Argumentation: Content, Structure, and Relationship with Essay Quality

Beata Beigman Klebanov[1], **Christian Stab**[2], **Jill Burstein**[1],**Yi Song**[1],
Binod Gyawali[1], **Iryna Gurevych**[2,3]
[1] Educational Testing Service
660 Rosedale Rd, Princeton NJ, 08541, USA
{bbeigmanklebanov,jburstein,ysong,bgyawali}@ets.org
[2]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[3]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract

In this paper, we investigate the relationship between argumentation structures and (a) argument content, and (b) the holistic quality of an argumentative essay. Our results suggest that structure-based approaches hold promise for automated evaluation of argumentative writing.

1 Introduction

With the advent of the Common Core Standards for Education,[1] argumentation, and, more specifically, argumentative writing, is receiving increased attention, along with a demand for argumentation-aware *Automated Writing Evaluation* (AWE) systems. However, current AWE systems typically do not consider argumentation (Lim and Kahng, 2012), and employ features that address grammar, mechanics, discourse structure, syntactic and lexical richness (Burstein et al., 2013). Developments in *Computational Argumentation* (**CA**) could bridge this gap.

Recently, progress has been made towards a more detailed understanding of argumentation in essays (Song et al., 2014; Stab and Gurevych, 2014; Persing and Ng, 2015; Ong et al., 2014). An important distinction emerging from the relevant work is that between argumentative *structure* and argumentative *content*. Facility with the argumentation structure underlies the contrast between (1) and (2) below: In (1), claims are made without support; relationships between claims are not explicit; there is intervening irrelevant material. In (2), the argumentative structure is clear – there is a critical claim supported by a specific reason. Yet, is it in fact a good argument? When choosing a provider for trash collection, how relevant is the color of the trucks? In contrast, in (3) the argumentative structure is not very explicit, yet the argument itself, if the reader is willing to engage, is actually more pertinent to the case, content-wise. Example (4) has both the structure and the content.

(1) "*The mayor is stupid. People should not have voted for him. His policy will fail. The new provider uses ugly trucks.*"

(2) "*The mayor's policy of switching to a new trash collector service is flawed because he failed to consider the ugly color of the trucks used by the new provider.*"

(3) "*The mayor is stupid. The switch is a bad policy. The new collector uses old and polluting trucks.*"

(4) "*The mayor's policy of switching to a new trash collector service is flawed because he failed to consider the negative environmental effect of the old and air-polluting trucks used by the new provider.*"

Song et al. (2014) took the content approach, annotating essays for arguments that are pertinent to the argumentation scheme (Walton et al., 2008; Walton, 1996) presented in the prompt. Thus, a critique raising undesirable side effects (examples 3 and 4) is appropriate for a prompt where a policy is proposed, while the critique in (1) and (2) is not. The authors show, using the annotations, that raising pertinent critiques correlates with holistic essay scores. They build a content-heavy automated model; the model, however, does not generalize

[1]www.corestandards.org

Proceedings of the 3rd Workshop on Argument Mining, pages 70–75,
Berlin, Germany, August 7-12, 2016.

well across prompts, since different prompts use different argumentation schemes and contexts.

We take the structure-based approach that is independent of particular content and thus has better generalization potential. We study its relationship with the content-based approach and with overall essay quality. Our contributions are the answers to the following research questions:

1. whether the use of good argumentation structure correlates with essay quality;

2. while structure and content are conceptually distinct, they might in reality go together. We therefore evaluate the ability of the structure-based system to deal with content-based annotations of argumentation.

2 Related Work

Existing work in CA focuses on argumentation mining in various genres. Moens et al. (2007) identify argumentative sentences in newspapers, parliamentary records, court reports and online discussions. Mochales-Palau and Moens (2009) identify argumentation structures including claims and premises in court cases. Other approaches focus on online comments and recognize argument components (Habernal and Gurevych, 2015), justifications (Biran and Rambow, 2011) or different types of claims (Kwon et al., 2007). Work in the context of the IBM Debater project deals with identifying claims and evidence in Wikipedia articles (Rinott et al., 2015; Aharoni et al., 2014).

Peldszus and Stede (2015) identify argumentation structures in microtexts (similar to essays). They rely on several base classifiers and minimum spanning trees to recognize argumentative tree structures. Stab and Gurevych (2016) extract argument structures from essays by recognizing argument components and jointly modeling their types and relations between them. Both approaches focus on the structure and neglect the content of arguments. Persing and Ng (2015) annotate argument strength, which is related to content, yet what it is that makes an argument strong has not been made explicit in the rubric and the annotations are essay-level. Song et al. (2014) follow the content-based approach, annotating essay sentences for raising topic-specific critical questions (Walton et al., 2008).

Ong et al. (2014) report on correlations between argument component types and holistic essay scores. They report that rule-based approaches for identifying argument components can be effective for ranking but not rating. However, they used a very small data set. In contrast, we study the relationship between content-based and structure-based approaches and investigate whether argumentation structures correlate with holistic quality of essays using a large public data set.

