Applying Argumentation Schemes for Essay Scoring

Yi Song Michael Heilman Beata Beigman Klebanov Paul Deane
Educational Testing Service
Princeton, NJ, USA
{ysong, mheilman, bbeigmanklebanov, pdeane}@ets.org

Abstract

Under the framework of the argumentation scheme theory (Walton, 1996), we developed annotation protocols for an argumentative writing task to support identification and classification of the arguments being made in essays. Each annotation protocol defined argumentation schemes (i.e., reasoning patterns) in a given writing prompt and listed questions to help evaluate an argument based on these schemes, to make the argument structure in a text explicit and classifiable. We report findings based on an annotation of 600 essays. Most annotation categories were applied reliably by human annotators, and some categories significantly contributed to essay score. An NLP system to identify sentences containing scheme-relevant critical questions was developed based on the human annotations.

1. Introduction

In this paper, we analyze the structure of arguments as a first step in analyzing their quality. Argument structure plays a critical role in identifying relevant arguments based on their content, so it seems reasonable to focus first on identifying characteristic patterns of argumentation and the ways in which such arguments are typically developed when they are explicitly stated. It is worthwhile to classify the arguments in a text and to identify their structure when they are extended to include whole text segments (Walton, 1996; Walton, Reed, and Macagno, 2008), but it is not clear how far human annotation can go in analyzing argument structure.

An analysis of the effectiveness and full complexity of argument structure is different than the identification of generic elements that might compose an argument, such as claims (e.g., a thesis sentence), main reasons (e.g., supporting topic sentences), evidence (e.g., elaborating segments), and other components, such as the introduction and conclusion (Burstein, Kukich, Wolff, Lu, Chodorow, Braden-Harder, & Harris, 1998; Burstein, Marcu, and Knight, 2003; Pendar & Cotos, 2008). In contrast, here we focus on analyzing specific types of arguments, what the literature terms *argumentation schemes* (Walton, 1996). Argumentation schemes include schematic content and take into account a pattern of possible argumentation moves in a larger persuasive dialog. Understanding these argumentation schemes is important for understanding the logic behind an argument. *Critical questions* associated with a particular argumentation scheme provide a normative standard that can be used to evaluate the relevance of an argument's justificatory structure (van Eemeren and Grootendorst, 1992; Walton, 1996; Walton et al., 2008).

We aimed to lay foundations for the automated analysis of argumentation schemes, such as the identification and classification of the arguments in an essay. Specifically, we developed annotation protocols for writing prompts in an argument analysis task from a graduate school admissions test. The task was designed to assess how well a student analyzes someone else's argument, which is provided by the prompt. The student must critically evaluate the logical soundness of the given argument. The annotation categories were designed to map student responses to the scheme-relevant critical questions. We examined whether this approach provides a useful framework for describing argumentation and whether human annotators can apply it reliably and consistently. Furthermore, we have begun work on automating the annotation process by developing a system to predict whether sentences contain scheme-relevant critical questions.

2. Theoretical Framework

As Nussbaum (2011) notes, there have been critical advances in the study of informal argument,

Proceedings of the First Workshop on Argumentation Mining, pages 69–78,
Baltimore, Maryland USA, June 26, 2014.

which takes place within a social context involving dialog among people with different beliefs, most notably the development of theories that provide relatively rich schemata for classifying informal arguments, such as Walton (1996).

An argumentation scheme is defined as "a more or less conventionalized way of representing the relation between what is stated in the argument and what is stated in the standpoint" (van Eemeren and Grootendorst, 1992, p. 96). It is a strategic pattern of argumentation linking premises to a conclusion and illustrating how the conclusion is derived from the premises. This "internal structure" of argumentation reflects justificatory standards that can be used to help evaluate the reasonableness of an argument (van Eemeren and Grootendorst, 2004). Argumentation schemes should be distinguished from the kinds of structures postulated in Mann and Thompson's (1988) Rhetorical Structure Theory (RST) because they focus on relations inherent in the meaning of the argument, regardless of whether they are explicitly realized in the discourse.

