Discussion Tracker: Supporting Teacher Learning about Students' Collaborative Argumentation in High School Classrooms

Luca Lugini and **Christopher Olshefski** and **Ravneet Singh** and **Diane Litman** and **Amanda Godley**

University of Pittsburgh
Pittsburgh, PA, USA

Abstract

Teaching collaborative argumentation is an advanced skill that many K-12 teachers struggle to develop. To address this, we have developed Discussion Tracker, a classroom discussion analytics system based on novel algorithms for classifying argument moves, specificity, and collaboration. Results from a classroom deployment indicate that teachers found the analytics useful, and that the underlying classifiers perform with moderate to substantial agreement with humans.

1 Introduction

Collaborative argumentation in student dialogue is essential to individual learning as well as group problem-solving (Reznitskaya and Gregory, 2013). Strong collaborative argumentation is characterized by specific claims, supporting evidence, and reasoning about that evidence as well as by building upon, questioning, and debating ideas posed by others. However, teaching collaborative argumentation is an advanced skill that many high school teachers struggle to develop (Lampert et al., 2010), partially due to the practical challenge of keeping track of important features of students' talk while managing class and reflecting on students' talk when no record of it exists.

To address this challenge, we have developed Discussion Tracker (DT), a system that leverages natural language processing (NLP) to provide teachers with automatically generated data about three important dimensions of students' collaborative argumentation: argument moves, specificity and collaboration. Discussion Tracker includes visualizations, interactive coded transcripts, collaboration maps, analytics across discussions, and instructional planning. In contrast to teacher dashboards which largely focus on discussion analytics such as amount of student/teacher talk, teacher wait time, and teacher question type (Chen et al., 2014; Gerritsen et al., 2018; Pehmer et al., 2015; Blanchard et al., 2016), DT focuses on students' collaborative argumentation. In contrast to related NLP algorithms which largely focus on coding student essays (Ghosh et al., 2016; Klebanov et al., 2016; Nguyen and Litman, 2016), asynchronous online discussions (Swanson et al., 2015), and news articles (Li and Nenkova, 2015), DT's NLP algorithms address the challenges of coding transcripts of synchronous, face-to-face classroom discussions.

2 Description of Discussion Tracker (DT)

To use DT, a teacher first uploads a classroom discussion transcript. Next, NLP classifiers code the transcript using a previously developed scheme for representing three important dimensions of collaborative argumentation (Lugini et al., 2018; Olshefski et al., 2020): argument moves (claim, evidence, explanation), specificity (low, medium, high), and collaboration (new, agree, extension, challenge/probe). Student turns are the unit of analysis for collaboration. Argumentative Discourse Units (ADUs) — either entire turns, or segments within turns — are the argumentation and specificity units of analysis.

Each NLP classifier in DT was developed by training on a previously collected and freely available corpus[1] of collaborative argumentation (Olshefski et al., 2020) using transformer-based neural networks.

[1]http://discussiontracker.cs.pitt.edu/ - we refer to this as corpus C1.

Proceedings of the 27th International Conference on Computational Linguistics, pages 53–58
Barcelona, Spain (Online), December 12, 2020.

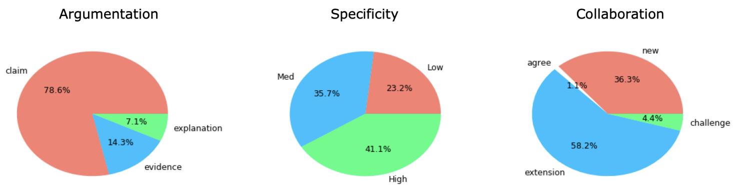

Figure 1: Partial screenshot of “Overview” page in Discussion Tracker.

