Annotating Student Talk in Text-based Classroom Discussions

Luca Lugini, Diane Litman, Amanda Godley, Christopher Olshefski
University of Pittsburgh
Pittsburgh, PA 15260
{lul32,dlitman,agodley,cao48}@pitt.edu

Abstract

Classroom discussions in English Language Arts have a positive effect on students' reading, writing, and reasoning skills. Although prior work has largely focused on teacher talk and student-teacher interactions, we focus on three theoretically-motivated aspects of high-quality student talk: argumentation, specificity, and knowledge domain. We introduce an annotation scheme, then show that the scheme can be used to produce reliable annotations and that the annotations are predictive of discussion quality. We also highlight opportunities provided by our scheme for educational and natural language processing research.

1 Introduction

Current research, theory, and policy surrounding K-12 instruction in the United States highlight the role of student-centered disciplinary discussions (i.e. discussions related to a specific academic discipline or school subject such as physics or English Language Arts) in instructional quality and student learning opportunities (Danielson, 2011; Grossman et al., 2014). Such student-centered discussions – often called "dialogic" or "inquiry-based" – are widely viewed as the most effective instructional approach for disciplinary understanding, problem-solving, and literacy (Elizabeth et al., 2012; Engle and Conant, 2002; Murphy et al., 2009). In English Language Arts (ELA) classrooms, student-centered discussions about literature have a positive impact on the development of students' reasoning, writing, and reading skills (Applebee et al., 2003; Reznitskaya and Gregory, 2013). However, most studies have focused on the role of teachers and their talk (Bloome et al., 2005; Elizabeth et al., 2012; Michaels et al., 2008) rather than on the aspects of student talk that contribute to discussion quality.

Additionally, studies of student-centered discussions rarely use the same coding schemes, making it difficult to generalize across studies (Elizabeth et al., 2012; Soter et al., 2008). This limitation is partly due to the time-intensive work required to analyze discourse data through qualitative methods such as ethnography and discourse analysis. Thus, qualitative case studies have generated compelling theories about the specific features of student talk that lead to high-quality discussions, but few findings can be generalized and leveraged to influence instructional improvements across ELA classrooms.

As a first step towards developing an automated system for detecting the features of student talk that lead to high quality discussions, we propose a new annotation scheme for student talk during ELA "text-based" discussions - that is, discussions that center on a text or piece of literature (e.g., book, play, or speech). The annotation scheme was developed to capture three aspects of classroom talk that are theorized in the literature as important to discussion quality and learning opportunities: *argumentation* (the process of systematically reasoning in support of an idea), *specificity* (the quality of belonging or relating uniquely to a particular subject), and *knowledge domain* (area of expertise represented in the content of the talk). We demonstrate the reliability and validity of our scheme via an annotation study of five transcripts of classroom discussion.

2 Related Work

One discourse feature used to assess the quality of discussions is students' argument moves: their claims about the text, their sharing of textual evidence for claims, and their warranting or reasoning to support the claims (Reznitskaya et al., 2009; Toulmin, 1958). Many researchers view student

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 110–116
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

reasoning as of primary importance, particularly when the reasoning is elaborated and highly inferential (Kim, 2014). In Natural Language Processing (NLP), most educationally-oriented argumentation research has focused on corpora of student persuasive essays (Ghosh et al., 2016; Klebanov et al., 2016; Persing and Ng, 2016; Wachsmuth et al., 2016; Stab and Gurevych, 2017; Nguyen and Litman, 2018). We instead focus on multi-party spoken discussion transcripts from classrooms. A second key difference consists in the inclusion of the warrant label in our scheme, as it is important to understand how students explicitly use reasoning to connect evidence to claims.

Educational studies suggest that discussion quality is also influenced by the specificity of student talk (Chisholm and Godley, 2011; Sohmer et al., 2009). Chisholm and Godley found that as specificity increased, the quality of students' claims and reasoning also increased. Previous NLP research has studied specificity in the context of professionally written newspaper articles (Li and Nenkova, 2015; Li et al., 2016; Louis and Nenkova, 2011, 2012). While the annotation instructions used in these studies work well for general purpose corpora, specificity in text-based discussions also needs to capture particular relations between discussions and texts. Furthermore, since the concept of a sentence is not clearly defined in speech, we annotate argumentative discourse units rather than sentences (see Section 3).