Consider, for instance, *argument from consequences*, which applies when the primary claim argues for or against a proposed policy (i.e., course of action) by citing positive or negative consequences that would follow if the policy were adopted (Walton, 1996). Elaborations of an argument from consequences are designed to defend against possible objections. For instance, an opponent could claim that the claimed consequences are not probable; or that they are not desirable; or that they are less important than other, undesirable consequences. Thus a sophisticated writer, in elaborating an *argument from consequences*, may provide information to reinforce the idea that the argued consequences are probable, desirable, and more important than any possible undesired effects. These moves correspond to what the literature calls *critical questions*, which function as a standard for evaluating the reasonableness of an argument based on its argumentation schemes (Walton, 1996).

Walton and his colleagues (2008) analyzed over 60 argumentation schemes, and identified critical questions associated with certain schemes as the logical moves in argumentative discourse. The range of possible moves is quite large, especially when people use multiple schemes. There have been several efforts to annotate corpora with argumentation scheme information to support future machine learning efforts (Mochales and Ieven, 2009; Palau and Moens, 2009; Rienks, Heylen, and Van der Weijden, 2005; Verbree, Rienks, and Heylen, 2006), to support argument representation (Atkinson, Bench-Capon, and McBurney, 2006; Rahwan, Banihashemi, Reed, Walton, and Abdallah, 2010), and to teach argumentative writing (Ferretti, Lewis, and Andrews-Weckerly, 2009; Nussbaum and Schraw, 2007; Nussbaum and Edwards, 2011; Song and Ferretti, 2013). In addition, Feng and Hirsh (2011) used the argumentation schemes to reconstruct the implicit parts (i.e., unstated assumptions) of the argument structure. In many previous studies, the data sets on argumentation schemes were relatively small and the inter-rater agreement was not measured.

We are particularly interested in exploring the relationship between the use of scheme-relevant critical questions and essay quality, as measured by holistic essay scores. The difference between an expert and a novice is that the expert knows which critical questions should be asked when the dynamic of the argument requires them, while the novice misses the essential moves to ask critical questions that help evaluate if the argument is valid or reasonable. Often, students presume information and fail to ask questions that would reveal potential fallacies. For example, they might use quotations from books, arguments from TV programs, or opinions posted online without evaluating whether the information is adequately supported by evidence.

Critically evaluating arguments is considered an important skill in college and graduate school. For example, a widely accepted graduate admissions test has a task to assess students' critical thinking and analytical writing skills. In this argument analysis task, students should demonstrate skills in critiquing other people's arguments, such as identifying unwarranted assumptions or discussing what specific evidence is need to support the argument. They must communicate their evaluation of the arguments clearly to the audience. To accomplish this task successfully, students need to evaluate the arguments against appropriate criteria. Therefore, their essays could be analyzed using an annotation approach based on the theory of argumentation schemes and critical questions.

Our research questions were as follows:

1. Can this scheme-based annotation approach be applied consistently by annotators to a corpus of argumentative essays?
2. Do annotation categories based on the theory of argumentation schemes contribute

significantly to the prediction of essay scores?

3. Can we use NLP techniques to train an automated classifier for distinguishing sentences that raise critical questions from sentences that contain no critical questions?

3 Development of Annotation Protocols

Although Walton's argumentation schemes provided a good framework for analyzing arguments, it was challenging to apply them in some cases of argument essays because various interpretations could be made on some argument structures. For instance, people were often confused with *argument from consequences*, *argument from correlation to cause*, and *argument from cause to effect* because all these three types of arguments indicate a causal relationship. While it is good that Walton tried to identify variations of a causal relationship, a side effect is that some schemes are not so distinguishable from each other, especially for someone who is not an expert in logic. This ambiguity makes it difficult to apply his theory directly to annotation. Thus, we modified Walton's schemes and created new schemes when necessary to achieve exclusive annotation categories and capture the features in the argument analysis task.

