Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays

Zixuan Ke **Hrishikesh Inamdar** **Hui Lin** **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{zixuan,hui,vince}@hlt.utdallas.edu, hai160030@utdallas.edu

Abstract

While the vast majority of existing work on automated essay scoring has focused on holistic scoring, researchers have recently begun work on scoring specific dimensions of essay quality. Nevertheless, progress in dimension-specific essay scoring research is hindered in part by the lack of annotated corpora. To facilitate advances in this area of research, we design a rubric for scoring an important, yet unexplored dimension of persuasive essay quality, thesis strength, and annotate a corpus of essays with thesis strength scores. We additionally identify the attributes that could impact thesis strength and annotate the essays with the values of these attributes, which, when predicted by computational models, could provide feedback to students on why her essay receives a particular thesis strength score.

1 Introduction

Recent work on automated essay scoring has largely focused on *holistic* scoring, which summarizes the quality of an essay with a single score (e.g., Taghipour and Ng (2016), Dong et al. (2017), Wang et al. (2018)). There are at least two reasons for this focus. First, corpora manually annotated with holistic scores such as the one used in the Kaggle-sponsored ASAP competition[1] are publicly available, facilitating the training and evaluation of holistic essay scoring engines. Second, holistic scoring technologies are commercially valuable: being able to successfully automate the scoring of the millions of essays written for aptitude tests such as SAT, GRE, and GMAT every year can save a lot of manual grading effort.

However, holistic essay scoring technologies are far from adequate for use in classroom settings, where providing students with feedback on how to improve their essays is of utmost importance. Specifically, merely returning a low holistic score to an essay provides essentially no feedback to its author on which aspect(s) of the essay contributed to the low score and how it can be improved. Recently, researchers have attempted to score a particular dimension of essay quality such as coherence (Miltsakaki and Kukich, 2004), technical errors, relevance to prompt (Higgins et al., 2004; Louis and Higgins, 2010; Persing and Ng, 2014), organization (Persing et al., 2010), thesis clarity (Persing and Ng, 2013), and argument persuasiveness (Persing and Ng, 2015; Ke et al., 2018). Automated systems that provide instructional feedback along multiple dimensions of essay quality such as *Criterion* (Burstein et al., 2004) have also begun to emerge. Providing scores along different dimensions of essay quality could help an author identify which aspects of her essay need improvements. Unfortunately, progress in dimension-specific essay scoring research is hampered in part by the lack of annotated corpora needed to train and evaluate systems for scoring essays along specific dimensions of essay quality.

Motivated by this observation, we aim to contribute to dimension-specific essay scoring research in this paper by creating the resources needed to empirically study *thesis strength*, a fundamental yet unexplored dimension of essay quality. Thesis strength refers to how strong the thesis statement in a persuasive essay is. A thesis statement summarizes the main point the author is trying to argue for in her essay in the form of a *claim* (i.e., a statement that is controversial and therefore can be argued) and states why the essay is important and worth reading. Hence, in addition to being clear, concise, specific, and relevant to the prompt the essay is written for, a strong thesis statement should briefly provide evidences for the author's claim, justifications for the importance of the claim, and possibly a roadmap for the essay.

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

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3994–4004
Florence, Italy, July 28 - August 2, 2019.

A strong thesis statement can help lay a strong foundation for the rest of the essay by organizing its content, improving its comprehensibility, and ensuring its relevance to the prompt. In contrast, an essay with a weak thesis statement lacks focus. Hence, an essay's thesis strength can be expected to have a strong influence on its holistic score.

To facilitate the computational study of thesis strength scoring in student essays, we design a rubric and use it to annotate a corpus of 1021 persuasive student essays with their thesis strength scores. One may argue that the feedback provided by a thesis strength score is limited: if a student receives a low score, she may still not know why her score is low. To address this concern, we identify the attributes that could impact thesis strength, design a scoring rubric for each of them and annotate the essays in our corpus with the values of these attributes. Not only can these attributes serve to explain a thesis strength score, but they could provide additional feedback to a student on why she receives a particular thesis strength score when predicted by a computational model.