The knowledge domain of student talk may also matter, that is, whether the talk focuses on disciplinary knowledge or lived experiences. Some research suggests that disciplinary learning opportunities are maximized when students draw on evidence and reasoning that are commonly accepted in the discipline (Resnick and Schantz, 2015), although some studies suggest that evidence or reasoning from lived experiences increases discussion quality (Beach and Myers, 2001). Previous related work in NLP analyzed evidence type for argumentative tweets (Addawood and Bashir, 2016). Although the categories of evidence type are different, their definition of evidence type is in line with our definition of knowledge domain. However, our research is distinct from this research in its application domain (i.e. social media vs. education) and in analyzing knowledge domain for all argumentative components, not only those containing claims.

3 Annotation Scheme

Our annotation scheme[1] uses argument moves as the unit of analysis. We define an argument move as an utterance, or part of an utterance, that contains an argumentative discourse unit (ADU) (Peldszus and Stede, 2013). Like Peldszus and Stede (2015), in this paper we use transcripts already segmented into argument moves and focus on the steps following segmentation, i.e., labeling argumentation, specificity, and knowledge domain. Table 1 shows a section of a transcribed classroom discussion along with labels assigned by a human annotator following segmentation.

3.1 Argumentation

The argumentation scheme is based on (Lee, 2006) and consists of a simplified set of labels derived from Toulmin's (1958) model: (i) *Claim*: an arguable statement that presents a particular interpretation of a text or topic. (ii) *Evidence*: facts, documentation, text reference, or testimony used to support or justify a claim. (iii) *Warrant*: reasons explaining how a specific evidence instance supports a specific claim. Our scheme specifies that warrants must come after claim and evidence, since by definition warrants cannot exist without them.

The first three moves in Table 1 show a natural expression of an argument: a student first claims that Willy's wife is only trying to protect him, then provides a reference as evidence by mentioning something she said to her kids at the end of the book, and finally explains how not caring about her kids ties the evidence to the initial claim. The second group shows the same argument progression, with evidence given as a direct quote.

3.2 Specificity

Specificity annotations are based on (Chisholm and Godley, 2011) and have the goal of capturing text-related characteristics expressed in student talk. Specificity labels are directly related to four distinct elements for an argument move: (1) it is specific to one (or a few) character or scene; (2) it makes significant qualifications or elaborations; (3) it uses content-specific vocabulary (e.g. quotes from the text); (4) it provides a chain of reasons. Our annotation scheme for specificity includes three labels along a linear scale: (i) *Low*:

[1] The coding manual is in the supplemental material.

Move	Stu	Argument Move	Argument	Specificity	Domain
23	S1	She’s like really just protecting Willy from everything.	claim	medium	disciplinary
24	S1	Like at the end of the book remember how she was telling the kids to leave and never come back.	evidence	medium	disciplinary
25	S1	Like she’s not even caring about them, she’s caring about Willy.	warrant	medium	disciplinary
41	S2	It’s like she’s concerned with him trying to [inaudible] and he’s concerned with trying to make her happy, you know? So he feels like he’s failing when he’s not making her happy like	claim	high	disciplinary
42	S2	”Let’s bring your mother some good news”	evidence	high	disciplinary
43	S2	but she knew that, there wasn’t any good news, so she wanted to act happy so he wouldn’t be in pain.	warrant	high	disciplinary
55	S3	Some people they just ask for a job is just like, some money.	evidence	low	experiential

Table 1: Examples of argument moves and their respective annotations from a discussion of the book *Death of a Salesman*. As shown by the argument move numbers, boxes for students S1, S2, and S3 indicate separate, non contiguous excerpts of the discussion.

statement that does not contain any of these elements. (ii) *Medium*: statement that accomplishes one of these elements. (iii) *High*: statement that clearly accomplishes at least two specificity elements. Even though we do not explicitly use labels for the four specificity elements, we found that explicitly breaking down specificity into multiple components helped increase reliability when training annotators.