In this paper, we illustrate our annotation protocols on a policy argument because over half of the argument analysis prompts for the assessment we are working with deal with policy issues (i.e., issues involve the possibility of putting a practice into place). Here, we use the "Patriot Car" prompt as an example.

> The following appeared in a memorandum from the new president of the Patriot car manufacturing company.
>
> "In the past, the body styles of Patriot cars have been old-fashioned, and our cars have not sold as well as have our competitors' cars. But now, since many regions in this country report rapid increases in the numbers of newly licensed drivers, we should be able to increase our share of the market by selling cars to this growing population. Thus, we should discontinue our oldest models and concentrate instead on manufacturing sporty cars. We can also improve the success of our marketing campaigns by switching our advertising to the Youth Advertising agency, which has successfully promoted the country's leading soft drink."

Test takers are asked to analyze the reasoning in the argument, consider any assumptions, and discuss how well any evidence that is mentioned supports the conclusion.

The prompt states that the new president of the Patriot car manufacturing company pointed out a problem that the body styles of Patriot cars have been old-fashioned and their cars have not sold as well as their competitors' cars. The president proposed a plan to discontinue their oldest models and to concentrate on manufacturing sporty cars. He believed that this plan will lead to an increase in their market share (i.e., the goal). This is a policy issue because it involves whether the plan of discontinuing oldest car models and manufacturing sporty cars should be put into place. This prompt shows a typical pattern of many argument analysis prompts about policy issues: (1) a problem is stated; (2) a plan is proposed; and (3) a desirable goal will be achieved if the plan is implemented. Thus, we created a *policy* scheme that includes these three major components (i.e., problem, plan, and goal), and a causal relationship that bridges the plan to the goal in the policy scheme. Therefore, a *causal* scheme appears in a policy argument to represent the causal relationship from the proposed plan to the goal. This part is different from Walton's analysis. He uses the *argument from consequences* scheme for policy arguments, but it created confusions when applying it to annotation, especially when students unconsciously use the word "cause" to introduce a potential consequence that follows a policy. In addition, our *causal* scheme combines the argument from *correlation to cause* scheme and the argument from *cause to effect* scheme specified by Walton.

Accordingly, we revised or re-arranged some of the critical questions in Walton's theory. For example, challenges to arguments that use a *policy* scheme fall into the following six categories: (a) problem; (b) goal; (c) plan implementation; (d) plan definition; (e) side effect; and (f) alternative plan. When someone writes that the president should re-evaluate whether this is really a problem, it matches the question in the "problem" category; when someone questions if there is an alternative plan that could also help achieve the goal and is better than the plan proposed by the president, it should be categorized as a challenge in "alternative plan." We call these "specific questions" because they are attached to a par-

Scheme	Category	Critical Question
Policy	Problem	Is this really a problem? Is the problem well-defined?
	Goal	How desirable is this goal? Are there specific conflicting goals we do not wish to sacrifice?
	Plan Implementation	Is it practically possible to carry out this plan?
	Plan Definition	Is the plan well defined?
	Side Effects	Are there negative side effects that should be taken into account if we carry out our plan?
	Alternative plan	Are there better alternatives that could achieve the goal?
Causal	Causal Mechanism	Is there really a correlation? Is the correlation merely a coincidence (invalid causal relationship)? Are there alternative causal factors?
	Causal Efficacy	Is the causal mechanism strong enough to produce the desired effects?
	Applicability	Does this causal mechanism apply?
	Intervening Factors	Are there intervening factors that could undermine the causal mechanism?
Sample	Significance	Are the patterns we see in the sample clear-cut enough (and in the right direction) to support the desired inference?
	Representativeness	Is there any reason to think that this sample might not be representative of the group about which we wish to make an inference?
	Stability	Is there any reason to think this pattern will be stable across all the circumstances about which we wish to make an inference?
	Sample Size	Is there any reason to think that the sample may not be large enough and reliable enough to support the inference we wish to draw?
	Validity	Is the sample measured in a way that will give valid information on the population attributes about which we wish to make inferences?
	Alternatives	Are there external considerations that could invalidate the claims?

