizable learning analytics design: A generic model and writing analytics evaluations. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 210–219. ACM.

[Somasundaran et al., 2014] Somasundaran, S., Burstein, J., and Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers*, pages 950–961.

[Song et al., 2017] Song, W., Wang, D., Fu, R., Liu, L., Liu, T., and Hu, G. (2017). Discourse mode identification in essays. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 112–122.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

[Stab and Gurevych, 2014] Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

[Taghipour and Ng, 2016] Taghipour, K. and Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

[Tay et al., 2018] Tay, Y., Phan, M. C., Tuan, L. A., and Hui, S. C. (2018). Skipflow: incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Taylor et al., 2003] Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer.

[Theano Development Team, 2016] Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.

[Uto et al., 2020] Uto, M., Xie, Y., and Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088.

[Vajjala, 2018] Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

[Witten et al., 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Woods et al., 2017] Woods, B., Adamson, D., Miel, S., and Mayfield, E. (2017). Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2071–2080. ACM.

[Xie et al., 2012] Xie, S., Evanini, K., and Zechner, K. (2012). Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–111. Association for Computational Linguistics.

[Yamamoto et al., 2019] Yamamoto, M., Umemura, N., and Kawano, H. (2019). Proposal of japanese vocabulary difficulty level dictionaries for automated essay scoring support system using rubric. *Journal of the Operations Research Society of China*, pages 1–17.

[Yannakoudakis and Briscoe, 2012] Yannakoudakis, H. and Briscoe, T. (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.

[Yannakoudakis et al., 2011] Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.

[Yuxin et al., 2018] Yuxin, P., Xiangteng, H., and Junjie, Z. (2018). Object-part attention model for fine-grained image classification. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 27(3):1487–1500.

[Zesch et al., 2015a] Zesch, T., Heilman, M., and Cahill, A. (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.

[Zesch et al., 2015b] Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (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.

[Zhang and Litman, 2017] Zhang, H. and Litman, D. (2017). Word embedding for response-to-text assessment of evidence. In *Proceedings of ACL 2017, Student Research Workshop*, pages 75–81.

[Zhang and Litman, 2018] Zhang, H. and Litman, D. (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.

[Zhang and Litman, 2020] Zhang, H. and Litman, D. (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.

[Zhang et al., 2019] Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L., Howe, E., and Quintana, R. (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.