An important yet often under-emphasized issue is which corpus of essays we should annotate. We envision that in the long run, substantial progress in this area of research can only be made if different researchers on automated essay grading create their annotations on the same corpus of essays. For instance, having a corpus of essays that are scored along different dimensions of quality, such as organization, prompt adherence, and thesis strength will facilitate the study of how these dimensions interact with each other to produce a holistic score. As another example, researchers working on automated essay *revision* (Zhang et al., 2017), where the goal is to revise, for instance, a thesis statement or an argument in an essay to make it stronger, would benefit from having the thesis strength scores we annotate. Specifically, the first step in deciding *how* to revise a thesis statement to make it stronger is to understand *why* it is weak, and the aforementioned attributes that we propose to annotate will provide insights into what makes a thesis statement weak and subsequently how to revise it. So, having both the attributes and the revised thesis annotated on the same set of essays will allow researchers to study how they interact and facilitate the design of joint models that capture such interactions. Unfortunately, existing essay annotations are spread over different corpora, some of which are not even publicly available. With this in mind, we choose to annotate a corpus of essays that have recently been scored along several dimensions of essay quality, the ICLE corpus (Granger et al., 2009). To stimulate research in thesis strength, we will make all of our annotations publicly available.[2]

2 Related Work

In this section, we provide an overview of the popularly used annotated essay corpora for *scoring*.

Holistic scoring. As mentioned before, the ASAP corpus, which was produced as part of a Kaggle competition, has recently been used extensively to evaluate holistic essay scoring systems. It contains holistically scored essays written for eight prompts by American students from grades 7 through 10, with 1190−3000 essays for each prompt. CLC-FCE (Yannakoudakis et al., 2011) is a relatively small corpus that contains 1244 essays written for 10 prompts by ESOL test takers. Each essay is not only holistically scored but also annotated with different kinds of errors, and therefore the corpus can also be used for grammatical error detection and correction. A Swedish corpus containing 1702 holistically scored essays written for 19 prompts by high school students is also publicly available (Östling et al., 2013).

Dimension-specific scoring. The Argument Annotated Essays corpus contains 402 essays taken from *essayforum2*, a site offering feedback to students wishing to improve their ability to write persuasive essays for tests (Stab and Gurevych, 2014). Each essay in the corpus is annotated with its argumentative structure (i.e., argument components such as claims and premises as well as the relationships between them (e.g., support, attack)). The corpus has been used extensively to evaluate argument mining systems. Recently, Carlile et al. (2018) annotated each argument in 102 essays randomly selected from the corpus with its persuasiveness score.

There are two corpora of essays that are scored along multiple dimensions of quality. Horbach et al. (2017) annotated a corpus of 2200 German essays written by prospective university students. Each essay is a summary of a given news article and is manually scored w.r.t. coherence, organiza-

[2] Our annotated corpus is publicly available at the website `http://www.hlt.utdallas.edu/~zixuan/EssayScoring`.

tion, argumentation, style, and grammar. Neither the essays nor the annotations are publicly available, however. The second corpus is the International Corpus of Learner English (ICLE) (Granger et al., 2009). ICLE is composed of essays written by university undergraduates. Approximately 1000 persuasive essays in the corpus are manually scored w.r.t. organization (Persing et al., 2010), thesis clarity (Persing and Ng, 2013), prompt adherence (Persing and Ng, 2014), and argument persuasiveness (Persing and Ng, 2015). Given the public availability of these scores, we believe that it is beneficial to additionally score the ICLE essays w.r.t. thesis strength.

At first glance, the aforementioned thesis clarity dimension studied by Persing and Ng (2013) appears to resemble the thesis strength dimension. Despite the fact that both dimensions are concerned with an essay's thesis, thesis clarity refers to how *clear* an essay's thesis is and can be viewed as an *attribute* that could affect thesis strength: intuitively, if a thesis statement is not clear, then it is unlikely to be strong. As we will see, besides thesis clarity, there are many attributes that could impact thesis strength. Argument persuasiveness, another essay scoring dimension studied by Persing and Ng (2015), also appears to be relevant to thesis strength since it refers to the persuasiveness of the argument an essay makes for its thesis. To see the difference between these two dimensions, recall that whether an argument is persuasive or not depends in part on how strong the supporting evidences are for its claim. In contrast, while a thesis statement is expected to provide evidences in support of the claim it states, the strength of the thesis statement does *not* depend on the strength of the support. In other words, while persuasiveness is adversely affected by the presence of *weak* evidences in the argument, thesis strength is adversely affected by the *absence* of evidences rather than the presence of weak evidences.