Table 1: Annotation protocols for three types of argumentation schemes

ticular prompt. In other words, specific questions are content dependent. Each category also includes one or more "general questions" that can be asked for any argument using the same argumentation scheme, and in this case, it is the *policy* scheme.

We have developed annotation protocols for various argumentation schemes. Table 1 includes part of the annotation protocols (i.e., scheme, category, and general critical questions) for three argumentation schemes: the *policy* argument scheme, the *causal argument* scheme, and the *argument from a sample* scheme. This study focuses on these three argumentation schemes and 16 associated categories.

4 Application of the Annotation Approach

This section focuses on applying the annotation approach and the following research question: Can this scheme-based annotation approach be applied consistently by raters to a corpus of argumentative essays?

4.1 Annotation Rules

The first step of the annotation is reading the entire essay. It is important to understand the writer's major arguments and the organization of the essay. Next, the annotator will identify and highlight any text segment (e.g., paragraph, sentence, or clause) that addresses a critical question. Usually, the minimal text segment is at the sentence-level, but it could be the case that the selection is at the phrase-level when a sentence includes multiple points that match more than one critical question. Thirdly, for a highlighted unit, the annotator will choose a topic, a category, and a second topic, if applicable. Only one category label can be assigned to each selected text unit.

"Generic" information will not be selected or assigned an annotation label. Generic information includes restatements of the text in the prompt, general statements that do not address any specific questions, rhetoric attacks, and irrelevant information. Note that this notion of generic information is related to "shell language," as described by Madnani et al (2012). However, our definition here focuses more closely on sentences that do not raise critical questions. Surface errors (e.g., grammar and spelling) can be

ignored if they do not prevent people from understanding the meaning of the essay. Here is an example of annotated text.

> As stated by the president, there is a rapid increase in the number of newly licensed drivers which would be a marketable target. [However, there was no concrete evidence that these newly licensed drivers favored sporty cars over other model types.]$_{\text{Causal Applicability}}$ [On a similar note, there was no anecdotal evidence demonstrating that lack of sales was contributed to the old-fashion body styles of the Patriot cars.]$_{\text{Causal Mechanism}}$ [There could be numerous other factors contributing to their lack of sales: prices are not competitive, safety ratings are not as high, features are not as appealing. The best way to tackle this problem is to send out researches and surveys to get the opinions of consumers.]$_{\text{Causal Mechanism}}$

4.2 Annotation Tool

The annotation interface includes the following elements:

1. the original writing prompt;
2. topics that the prompt addresses;
3. categories associated with critical questions relevant to that type of argument;
4. general critical questions that can be used across prompts that possess the same argumentation scheme; and
5. specific critical questions for this particular prompt.

The annotators highlight text segments to be annotated and then clicked a button to choose a topic (e.g., body style versus advertising agency in the Patriot Car prompt) and a category to identify which critical questions were addressed.

4.3 Data and Annotation Procedures

In this section, we report our annotation on two selected argument analysis prompts in an assessment for graduate school admissions. The actual prompts are not included here because they may be used in future tests. Both prompts deal with policy issues and are involved in causal reasoning, but the second prompt also has a *sample* scheme (see Table 1). For each prompt, we randomly selected 300 essays to annotate. These essays were written between 2008 and 2010.