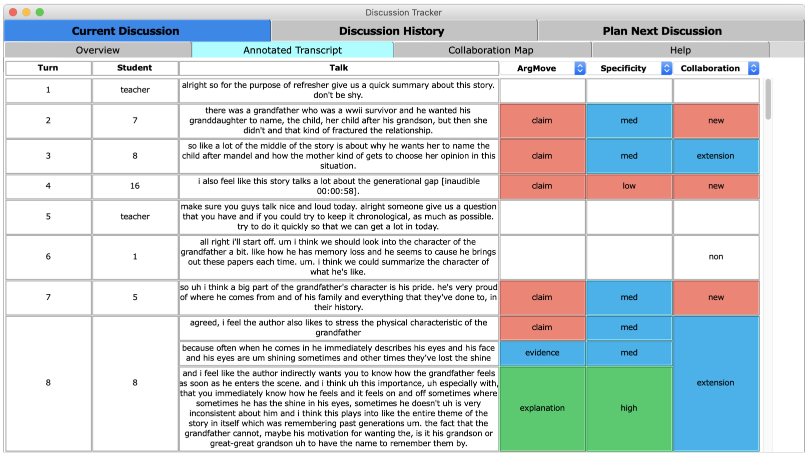

Figure 2: Screenshot of “Annotated Transcript” page in Discussion Tracker.

A pretrained BERT model (Devlin et al., 2019; Wolf et al., 2019) is used to generate word embeddings for each word in an ADU (or turn, for collaboration). An average pooling layer is then used to compute the final embedding for the target ADU. For predicting specificity, a softmax classifier is applied to the target ADU embedding. For predicting argument moves, the target ADU as well as a window of surrounding ADUs are embedded, then concatenated to form the final feature vector. A softmax layer is applied on top of the feature vector to complete the argument move classifier. This improves our prior argumentation models (Lugini and Litman, 2018) by using a pre-trained neural network and adding context information (Lugini and Litman, 2020). The collaboration classifier is slightly more complex since collaboration labels depend on the relationship between a target turn and a particular reference turn. For the purpose of this work we assume that the target turn is already provided in the input transcript. A pretrained BERT model and average pooling layer are used to generate embeddings for the target and reference turns. An element-wise multiplication between the two embeddings is performed, yielding the feature vector used by a softmax classifier.

All models use the *bert-base-uncased* BERT variant from the HuggingFace (Wolf et al., 2019) library, which results in the smallest available dimensionality to keep computational complexity to a minimum. The three models were built using the Keras library (Chollet and others, 2015). The *Adam* optimizer was used, as well as early stopping to automatically determine the number of epochs for training by monitoring validation loss (the validation set was chosen randomly and consisted of 10% of the initial training set for each fold).

After classification, all discussion analytics are automatically generated from the NLP codes. The DT overview screen (Figure 1) includes pie charts indicating the distribution of the codes for students' argument moves, specificity, and collaboration. Other screens include interactive coded transcripts (Figure 2), collaboration maps (Figure 3), identification of strengths and weaknesses to support teacher goal-setting (Figure 4), and a history page (not shown) that compares the code distributions across discussions.

We initially implemented a desktop version of DT using Python and Tkinter. The screenshots in the figures and the usability evaluation below are based on this version. To make DT more portable across hardware and to allow teachers to easily use DT on multiple machines (e.g., school, home), we now have

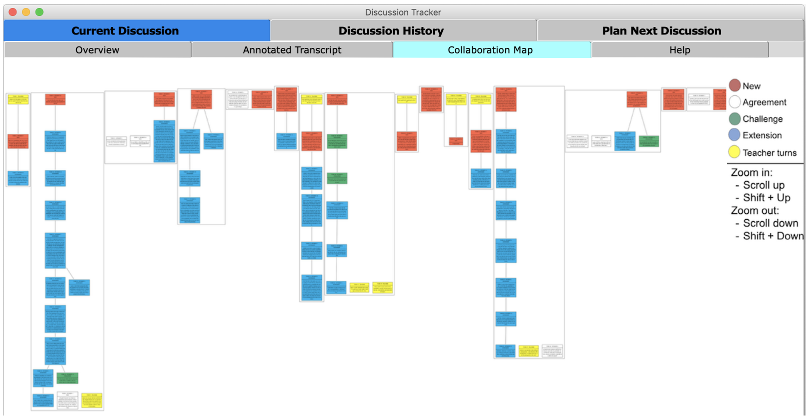

Figure 3: Screenshot of "Collaboration Map" page in Discussion Tracker.