3 Corpus

As mentioned above, we use as our corpus the 4.5 million word ICLE, which consists of more than 6000 essays on a variety of writing topics written by university undergraduates from 16 countries and 16 native languages who are learners of English as a Foreign Language. 91% of the ICLE texts are written in response to prompts that trigger persuasive essays. We selected 1021 persuasive essays to annotate. As discussed above, since it is beneficial to have a corpus of essays annotated along multiple dimensions of quality, these 1021 essays are selected to maximize the overlap with those previously annotated by Persing and Ng, as mentioned above. These essays were written for 13 prompts and have 7.6 paragraphs, 31.3 sentences, and 680.9 tokens on average.

sentences in thesis	0	1	2	3	4
essays	228	411	260	97	25

Table 1: Distribution of essays over the number of sentences in the thesis statement.

4 Annotation

4.1 Annotation Scheme

For each of the 1021 essays, we produce three kinds of annotations. We (1) identify its thesis statement (if any), and score (2) its strength as well as (3) the attributes that could impact its strength.

Thesis statement identification. According to the definitions collected from different essay writing resources (Anson and Schwegler, 2011; Ruszkiewicz, 2010; Lunsford, 2015; Ramage et al., 2018), a thesis statement offers a concise summary of the main idea of an essay. It is usually expressed in one sentence and can be reiterated elsewhere. It often includes the stance of the author and usually leads the whole (or at least part of an) essay. It helps organize and develop the body of the essay, letting the readers know what the writer's statement is and what it aims to prove.

Since thesis statements are typically realized as sentences, we take a sentence as the basic unit of our annotation. Table 1 shows the distribution of the 1021 essays over the number of sentences in a thesis statement. As we can see, 228 of our essays contains no thesis statement, whereas approximately 40% of them contain exactly one sentence.

Thesis strength scoring. We develop a rubric for scoring the strength of an essay's thesis statement. Motivated by the rubric typically used for scoring essays written for standardized aptitude tests such as GRE, we evaluate a thesis statement's strength using a numerical score from 1 to 6, with a score of 6 indicating a very strong thesis and a score of 1 indicating the absence of a thesis in the corresponding essay. A description of each score can be found in the rubric shown in Table 2.

Attribute scoring. Aiming to provide feedback to a student on why she receives a particular thesis

Score	Description
6	A very strong thesis: Little can be done to strengthen the thesis.
5	A strong thesis: Only minor changes can be made to strengthen the thesis.
4	A decent thesis: The thesis is generally good, though it can be strengthened in various aspects.
3	A poor, understandable thesis: It may only be partially clear or contain severe errors that detract from its strength.
2	It is unclear what the author is trying to argue in the thesis (e.g., the thesis is not understandable; it is not relevant to the prompt; the thesis presents opposing views).
1	The essay presents no thesis of any kind.

Table 2: Description of the Thesis Strength scores.

Score	Description
3	Arguable: The thesis expresses the author's stance and opinion w.r.t. to the essay's topic and contains a controversial statement that should not be accepted by readers without additional support.
2	Confusing: The thesis appears to present conflicting views, or the author fails to express her stance.
1	Unarguable: The thesis merely describes some events or facts.

Table 3: Description of the Arguability scores.

Score	Description
3	Specific: The thesis addresses the question of ”what is the opinion expressed in the thesis” specifically. No concept in the thesis needs to be more specific in order to adequately answer this question.
2	Partially specific: The thesis addresses the question of ”what is the opinion expressed in the thesis” broadly. One concept needs to be more specific in order to adequately answer this question.
1	General: The thesis addresses the question of ”what is the opinion expressed in the thesis” very broadly. More than one concept needs to be made more specific in order to adequately answer this question.

Table 4: Description of the Specificity scores.

Score	Description
3	Clear: Readers can easily understand what the opinion is.
2	Moderately clear: Readers have some difficulty understanding what the opinion is.
1	Not understandable: Readers can hardly understand what the opinion is.

Table 5: Description of the Clarity scores.

strength score, we identify a set of 10 attributes that could impact a thesis statement's strength. Since these are attributes of a thesis statement, they are computed solely based on a thesis statement. Below we describe these 10 attributes.