Four annotators with linguistics backgrounds who were not co-authors of the paper received training on the annotation approach. Training focused on the application to specific prompts because each prompt had a specific annotation protocol that covers the argumentation schemes and how they relate to the prompt's topics. The first author delivered the training sessions, and helped resolve differences of opinion during practice annotation rounds. After training and practice, the annotators annotated 20 pilot essays for a selected prompt to test their agreement. This pilot stage gave us another chance to find and clarify any confusion about the annotation categories. After that, the annotators worked on the sampled set of 300 essays, and these annotations were then used for analyses. For each prompt, 40 essays were randomly selected, and all 4 annotators annotated these 40 essays to check the inter-annotator agreement. For the experiments described later that involve the multiply-annotated set, we used the annotations from the annotator who seemed most consistent.

4.4 Inter-Annotator Agreement

To compute human-human agreement, we automatically split the essays into sentences. For each sentence, we computed the annotations that overlapped with at least part of the tence. Then, for each category, we computed human-human agreement across all sentences about whether that category should be marked or not. We also created a "Generic" label, as discussed in section 4.1, for sentences that were not marked by any of the other labels.

We computed two inter-annotator agreement statistics. Our primary statistic is Cohen's *kappa* between pairs of raters. Four annotators generated 6 pairs of *kappa* values, and in this report we only report the average *kappa* value for each annotation category. As an alternative statistic, we computed Krippendorff's *alpha*, a chance-corrected statistic for calculating the inter-annotator agreement between multiple coders (four annotators in our case), which is similar to multi *kappa* (Krippendorff, 1980).

Table 2 shows the *kappa* and *alpha* values for each annotation category, excluding those that were rare. To identify rare categories, we averaged the numbers of sentences annotated under a category among four annotators, which indicated how many sentences were annotated under this category in 40 essays. If the number was lower than 10, which means that no more than one sentence was annotated in every four essays, then

the category was considered rare. Most rare categories had low inter-rater agreement, which is not surprising. It is not realistic to require annotators to always agree about rare categories.

From Table 2, we can see that the *kappa* value and the alpha value on the same category were close. The inter-annotator agreement on the "generic" category varied little across the two prompts (*kappa*: 0.572-0.604; *alpha*: 0.571-0.603), which indicates that the annotators had a fairly good agreement on this category. The annotators had good agreements on most of the commonly used categories (*kappa* ranged from 0.549 to 0.848, and *alpha* ranged from 0.537 to 0.843) except the "plan definition" under the *policy* scheme in prompt B (both *kappa* and alpha *values* were below 0.400). The major reason for this disagreement is that one annotator marked a significantly higher number of sentences (more than double) for this category than others did.

Prompt	Category	*Kappa*	*Alpha*
Prompt A			
	Generic	0.572	0.571
	Policy : Problem	0.644	0.640
	Policy : Side Effects	0.612	0.609
	Policy : Alternative Plan	0.665	0.666
	Causal : Causal Mechanism	0.680	0.676
	Causal : Applicability	0.557	0.555
Prompt B			
	Generic	0.604	0.603
	Policy : Problem	0.848	0.843
	Policy : Plan Definition	0.346	0.327
	Causal : Causal Mechanism	0.620	0.622
	Causal : Applicability	0.767	0.769
	Sample : Validity	0.549	0.537

Table 2: Inter-annotator agreement

5 Essay Score and Annotation Features

This section explores the second research question: Do annotation categories based on the theory of argumentation schemes contribute significantly to the prediction of essay scores? Answering this question would tell us whether we capture an important construct of the argument analysis task by recognizing these argumentation features. Specifically, we tested whether these features add predictive value to a model based the state-of-the-art e-rater essay scoring system (Burstein, Tetreault, and Madnani, 2013).

To explore the relationship between annotation categories and essay quality, we ran a multiple regression analysis for each prompt. Essay quality was the dependent variable and was measured by a final human score, on a scale from 0 to 6. The independent variables were nine high-level e-rater features and the annotation categories relevant to a prompt (Prompt A: 10 categories; Prompt B 16 categories). The e-rater features were designed to measure different aspects of writing (grammar, mechanics, style, usage, word choice, word length, sentence variety, development, and organization). We computed the percentage of sentences that were marked as belonging to each category (i.e., the number of sentences in a category divided by the total number of sentences) to factor out essay length.