In the literature on the development of argumentation skills, an emphasis is made on both the structure, namely, the need to support one's position with reasons and evidence (Ferretti et al., 2000), and on the content, namely, on evaluating the effectiveness of arguments. For example, in a study by Goldstein et al. (2009), middle-schoolers compared more and less effective rebuttals to the same original argument.

3 Argumentation Structure Parser

For identifying argumentation structures in essays, we employ the system by Stab and Gurevych (2016) as an off-the-shelf argument structure parser. The parser performs the following steps:
Segmentation: Separates argumentative from non-argumentative text units; identifies the boundaries of argument components at token-level.
Classification: Classifies each argument component as Claim, Major Claim or Premise.
Linking: Identifies links between argument components by classifying ordered pairs of components in the same paragraph as either linked or not.
Tree generation: Finds tree structures (or forests) in each paragraph which optimize the results of the the previous analysis steps.
Stance recognition: Classifies each argument component as either *for* or *against* in order to discriminate between supporting or opposing argument components and argumentative support and attack relations respectively.

4 Experiment 1: Content vs Structure

4.1 Data

We use data from Song et al. (2014) – essays written for a college-level exam requiring test-takers to criticize an argument presented in the prompt. Each sentence in each essay is classified as generic (does not raise a critical question appropriate for the argument in the prompt) or non-generic (raises an apt critical question); about 40% of sentences are non-generic. Data sizes are shown in Table 1.

Prompt	Train			Test	
	#Es-says	#Sen-tences	Non-generic	#Es-says	#Sen-tences
A	260	4,431	42%	40	758
B	260	4,976	41%	40	758

Table 1: Data description for Experiment 1.

4.2 Selection of Structural Elements

We use the training data to gain a better understanding of the relationship between structural and content aspects of argumentation. Each selection is evaluated using kappa against Song et al. (2014) generic vs non-generic annotation.

Our first hypothesis is that any sentence where the parser detected an argument component (any claim or premise) could contain argument-relevant (non-generic) content. This approach yields kappa of 0.24 (prompt **A**) and 0.23 (prompt **B**).

We observed that the linking step in the parser's output identified many cases of singleton claims – namely, claims not supported by an elaboration. For example, "*The county is following wrong assumptions in the attempt to improve safety*" is an isolated claim. This sentence is classified as "generic", since no specific scheme-related critique is being raised. Removing unsupported claims yields kappas of 0.28 (A) and 0.26 (B).

Next, we observed that even sentences that contain claims that are supported are often treated as "generic". Test-takers often precede a specific critique with one or more claims that set the stage for the main critique. For example, in the following 3-sentence sequence, only the last is marked as raising a critical question: "*If this claim is valid we would need to know the numbers. The whole argument in contingent on the reported accidents. Less reported accidents does not mean less accidents.*" The parser classified these as Major Claim, Claim, and Premise, respectively. Our next hypothesis is that it is the premises, rather than the claims, that are likely to contain specific argumentative content. We predict that only sentences containing a premise would be "non-generic." This yields a substantial improvement in agreement, reaching kappas of 0.34 (A) and 0.33 (B).

Looking at the overall pattern of structure-based vs content-based predictions, we note that the structure-based prediction over-generates: The ratio of false-positives to false-negatives is 2.9 (A) and 3.1 (B). That is, argumentative structure without argumentative content is about 3 times more common than the reverse. False positives include sentences that are too general ("*Numbers are needed to compare the history of the roads*") as well as sentences that have an argumentative form, but fail to make a germane argument ("*If accidents are happening near a known bar, drivers might be under the influence of alcohol*").

Out of all the false-negatives, 30% were cases where the argument parser predicted no argumentative structures at all (no claims of any type and no premises). Such sentences might not have a clear argumentative form but are understood as making a critique in the context. For example, "*What was it 3 or 4 years ago?*" and "*Has the population gone up or down?*" look like fact-seeking questions in terms of structure, but are interpreted in the context as questioning the causal mechanism presented in the prompt. Overall, in 9% of all non-generic sentences the argument parser detected no claims or premises.

4.3 Evaluation

Table 2 shows the evaluation of the structure-based predictions (classifying all sentences with a Premise as non-generic) on test data, in comparison with the published results of Song et al. (2014), who used content-heavy features (such as word ngrams in the current, preceding, and subsequent sentence). The results clearly show that while the structure-based prediction is inferior to content-based one when the test data are essays responding to the same prompt as the training data, the off-the-shelf structure-based prediction is on-par with content-based prediction on the cross-prompt evaluation. Thus, when the content is expected to shift, falling back to structure-based prediction is potentially a reasonable strategy.

System	Train	Test	κ
Song et al. (2014)	A	A	.410
Song et al. (2014)	B	B	.478
Song et al. (2014)	A	B	.285
Song et al. (2014)	B	A	.217
Structure-based (Premises)	–	A	.265
Structure-based (Premises)	–	B	.247

Table 2: Evaluation of content-based (Song et al., 2014) and structure-based prediction on content-based annotations.