Arguability concerns whether the claim underlying the thesis can be supported or refuted with evidences. *Specificity* concerns the narrowness of the concepts referred to in a thesis statement. Concepts that are specific are more believable because they indicate an author's depth of knowledge about a subject matter. *Clarity* is how clear and understandable the thesis is. *Relevance to Prompt* is the extent to which the thesis is relevant to the prompt. *Conciseness* is how concise the idea underlying the thesis is expressed. *Eloquence* is how well the author uses language to convey ideas, similar to fluency. *Confidence* refers to how confident the author is in the truthfulness of her thesis. *Direction of Development* refers to the extent to which the thesis directs the essay's development. *Justification of Opinion* refers to the extent to which the author justifies her opinion(s) expressed in the thesis. Finally, *Justification of Importance or Interest* refers to the extent to which the author justifies why her thesis is important and/or interesting.

The rubrics for scoring these attributes are shown in Tables 3−12. Each attribute is scored on a scale of 1−3. While the meaning of these scores differs from one attribute to another, generally speaking, '1' means ”no”, '2' means ”partially”, and '3' means ”yes”. We hypothesize that a high score on any of these attributes would have a positive impact on the thesis strength score. Since these attributes are associated with a thesis statement, for essays that do not have a thesis statement, the values of these attributes will be undefined (i.e., the attributes will not be scored).

4.2 Annotation Procedure

Our 1021 essays are annotated by three human annotators. We first familiarized them with the definition of a thesis statement stated in the previous subsection as well as the rubrics for scoring thesis strength and the 10 attributes, and then trained

Score	Description
3	Relevant: The thesis is clearly relevant to the prompt.
2	Partially relevant: The thesis is only partially relevant to the prompt.
1	Irrelevant: The thesis does not respond to the prompt.

Table 6: Description of the Relevance to Prompt scores.

Score	Description
3	Concise: All concepts in the thesis are expressed in the most effective words or phrases: a) no word or phrase can be replaced with a more powerful one without losing any of its value; b) no word or phrase can be deleted without losing any of its value; c) no sentence can be easily inserted into another sentence without losing any of its value.
2	Partially concise: There is a concept in the thesis that is not expressed in the most effective words or phrases.
1	Verbose: More than one concept in the thesis is not expressed in the most effective words.

Table 7: Description of the Conciseness scores.

Score	Description
3	Demonstrates mastery of English: There are no grammatical errors that detract from the meaning of the sentence. Exhibits a well thought out, flowing sentence structure that is easy to read and conveys the idea exceptionally well.
2	Demonstrates competence in English: There might be a few noticeable but forgivable errors, such as an incorrect verb tense or unnecessary pluralization. Demonstrates a typical vocabulary and a simple sentence structure.
1	Demonstrates poor understanding of sentence composition and/or poor vocabulary: The choice of words or grammatical errors force the reader to reread the sentence before moving on.

Table 8: Description of the Eloquence scores.

Score	Description
3	Confident: The author has a firm attitude to all of her opinions and takes an authoritative stance. No statement can reduce her credibility and weaken her statement.
2	Occasionally confident: The author has a firm attitude to some of her opinions.
1	Not Confident: The author does not have a firm attitude to any of her opinions.

Table 9: Description of the Confidence scores.

Score	Description
3	Clear roadmap: The thesis suggests full paths for the essay's development and informs readers of what will be discussed in the body of the essay. It addresses the question of ”what can be expected from the essay”.
2	Partially clear roadmap: The thesis suggests some paths for the essay's development and informs readers of what will be discussed in its body. It partially addresses the question of ”what can be expected from the essay”.
1	No roadmap: The thesis fails to suggest paths for the essay's development and does not inform readers of what will be discussed in its body. It fails to address the question of ”what can be expected from the essay”.

Table 10: Description of the Direction of Development scores.

Score	Description
3	Well justified: The thesis justifies the author's opinion(s) regardless of how convincing the justification is.
2	Partially justified: The thesis justifies some of the author's opinion(s).
1	Unjustified: The thesis fails to justify the author's opinion(s).

Table 11: Description of the Justification of Opinion scores.

Score	Description
3	Well justified: The thesis justifies why every opinion expressed in it is important or interesting.
2	Partially justified: The thesis justifies why some of the opinions expressed in the thesis are important or interesting.
1	Unjustified: The thesis fails to justify why the author's opinion is important or interesting.