The first three argument moves in Table 1 all contain the first element, as they refer to select characters in the book. However, no content-specific vocabulary, clear chain of reasoning, or significant qualifications are provided; therefore all three moves are labeled as medium specificity. The fourth move, however, accomplishes the first and fourth specificity elements, and is labeled as high specificity. The fifth move is also labeled high specificity since it is specific to one character/scene, and provides a direct quote from the text. The last move is labeled as low specificity as it reflects an overgeneralization about all humans.

3.3 Knowledge Domain

The possible labels for knowledge domain are: (i) *Disciplinary*: the statement is grounded in knowledge gathered from a text (either the one under discussion or others), such as a quote or a description of a character/event. (ii) *Experiential*: the statement is drawn from human experience, such as what the speaker has experienced or thinks that other humans have experienced.

In Table 1 the first six argument moves are labeled as disciplinary, since the moves reflect knowledge from the text currently being discussed. The last move, however, draws from a student’s experience or perceived knowledge about the real world.

4 Reliability and Validity Analyses

We carried out a reliability study for the proposed scheme using two pairs of expert annotators, P1 and P2. The annotators were trained by coding one transcript at a time and discussing disagreements. Five text-based discussions were used for testing reliability after training: pair P1 annotated discussions of *The Bluest Eye*, *Death of a Salesman*, and *Macbeth*, while pair P2 annotated two separate discussions of *Ain’t I a Woman*. 250 argument moves (discussed by over 40 students and consisting of over 8200 words) were annotated. Inter-rater reliability was assessed using Cohen’s kappa:

Moves	Argumentation (kappa)	Specificity (qwkappa)	Domain (kappa)
169	0.729	0.874	0.980
81	0.725	0.930	1

Table 2: Inter-rater reliability for pairs P1 and P2.

Argumentation	evidence	warrant	claim
evidence	25	5	0
warrant	6	92	12
claim	0	2	27
Specificity	**low**	**medium**	**high**
low	59	5	3
medium	5	25	2
high	1	6	63
Knowledge Domain	**disciplinary**	**experiential**	
disciplinary	138	1	
experiential	0	30	

Table 3: Confusion matrices for argumentation, specificity, and knowledge domain, for annotator pair P1.

unweighted for argumentation and knowledge domain, but quadratic-weighted for specificity given its ordered labels.

Table 2 shows that kappa for argumentation ranges from $0.61 - 0.8$, which generally indicates substantial agreement (McHugh, 2012). Kappa values for specificity and knowledge domain are in the $0.81 - 1$ range which generally indicates almost perfect agreement (McHugh, 2012). These results show that our proposed annotation scheme can be used to produce reliable annotations of classroom discussion with respect to argumentation, specificity, and knowledge domain.

Table 3 shows confusion matrices[2] for annotator pair P1 (we observed similar trends for P2). The argumentation section of the table shows that the largest number of disagreements happens between the claim and warrant labels. One reason may be related to the constraint we impose on warrants - they require the existence of a claim and evidence. If a student tries to provide a warrant for a claim that happened much earlier in the discussion, the annotators might interpret the warrant as new claim. The specificity section shows relatively few low-high label disagreements as compared to low-med and med-high. This is also reflected in the quadratic-weighted kappa as low-high disagreements will carry a larger penalty (unweighted kappa is 0.797). The main reasons for disagreements over specificity labels come from two of the four specificity elements discussed in Section 3.2: whether an argument move is related to one character or scene, and whether it provides a chain of reasons. With respect to the first of these two elements we observed disagreements in argument moves containing pronouns with an ambiguous reference. Of particular note is the pronoun *it*. If we consider the argument move *"I mean even if you know you have a hatred towards a standard or whatever, you still don't kill it"*, the pronoun *it* clearly refers to something within the move (i.e. the standard) that the student themselves mentioned. In contrast, for argument moves such as *"It did happen"* it might not be clear to what previous move the pronoun refers, therefore creating confusion on whether this specificity element is accomplished. Regarding specificity element (4) we found that it was easier to determine the presence of a chain of reasons when discourse connectives (e.g. because, therefore) were present in the argument move. The absence of explicit discourse connectives in an argument move might drive annotators to disagree on the presence/absence of a chain of reasons, which is likely to result in a different specificity label. Additionally, annotators found that shorter turns at talk proved harder to annotate for specificity. Finally, as we can see from the third section in the table, knowledge domain has the lowest disagreements with only one.