Note that the generic category was negatively correlated with the essay score in both prompts, since it included responses judged irrelevant to the scheme-relevant critical questions. In other words, the generic responses are the parts of the text that do not present specific critical evaluations of the arguments in a given prompt. For the purposes of our evaluation, we used the inverse feature labeled "all critical questions": the proportion of the text that actually raises some critical question (i.e., is not generic), regardless of scheme. We believe this formulation more transparently expresses the underlying mechanism relating the feature to essay quality.

For each prompt, we split the 300 essays into two data sets: the training set and the testing set. The testing set had the 40 essays that were annotated by all four annotators, and the training set had the remaining 260. We trained three models with stepwise regression on the training set and evaluated them on the testing set:

1. A model that included only the e-rater features to examine how well the e-rater model works ("baseline")
2. A model with the baseline features and all the annotation category percentage variables except for the "generic" category variable ("baseline + categories")
3. A model with the baseline features and a feature corresponding to the inverse of the "generic" category ("baseline + all critical questions").

Table 3 presents the Pearson correlation coefficient *r* values for comparing model predictions

to human scores for each of the models. In prompt A, three annotation categories (causal mechanism, applicability, and alternative plan) were selected by the stepwise regression because they significantly contributed to the essay score above the nine e-rater features. This model showed higher test set correlations than the baseline model ($\Delta r = .014$). The model with the general argument feature ("all critical questions") showed a similar increase ($\Delta r = .014$).

	Training Set *r*	Testing Set *r*	Testing Set Δr
Prompt A			
baseline	.838	.852	---
baseline + specific categories	.852	.866	.014
baseline + all critical questions	.858	.866	.014
Prompt B			
baseline	.818	.761	---
baseline + specific categories	.835	.817	.056
baseline + all critical questions	.845	.821	.060

Table 3: Performance of essay scoring models with and without argumentation features

Similar observations apply to prompt B. The causal mechanism category added prediction significantly above e-rater with an increase ($\Delta r = .056$). The model containing the general argument feature ("all critical questions") performed slightly better ($\Delta r = .060$).

These results suggest that annotation categories based on argumentation schemes contribute additional useful information about essay quality to a strong baseline essay scoring model. In the next section, we report on preliminary experiments testing whether these annotations can be automated, which would almost certainly be necessary for practical applications.

6 Argumentation Schemes NLP System

We developed an NLP system for automatically identifying the presence of scheme-relevant critical questions in essays, and we evaluated this system with annotated data from the two selected argument prompts. This addresses the third research question: Can we use NLP techniques to train an automated classifier for distinguishing sentences that raise critical questions from sentences that contain no critical questions?

6.1 Modeling

In this initial development of the NLP system, we focused on the task of predicting whether a sentence raises any critical questions or none (i.e., generic vs. nongeneric). As such, the task was binary classification at the level of the sentence. The system we developed uses the SKLL tool[1] to fit L2-penalized logistic regression models with the following features:

- Word *n*-grams: Binary indicators for the presence of contiguous subsequences of *n* words in the sentence. The value of *n* ranged from 1 to 3. These features had value 1 if a particular *n*-gram was present in a sentence and 0 otherwise.
- word *n*-grams of the previous and next sentences: These are analogous to the word *n*-gram features for the current sentence.
- sentence length bins: Binary indicators for whether the sentence is longer than 2*t* word tokens, where *t* ranges from 1 to 10.
- sentence position: The sentence number divided by the number of sentences in text.
- part of speech tags: Binary indicators for the presence of words with various parts of speech, as predicted by NLTK 2.0.4.
- prompt overlap: Three features based on lexical overlap between the sentence and the prompt for the essay: a) the Jaccard similarity between the sets of word *n*-grams in the sentence and prompt ($n = 1, 2, 3$), b) the Jaccard similarity between the sets of word unigrams (i.e., just $n = 1$) in the sentence and prompt, and c) the Jaccard similarity between the sets of "content" word unigrams in the sentence and prompt (for this, content words were defined as word tokens that contained only numbers and letters and did not appear in NLTK's English stopword list).