Table 12: Description of the Justification of Importance/Interest scores.

them on 10 essays (not included in our corpus). After that, they were asked to identify the thesis statements in a randomly selected subset of 120 essays and discuss the resulting annotations to resolve any discrepancies. After they agreed on the thesis statements in these 120 essays, they were asked to score the strength of each of these statements and the associated attributes. Discrepancies were resolved through open discussion. Finally, the remaining essays were partitioned into three

Attribute	1	2	3	4	5	6
Thesis Strength	228	73	163	247	191	119
Arguability	5	32	756	–	–	–
Specificity	13	78	702	–	–	–
Clarity	11	28	754	–	–	–
Relevance	103	171	519	–	–	–
Conciseness	5	46	742	–	–	–
Eloquence	25	125	643	–	–	–
Confidence	40	23	730	–	–	–
Dir. of Dev.	734	4	55	–	–	–
Just. Opinion	533	24	236	–	–	–
Just. Imp./Int.	675	14	104	–	–	–

Table 13: Distribution of scores for Thesis Strength and the attributes.

Attribute	α
Thesis Strength	.635
Arguability	.657
Specificity	.530
Clarity	.748
Relevance to Prompt	.550
Conciseness	.581
Eloquence	.532
Confidence	.624
Direction of Development	.856
Justification of Opinion	.787
Justification of Importance/Interest	.635

Table 14: Inter-annotator agreement on each attribute in terms of Krippendorff's α.

sets of roughly the same size, and each annotator received one set to annotate. As mentioned before, attribute scoring was not performed on essays that do not have a thesis statement.

The resulting distributions of scores for Thesis Strength and the attributes are shown in Table 13.

4.3 Inter-Annotator Agreement

We measure inter-annotator agreement on the aforementioned 120 triply-annotated essays using Krippendorff's α (Krippendorff, 1980).

For thesis statement identification, we measure agreement at the sentence level (i.e., whether a sentence is correctly marked as ”thesis” or ”not thesis”). Agreement is substantial: α is .816.

Agreement results on Thesis Strength and attribute scoring are shown in Table 14. As we can see, all attributes exhibit an agreement above 0.5, showing a correlation much more significant than random chance. Thesis Strength has an agreement of 0.635, which suggests that it can be agreed upon in a reasonably general sense. The attributes that have the highest α values are Direction of Development (0.856), Justification of Opinion (0.787) and Clarity (0.748), whereas the ones that have the lowest α values are Specificity (0.530), Eloquence (0.532), and Relevance to Prompt (0.550).

Attribute	PC	p-value
Arguability	.134	.000
Specificity	.139	.000
Clarity	.187	.000
Relevance to Prompt	.712	.000
Conciseness	.017	.631
Eloquence	.045	.204
Confidence	.094	.008
Direction of Development	.123	.001
Justification of Opinion	.420	.000
Justification of Importance/Interest	.206	.000

Table 15: Pearson's Correlation of each attribute with Thesis Strength and the corresponding p-value.

To better understand these agreement numbers, recall that each attribute is scored on a scale of $1-3$, where '1' generally means ”no”, '2' generally means ”partially”, and '3' generally means ”yes”. We found that the ”mostly yes” and ”mostly no” cases are the most difficult for the annotators to agree on. Specifically, some annotators translate a ”mostly yes” to a '3' while others translate it to a '2', and similarly for a ”mostly no”. As a result, attributes that have fewer ambiguous cases (i.e., the '2' cases) tend to have a higher agreement.

4.4 Analysis of Annotations

In this subsection, we conduct several experiments in order to gain insights into our annotations.

Correlation between Thesis Strength and the attributes. To understand whether the 10 attributes we annotated are indeed useful for predicting Thesis Strength, we compute the Pearson's Correlation Coefficient (PC) between Thesis Strength and each of the attributes along with the corresponding p-values. Results are shown in Table 15. As hypothesized in Section 4.1, all attributes are positively correlated with Thesis Strength, even though two of the correlations (the ones concerning Eloquence and Conciseness) are statistically insignificant at the $p < 0.01$ level. Among the eight statistically significant correlations, we see that Relevance to Prompt is highly correlated with Thesis Strength, having a PC value of 0.712. Justification of Opinion, though having a PC value of only 0.420, has a higher correlation with Thesis Strength than any of the remaining six attributes. In fact, the remaining six attributes are all very weakly correlated with Thesis Strength, having PC values that fall roughly between 0.1 and 0.2.