We also (Godley and Olshefski, 2017) explored the validity of our coding scheme by comparing our annotations of student talk to English Education experts' evaluations (quadratic-weighted kappa of 0.544) of the discussion's quality. Using stepwise regressions, we found that the best model of discussion quality (R-squared of 0.432) included all three of our coding dimensions: argumentation, specificity, and knowledge domain.

5 Opportunities and Challenges

Our annotation scheme introduces opportunities for the educational community to conduct futher research on the relationship between features of student talk, student learning, and discussion quality. Although Chisholm and Godley (2011) and we found relations between our coding constructs and

[2] The class distributions for argumentation and specificity labels vary significantly across transcripts, as can be seen in (Lugini and Litman, 2017) and (Godley and Olshefski, 2017).

discussion quality, these were small-scale studies based on manual annotations. Once automated classifiers are developed, such relations between talk and learning can be examined at scale. Also, automatic labeling via a standard coding scheme can support the generalization of findings across studies, and potentially lead to automated tools for teachers and students.

The proposed annotation scheme also introduces NLP opportunities and challenges. Existing systems for classifying specificity and argumentation have largely been designed to analyze written text rather than spoken discussions. This is (at least in part) due to a lack of publicly available corpora and schemes for annotating argumentation and specificity in spoken discussions. The development of an annotation scheme explicitly designed for this problem is the first step towards collecting and annotating corpora that can be used by the NLP community to advance the field in this particular area. Furthermore, in text-based discussions, NLP methods need to tightly couple the discussion with contextual information (i.e., the text under discussion). For example, an argument move from one of the discussions mentioned in Section 4 stated *"She's saying like free like, I don't have to be, I don't have to be this salesman's wife anymore, your know? I don't have to play this role anymore."* The use of the term *salesman* shows the presence of specificity element (3) (see Section 3.2) because the text under discussion is indeed *Death of a Salesman*. If the students were discussing another text, the mention of the term *salesman* would not indicate one of the specificity elements, therefore lowering the specificity rating. Thus, using existing systems is unlikely to yield good performance. In fact, we previously (Lugini and Litman, 2017) showed that while using an off-the-shelf system for predicting specificity in newspaper articles resulted in low performance when applied to classroom discussions, exploiting characteristics of our data could significantly improve performance. We have similarly evaluated the performance of two existing argument mining systems (Nguyen and Litman, 2018; Niculae et al., 2017) on the transcripts described in Section 4. We noticed that since the two systems were trained to classify only claims and premises, they were never able to correctly predict warrants in our transcripts. Additionally, both systems classified the overwhelming majority of moves as premise, resulting in negative kappa in some cases. Using our scheme to create a corpus of classroom discussion data manually annotated for argumentation, specificity, and knowledge domain will support the development of more robust NLP prediction systems.

6 Conclusions

In this work we proposed a new annotation scheme for three theoretically-motivated features of student talk in classroom discussion: argumentation, specificity, and knowledge domain. We demonstrated usage of the scheme by presenting an annotated excerpt of a classroom discussion. We demonstrated that the scheme can be annotated with high reliability and reported on scheme validity. Finally, we discussed some possible applications and challenges posed by the proposed annotation scheme for both the educational and NLP communities. We plan to extend our annotation scheme to label information about collaborative relations between different argument moves, and release a corpus annotated with the extended scheme.

Acknowledgements

We want to thank Haoran Zhang, Tazin Afrin, and Annika Swallen for their contribution, and all the anonymous reviewers for their helpful suggestions.