6.2 Experiments

For these experiments, we used the training and testing sets described in Section 5. We trained models on the training data for each prompt individually and on the combination of the training data for both prompts. To measure generalization across prompts, we tested these models on the testing data for each prompt and on the combina-

[1] https://github.com/EducationalTestingService/skll

tion of the testing data for the two prompts. We evaluated performance in terms of unweighted Cohen's *kappa*. The results are in Table 4.

Training	Testing	*Kappa*
combined	combined	.438
Prompt A		.350
Prompt B		.346
combined	Prompt A	.379
Prompt A		.410
Prompt B		.217
combined	Prompt B	.498
Prompt A		.285
Prompt B		.478

Table 4: Performance of the NLP Model

The model trained on data from both prompts performed relatively well compared to the other models. For the testing data for prompt B, the combined model outperformed the model trained on just data from prompt B. However, the prompt-specific model for prompt A slightly outperformed the combined model on the testing data for prompt A.

Although the performance of models trained with data from one prompt and tested with data from another prompt did not perform as well, there is evidence of some generalization across prompts. The model trained on data from prompt B and tested on data from prompt A had *kappa* = 0.217; the model trained on data from prompt A and tested on data from prompt B had *kappa* = 0.285. Of course, these human-machine agreement values were somewhat lower than human-human agreement values (0.572 and 0.604, respectively), leaving substantial room for improvement in future work.

We also examined the most strongly weighted features in the combined model. We observed that multiple hedge words (e.g., "perhaps", "may") had positive weights, which associated with the "generic" class. We also observed that words related to argumentation (e.g., "conclusions", "questions") had negative weights, which associated them with the nongeneric class, as one would expect. One issue of concern is that some words related to the specific topics discussed in the prompts received high weights as well, which may limit generalizability.

7 Conclusion

Our research focused on identification and classification of argumentation schemes in argumentative text. We developed annotation protocols that capture various argumentation schemes. The annotation categories corresponded to scheme-relevant critical questions, and for text segments that do not contain any critical questions, we assigned a "generic" category. In this paper, we reported the results based on an annotation of a large pool of student essays (both high-quality and low-quality essays). Results showed that most of the common annotation categories (e.g. causal mechanism, alternative plan) can be applied reliably by the four annotators.

However, the annotation work is labor-intensive. People need to receive sufficient training to apply the approach consistently. They must not only identify meaningful chunks of textual information but also assign the right annotation category label for the selected text. Despite these complexities, it is a worthwhile investigation. Developing a systematic classification of argument structures not only plays a critical role in this project, but also has a potential contribution to other assessments on argumentation skills aligned with the Common Core State Standards. This work would help improve the current automated scoring techniques for argumentative essays because this annotation approach takes into account the argument structure and its content.

We ran regression analyses and found that manual annotations grounded in the argumentation schemes theory predict essay quality. Our data showed that features based on manual argument scheme annotations significantly contributed to models of essay scores for both prompts. This is probably because our approach focused on the core of argumentation, rather than surface or word-level features (e.g., mechanics, grammar, usage, style, essay organization, and vocabulary) examined by the baseline model.

Furthermore, we have implemented an automated system for predicting the human annotations. This system focused only on predicting whether or not a sentence raises any critical questions (i.e., generic vs. nongeneric). In the future, we plan to test whether features based on automated annotations make contributions to essay scoring models that are similar to the contributions of manual annotations. We also plan to work on detecting specific critical questions and adding additional features, such as features from Feng and Hirst (2011).