Predicting Thesis Strength using gold attributes. Next, we conduct an oracle experiment to determine how well these 10 attributes, when used in combination, can explain Thesis Strength. Specifically, we train two models on the 793 essays that have a thesis statement to score a thesis statement's strength using the *gold* attributes as features. The first model is a linear SVM regressor trained using the scikit-learn package (Pedregosa et al., 2011) with default learning parameters except for C (the regularization parameter), which is tuned on development data using grid search. The second model is a neural network trained using Keras (Chollet et al., 2015). The network passes the attribute vector through two dense layers, one for reducing the vector's dimension to 150 and the other for scoring. It uses mean absolute error as the loss function, Leaky ReLU as the activation function, rmsprop as the optimizer, and early stopping with patience = 10.

We report results obtained using five-fold cross validation.[3] The SVM regressor yields a PC score (the Pearson Correlation between the system's predictions and the gold scores) of 0.758 and a ME score (the mean absolute distance between the system's prediction and the gold score) of 0.520, whereas the neural network yields a PC score of 0.749 and a ME score of 0.575. Since PC is a correlation metric, higher correlation implies better performance. In contrast, ME is an error metric, so lower scores imply better performance. The large PC values and the relatively small ME values that we obtained in these experiments provide suggestive evidence that these attributes, when used in combination, can largely explain the strength of a thesis statement.

Attribute importance. The previous experiment allows us to conclude that the 10 attributes, when used in combination, can largely explain Thesis Strength. The question, then, is: which attributes are more useful than the others in scoring thesis strength? To answer this question, we train a linear SVM regressor on the 793 essays that have a thesis statement and examine the feature weight learned by the regressor for each attribute, as an attribute with a higher absolute weight has a higher impact on thesis strength scoring.

The feature weights and the bias term are shown in Table 16. As we can see, the weight associated with Relevance to Prompt is the highest, followed by those of Justification of Opinion and Justification of Importance/Interest. This is perhaps not surprising, as these attributes have the highest correlation with Thesis Strength.

Attribute	Weight
Arguability	0.000590
Specificity	0.000099
Clarity	0.000134
Relevance to Prompt	0.999998
Conciseness	0.000268
Eloquence	0.000061
Confidence	0.000070
Direction of Development	0.000192
Justification of Opinion	0.499983
Justification of Importance/Interest	0.499935
(Bias)	−0.003770

Table 16: Feature weights on the attributes obtained by training a linear SVM on these attributes to predict Thesis Strength.

What is perhaps somewhat surprising is that many of the attributes that are believed to be relevant for thesis strength scoring, such as Arguability, Specificity, and Clarity, have negligibly small weights. We believe that these counter-intuitive results can be attributed to the score distributions of these attributes. Looking at Tables 13 and 16, we can see that the attributes that have low weights all have skewed distributions. For instance, only 37 of the 793 essays have an Arguability score of less than 3. Having distributions that are skewed towards one value, these attributes are unlikely to be useful for thesis strength scoring. In contrast, the attributes that have higher weights all have comparatively less skewed distributions. For instance, Relevance to Prompt has a score distribution that is the least skewed among the 10 attributes. This kind of distribution offers the learner an opportunity to learn how the different scores of an attribute correlate with the thesis strength score.

In addition, note that the 10 attributes we identified account for nearly all attributes impacting thesis strength, as unenumerated attributes cost essays an average of only four-thousandths of a point on the six-point thesis strength scale.

Correlation with other scoring dimensions. Recall that annotating the ICLE essays (as opposed to essays in other corpora) would allow us to study the interactions between Thesis Strength and other scoring dimensions. Our next experiment exploits this benefit. Specifically, we compute the PC values between our Thesis Strength scores and the scores annotated by Persing and his colleagues (2010; 2013; 2014; 2015) along four dimensions, namely Thesis Clarity (how clear is the thesis?), Organization (how well-organized is

[3] In all five-fold cross-validation experiments in this paper, we use three folds for training, one fold for development (parameter tuning), and one fold for testing.

Scoring Dimension	PC	p**-value**	% Essays
Thesis Clarity	.301	.000	829/830
Organization	.091	.004	1002/1003
Adherence to Prompt	.205	.000	829/830
Persuasiveness	.222	.000	999/1000

Table 17: Correlation of Thesis Strength with other essay scoring dimensions and the corresponding p-value.

the essay?), Adherence to Prompt (how relevant is the essay's content to the prompt?), and Argument Persuasiveness (how persuasive is the argument the essay makes for its thesis?).