This work was supported by the Learning Research and Development Center at the University of Pittsburgh.

References

Aseel Addawood and Masooda Bashir. 2016. ”what is your evidence?” a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11.

Arthur N Applebee, Judith A Langer, Martin Nystrand, and Adam Gamoran. 2003. Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school english. *American Educational Research Journal*, 40(3):685–730.

Richard Beach and Jamie Myers. 2001. *Inquiry-based English instruction: Engaging students in life and literature*, volume 55. Teachers College Press.

David Bloome, Stephanie Power Carter, Beth Morton Christian, Sheila Otto, and Nora Shuart-Faris. 2005. *Discourse analysis and the study of classroom* *language and literacy events: A microethnographic perspective*. Lawrence Erlbaum.

James S Chisholm and Amanda J Godley. 2011. Learning about language through inquiry-based discussion: Three bidialectal high school students talk about dialect variation, identity, and power. *Journal of Literacy Research*, 43(4):430–468.

Charlotte Danielson. 2011. Evaluations that help teachers learn. *Educational leadership*, 68(4):35–39.

Tracy Elizabeth, Trisha L Ross Anderson, Elana H Snow, and Robert L Selman. 2012. Academic discussions: An analysis of instructional discourse and an argument for an integrative assessment framework. *American Educational Research Journal*, 49(6):1214–1250.

Randi A Engle and Faith R Conant. 2002. Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, 20(4):399–483.

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

Amanda Godley and Christopher Olshefski. 2017. The role of argument moves, specificity and evidence type in meaningful literary discussions across diverse secondary classrooms. Unpublished paper presented at Literacy Research Association 67th Annual Conference: Literacy Research for Expanding Meaningfulness.

Pam Grossman, Julie Cohen, Matthew Ronfeldt, and Lindsay Brown. 2014. The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, 43(6):293–303.

Il-Hee Kim. 2014. Development of reasoning skills through participation in collaborative synchronous online discussions. *Interactive Learning Environments*, 22(4):467–484.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75.

Carol D Lee. 2006. every good-bye aint gone: analyzing the cultural underpinnings of classroom talk. *International Journal of Qualitative Studies in Education*, 19(3):305–327.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, pages 2281–2287.

Junyi Jessy Li, Bridget ODaniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Annie Louis and Ani Nenkova. 2011. General versus specific sentences: automatic identification and application to analysis of news summaries. Technical Report MS-CIS-11-07, University of Pennsylvania.

Annie Louis and Ani Nenkova. 2012. A corpus of general and specific sentences from news. In *LREC*, pages 1818–1821.

Luca Lugini and Diane Litman. 2017. Predicting specificity in classroom discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Sarah Michaels, Catherine OConnor, and Lauren B Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297.

P Karen Murphy, Ian AG Wilkinson, Anna O Soter, Maeghan N Hennessey, and John F Alexander. 2009. Examining the effects of classroom discussion on students comprehension of text: A meta-analysis. *Journal of Educational Psychology*, 101(3):740.

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

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument Mining with Structured SVMs and RNNs. In *Proceedings of ACL*.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394.

Lauren B Resnick and Faith Schantz. 2015. Talking to learn: The promise and challenge of dialogic teaching. *Socializing Intelligence through academic talk and dialogue*, pages 441–450.

Alina Reznitskaya and Maughn Gregory. 2013. Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist*, 48(2):114–133.

Alina Reznitskaya, Li-Jen Kuo, Ann-Marie Clark, Brian Miller, May Jadallah, Richard C Anderson, and Kim Nguyen-Jahiel. 2009. Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education*, 39(1):29–48.

Richard Sohmer, Sarah Michaels, MC OConnor, and Lauren Resnick. 2009. Guided construction of knowledge in the classroom. *Transformation of knowledge through classroom interaction*, pages 105–129.

Anna O Soter, Ian A Wilkinson, P Karen Murphy, Lucila Rudge, Kristin Reninger, and Margaret Edwards. 2008. What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, 47(6):372–391.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Stephen Toulmin. 1958. *The uses of argument*. Cambridge: Cambridge University Press.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691.