Acknowledgements

We would like to thank Keelan Evanini, Jill Burstein, Aoife Cahill, and the anonymous reviewers of this paper for their helpful comments. We would also like to thank Michael Flor for helping set up the annotation interface, and Melissa Lopez, Matthew Mulholland, Patrick Houghton, and Laura Ridolfi for annotating the data.

References

Katie Atkinson, Trevor Bench-Capon, and Peter McBurney. 2006. Computational representation of practical argument. *Synthese, 152*: 157-206. Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. "Automated scoring using a hybrid feature identification technique." In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp. 206-210. Association for Computational Linguistics.

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

Jill Burstein, Joel Tetreault, and Nitin Madnani. 2013. The e-rater automated essay scoring system. In Sermis, M. D. and Burstein, J. (eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 55-67). New York: Routledge.

Vanessa W. Feng and Graeme Hirst. 2011. Classifying arguments by scheme. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR.

Ralph P. Ferretti, William E. Lewis, and Scott Andrews-Weckerly. 2009. Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology, 101*: 577-589.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA : Sage Publications.Mann, William C., and Sandra A. Thompson. 1988. "Rhetorical structure theory: Toward a functional theory of text organization." *Text* 8(3): 243-281.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying High Level Organizational Elements in Argumentative Discourse. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (pp. 20-28). Association for Computational Linguistics.

Raquel Mochales and Asgje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments?: a case study on the legal argumentation of the ECHR. In ICAIL '09: Proceedings of the 12th International Conference on Artificial Intelligence and Law.

Michael Nussbaum. 2011. Argumentation, dialogue theory, and probability modeling: alternative frameworks for argumentation research in education. *Educational Psychologist, 46*: 84-106.

Nussbaum, E. M. and Edwards, O.V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of the Learning Sciences, 20*, 443-488.

Palau, R.M. and Moens, M. F. 2009. Automatic argument detection and its role in law and the semantic web. In Proceedings of the 2009 conference on law, ontologies and the semantic web. IOS Press, Amsterdam, The Netherlands.Pendar, Nick, and Elena Cotos. 2008. "Automatic identification of discourse moves in scientific article introductions." In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 62-70. Association for Computational Linguistics.

Rahwan, I., Banihashemi, B., Reed, C. Walton, D., and Abdallah, S. (2010). Representing and classifying arguments on the semantic web. *The Knowledge Engineering Review*.

Rienks, R., Heylen, D., and Van der Weijden, E. 2005. Argument diagramming of meeting conversations. In A. Vinciarelli, J. Odobez (Ed.), Proceedings of Multimodal Multiparty Meeting Processing, Workshop at the 7th International Conference on Multimodal Interfaces (pp. 85–92). Trento, Italy.

Yi Song and Ralph P. Ferretti. 2013. Teaching critical questions about argumentation through the revising process: Effects of strategy instruction on college students' argumentative essays. *Reading and Writing: An Interdisciplinary Journal, 26*(1): 67-90.

Stephen E. Toulmin. 1958. *The uses of argument*. Cambridge University Press, Cambridge, UK.

Frans H. van Eemeren and Rob Grootendorst. 1992. *Argumentation, communication, and fallacies: A pragma-dialectical perspective.* Mahwah, NJ: Erlbaum.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A systematic theory of argumentation: A pragma-dialectical approach*. Cambridge, UK: Cambridge University Press.

Verbree, D., Rienks, H., and Heylen, D. (2006). First Steps Towards the Automatic Construction of Argument-Diagrams from Real Discussions. In Pro-

ceedings of the 2006 conference on Computational Models of Argument: Proceedings of COMMA 2006. IOS Press, Amsterdam, The Netherlands.

Douglas N. Walton. 1996. *Argumentation schemes for presumptive reasoning.* Mahwah, NJ: Lawrence Erlbaum.

Douglas N. Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation schemes.* New York, NY: Cambridge University Press.