Correlation results together with the corresponding p-values are shown in Table 17.[4] As we can see, all four correlations are relatively weak. The weak correlations are consistent with our intuition that these dimensions capture different aspects of essay quality. Among the four dimensions, Thesis Strength has the highest correlation with Thesis Clarity. This is not surprising, as a thesis is unlikely to be strong if it is not clear. A somewhat weaker correlation exists between Thesis Strength and Argument Persuasiveness. This is perhaps not surprising either. As mentioned above, Argument Persuasiveness scoring is partly based on the thesis. Intuitively, an argument is unlikely to be persuasive if the underlying thesis statement is weak, even though an unpersuasive argument does not necessarily imply a weak thesis statement. The remaining two dimensions, Organization and Adherence to Prompt, have very weak correlations with Thesis Strength. We speculate the low correlation has to do with the fact that both dimensions are scored based on the entire essay rather than just the thesis statement.

4.5 Additional Experiments

Next, we conduct preliminary experiments on thesis statement identification and thesis strength scoring to gauge the difficulty of these two tasks.

Thesis statement identification. We employ four systems, the first two of which are heuristic.
(1) *First Major Claim.* Given the close connection between a major claim and a thesis, we approximate the thesis statement identification task as a major claim identification problem. Specifically, we use Eger et al.'s (2017) argument mining system, which was trained on the Argument Annotated Essays Corpus mentioned in Section 2, for major claim identification, taking the first major claim identified in an essay as its thesis statement.
(2) *Keyword similarity.* Intuitively, sentences that resemble the prompt are more likely to be thesis sentences than those that do not. Hence, this system considers the k sentences that have the largest *keyword* overlap with the prompt as thesis sentences, where keywords are the important words in a prompt that we manually picked.
(3) *SVM.* As our first learning-based system, we employ a linear SVM as implemented in the scikit-learn package to train a classifier for determining whether a sentence is a thesis sentence or not, using the unigrams, bigrams, and trigrams extracted from the sentence as features. All parameters are set to their default values except for C, the regularization parameter, which is tuned to maximize F1 on development data using grid search.
(4) *Neural network.* Next, we train a neural network (NN) using Keras to determine whether a sentence is a thesis sentence or not. The NN takes as inputs the given sentence and the prompt for which the essay was written, each of which is represented as a sequence of 300-dimensional Facebook FastText pre-trained word embeddings (Bojanowski et al., 2017). (Out-of-vocabulary words are represented as zero vectors, and all inputs are padded to their maximum size by adding zeros to the end.) We employ two bidirectional LSTMs (Schuster and Paliwal, 1997) with an attention mechanism to create representations for the two input vectors. These two representations, together with their similarity (computed by taking their dot product), are concatenated. The resulting vector then goes through two dense layers, one for reducing the vector's dimension to 150 and the other for predicting whether the given sentence is a thesis sentence. The network uses categorical cross-entropy as the loss function, Leaky ReLU as the activation function (except for the output layer, which uses a softmax), rmsprop as the optimizer, and early stopping with patience = 10.

Table 18 shows the five-fold cross-validation results, which are expressed in terms of recall (R), precision (P), and F1 in identifying thesis sentences. While the SVM outperforms the other systems, its F1 score is only around 24%. These re-

[4] The last column of Table 17 shows the number of essays used to compute the PC value for each dimension. For instance, 829 of the 830 essays that Persing and Ng annotated with Thesis Clarity scores are annotated by us with Thesis Strength scores, and these 829 overlapping essays are used to compute the PC value between Thesis Clarity and Thesis Strength. Note that the percentage of overlap for each dimension is high, as we selected the essays to maximize the degree of overlap with those that Persing and Ng annotated.

System	R	P	F1
First Major Claim	.174	.182	.178
Keyword similarity (top 1)	.165	.239	.195
Keyword similarity (top 3)	.190	.097	.128
SVM	.245	.244	.245
NN	.222	.222	.222

Table 18: Thesis statement identification results.

sults suggest that thesis statement identification is a challenging task.[5]

Thesis strength scoring. We employ three simple systems, all of which are learning-based.