5 Experiment 2: Argumentation Structure and Essay Quality

Using argumentation structure and putting forward a germane argument are distinct, not only theoretically, but also empirically, as suggested by the results of Experiment 1. In this section, we evaluate to what extent the use of argumentation structures correlates with overall essay quality.

5.1 Data

We use a publicly available set of essays written for the TOEFL test in an argue-for-an-opinion-on-an-issue genre (Blanchard et al., 2013). Although this data was originally used for natural language identification experiments, coarse-grained holistic scores (3-grade scale) are provided as part of the LDC distribution. Essays were written by non-native speakers of English; we believe this increases the likelihood that fluency with argumentation structures would be predictive of the score. We sampled 6,074 essays for training and 2,023 for testing, both across 8 different prompts. In terms of distribution of holistic scores in the training data, 54.5% received the middle score, 11% – the low score, and 34.5% – the high score.

5.2 Features for essay scoring

Our set of features has the following essay-level aggregates: the numbers of any argument components, major claims, claims, premises, supporting and attacking premises, arguments against, arguments for, and the average number of premises per claim. Using the training data, we found that 90% Winsorization followed by a log transformation improved the correlation with scores for all features. The correlations range from 0.08 (major claims) to 0.39 (argument components).

5.3 Evaluation

To evaluate whether the use of argumentation structures correlates with holistic scores, we estimated a linear regression model using the nine argument features on the training data and evaluated on the test data. We use Cohen's kappa, as well as Pearson's correlation and quadratically-weighted kappa, the latter two being standard measures in essay scoring literature (Shermis and Burstein, 2013). Row "Arg" in Table 3 shows the results; argument structures have a moderate positive relationship with holistic scores.

More extensive use of argumentation structures is thus correlated with overall quality of an argumentative essay. However, argumentative fluency specifically is difficult to disentangle from fluency in language production in general manifested through the sheer length of the essay. In a timed test, a more fluent writer will be able to write more. To examine whether fluency in argumentation structures can explain additional variance in scores beyond that explained by general fluency (as approximated through the number of words in an essay), we estimated a length-only based linear regression model as well as a model that uses all the 9 argument structure features in addition to length. As shown in Table 3, the addition of argumentation structures yields a small improvement across all measures over a length-only model.

Model	κ	r	qwk
Arg	.195	.389	.344
Len	.365	.605	.518
Arg + Len	.389	.614	.540

Table 3: Prediction of holistic scores using argument structure features (Arg), length (Len), and argument structure features and length (Arg+Len). "qwk" stands for quadratically weighted kappa.

6 Conclusion & Future Work

In this paper, we set out to investigate the relationship between argumentation structures, argument content, and the quality of the essay. Our experiments suggest that (a) more extensive use of argumentation structures is predictive of better quality of argumentative writing, beyond overall fluency in language production; and (b) structure-based detection of argumentation is a possible fallback strategy to approximate argumentative *content* if an automated argument detection system is to generalize to new prompts. The two findings together suggest that the structure-based approach is a promising avenue for research in argumentation-aware automated writing evaluation.

In future work, we intend to improve the structure-based approach by identifying characteristics of argument components that are too general and so cannot be taken as evidence of germane, case-specific argumentation on the student's part (claims like "*More information is needed*"), as well as study properties of seemingly non-argumentative sentences that neverthe-

less have a potential for argumentative use in context (such as asking fact-seeking questions). We believe this would allow pushing the envelope of structure-based analysis towards identification of arguments that have a higher likelihood of being effective.

Acknowledgments

The work of Christian Stab and Iryna Gurevych has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus project AWS under grant No. 01|S12054.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 5(4):363–381.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Jill Burstein, Joel Tetreault, and Nitin Madnani. 2013. The e-rater automated essay scoring system. In M. Shermis and J Burstein, editors, *Handbook of Automated Essay Evaluation: Current Applications and Future Directions*. New York: Routledge.

Ralph Ferretti, Charles MacArthur, and Nancy Dowdy. 2000. The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, 93:694–702.

Marion Goldstein, Amanda Crowell, and Deanna Kuhn. 2009. What constitutes skilled argumentation and how does it develop? *Informal Logic*, 29:379–395.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP ' 15, pages 2127–2137, Lisbon, Portugal.

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, pages 76–81, Philadelphia, PA, USA.

Hyojung Lim and Jimin Kahng. 2012. Review of Criterion®. *Language Learning & Technology*, 16(2):38–45.

Raquel Mochales-Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, Barcelona, Spain.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, Stanford, CA, USA.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, MA, USA.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, page (to appear), Lisbon, Portugal.

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)*, ACL '15, pages 543–552, Beijing, China.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 440–450, Lisbon, Portugal.

Mark D. Shermis and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and Future Directions*. Routledge Chapman & Hall.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, MA, USA.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International*

Conference on Computational Linguistics, COLING '14, pages 1501–1510, Dublin, Ireland, August.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *arXiv preprint arXiv:1604.07370*.

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

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