(1) *Gold thesis statements without attributes.* Given the difficulty of thesis statement identification, we conduct our thesis strength scoring experiments using gold thesis statements. (In other words, we exclude essays without thesis statements in these experiments.) Specifically, we train a model that takes a gold thesis statement as input and predicts its strength score.

(2) *Gold thesis statements with predicted attributes.* This is a pipeline system in which (1) 10 models are first trained to independently predict the scores of the 10 attributes given the gold thesis statement, and then (2) a second model is trained to predict the thesis strength score using the 10 predicted attributes.

(3) *Entire essay without attributes.* In this system, we train a model to predict thesis strength based on the *entire* essay. In other words, we use as input all the sentences in the essay.

For each of these systems, we employ (1) a linear SVM regressor and (2) a NN as the underlying model. Specifically, to train the first system, the SVM/NN we use is the same as that in the thesis statement identification experiment except that (1) the input is the gold thesis and (2) the output is the thesis strength score. To train the second system, the 10 SVMs/NNs in the first step of the pipeline are the same as that in the first system except that the output is the attribute score, and the SVM/NN in the second step of the pipeline is the same as that in the SVM/NN experiments in Section 4.4 except that we use predicted rather than perfect attribute values as inputs. To train the third system, the SVM/NN is the same as that in the first system except that the input is the entire essay.

Experiment	Model	PC	ME
Gold thesis without attributes	SVM	.390	.921
	NN	.310	.978
Gold thesis with predicted attributes	SVM	.353	.962
	NN	.256	.951
Entire essay without attributes	SVM	.065	1.14
	NN	.128	.964

Table 19: Thesis strength scoring results.

Five-fold cross-validation results, which are expressed in terms of PC and ME, are shown in Table 19. Consider first the two thesis-based experiments (rows 1 and 2). As we can see, although gold thesis statements are used, the results are not particularly strong, with PC values less than 0.4 and ME values greater than 0.9. Moreover, the results in the first experiment are generally better than those in the second experiment. This suggests that not only do the noisily predicted attributes fail to benefit thesis strength scoring, but they actually hurt thesis strength scoring.[6] Furthermore, though of lesser importance, SVM generally performs better than NN. These results contrast with those in the essay-based experiment (row 3), where NN performs better than SVM. As we can see, both essay-based models substantially underperform their thesis-based counterparts. This suggests that accurate thesis statement identification is crucial for accurate scoring of thesis strength.

5 Conclusion

Since progress in dimension-specific essay scoring research is hampered in part by the scarcity of annotated corpora, we designed rubrics for manually scoring 1021 essays along a fundamental yet unexplored dimension of essay quality, thesis strength, as well as the attributes that could impact strength. We chose to annotate the essays in ICLE that have previously been scored along multiple dimensions in order to facilitate future developments of joint models that can capture the interactions among different dimensions. We believe our annotations will be a valuable resource to the NLP community.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was supported in part by NSF Grants IIS-1528037 and CCF-1848608.

[5] Other approaches to thesis statement identification exist. For instance, Burstein et al. (2003) require that their model be trained on essays where *each* sentence is annotated with its discourse function (e.g., thesis, rebuttal, elaboration, conclusion). Given the lack of such annotations in our corpus, we do not use these systems in our experiments.

[6] PC values for attribute prediction are fairly low, ranging from 0.02 to 0.11.

References

Chris M. Anson and Robert A. Schwegler. 2011. *The Longman Handbook for Writers and Readers*, 6th edition. New York: Longman.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The Criterion online writing evaluation service. *AI Magazine*, 25(3):27–36.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56 Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

François Chollet et al. 2015. Keras. `https://keras.io`.

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

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 185–192.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2017. Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4130–4136.

Klaus Krippendorff. 1980. Validity in content analysis. In E. Mochmann, editor, *Computerstrategien fr die kommunikationsanalyse*, pages 69–112. Frankfurt, Germany: Campus.

Annie Louis and Derrick Higgins. 2010. Off-topic essay detection using short prompt texts. In *Proceedings of the Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–95.

Andrea A. Lunsford. 2015. *The St. Martin's Handbook*, 8th edition. Boston: Bedford/St. Martin's.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

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

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

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

John D. Ramage, John C. Bean, and June Johnson. 2018. *The Allyn & Bacon Guide to Writing*, 8th edition. New York: Pearson.

John J. Ruszkiewicz. 2010. *The Scott, Foresman Handbook for Writers*, 9th edition. New York: Longman.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

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

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578.