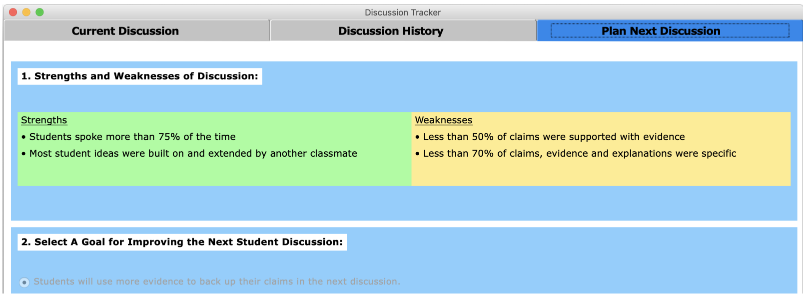

Figure 4: Partial screenshot of "Plan Next Discussion" page in Discussion Tracker.

a web version of DT[2]. This version is implemented in Python and uses the REMI package[3] to convert Python into HTML and launch a webserver to accept requests for the site and handle user input. With this setup it is easy to integrate the classifiers, implemented as a REST API on the same server hosting .

3 Evaluation

From January to March 2020, we collected data (corpus C2) to evaluate both teacher perceptions of DT as well as NLP classifier performance. In particular, the desktop version of DT was used by 18 high school English Language Arts teachers from 4 schools, where: 1) each teacher led a discussion about a literary text that was audio-recorded and observed by a researcher, 2) each teacher completed an online survey within a day, 3) experienced annotators[4] hand-coded transcripts of the discussion for the three dimensions of collaborative argumentation discussed above and uploaded them into the DT system, 4) within two weeks, researchers conducted a 45-minute cognitive interview (Voet and Wever, 2017) with each teacher while they were using DT to look at their students' discussion[5], and 5) the same day, teachers completed a second survey that mirrored the first with additional items for ratings of DT.

DT Usability. We measured teachers' perceptions of the overall usefulness of DT and of specific features/visualizations through Likert-scale items on the survey from step 5 above. Survey items were based on Holden and Rada's (2011) teacher survey of perceived usability of technology. To remove noise that might distract from this usability evaluation, we evaluated DT under the best possible NLP conditions by using the manual codings of collaborative argumentation from step 3 above to generate all analytics. The NLP codings are separately evaluated in the classifier discussion below. Table 1 indicates that teachers perceived DT to be very helpful for their learning about facilitating collaborative argumentation. For nine of the 13 items, all teachers selected either "Agree" or "Strongly agree" (a mean score of 4.5), and no item received a "Strongly disagree." Although the item "I find the system easy to

[2]Web app demo link and details and source at `discussiontracker.cs.pitt.edu`
[3]`https://github.com/dddomodossola/remi`
[4]Kappa for argumentation (0.971) and collaboration (0.578) and Quadratic Weighted Kappa for specificity (0.813).
[5]Teachers navigated DT with minimal training (a 15-minute, face-to-face demo).

Question	Mean	Question	Mean
The overview of the discussion is helpful.	4.67	I find the system easy to use.	4.11
The pie charts of different features of the student discussion are helpful.	4.78	The system helps me to recognize my students' strengths during discussion.	4.72
The annotated transcript of student discussion is helpful.	4.89	The system helps me to recognize my students' weakness during discussion.	4.72
The collaboration diagram is helpful.	4.22	The system gives me more insight into student learning than I usually get from thinking about the discussion.	4.67
The system-generated strengths and weaknesses are helpful.	4.44	The system encourages me to make more changes to my facilitation of discussion than I usually do.	4.28
The goal-setting is helpful.	4.56	Overall, Discussion Tracker is helpful for my teaching of literature discussions.	4.72
The instructional resources are helpful.	4.17		

Table 1: Teacher survey items and Likert score means.

Annotation	Distribution
Argumentation	claim (72%), evidence (18%), explanation (10%)
Specificity	low (29%), medium (36%), high (36%)
Collaboration	new (22%), agree (3%), extensions (54%), challenge/probe (21%)

Table 2: Descriptive statistics of gold-standard annotations in test corpus.

Code	N	Kappa	Macro F	Micro F
Argument Move	1942 ADUs	0.574	0.730	0.789
Specificity	1942 ADUs	0.727	0.688	0.679
Collaboration	1467 Turns	0.566	0.439	0.775

Table 3: Transfomer-based neural classification results.

use" received the lowest score (4.11), all teachers either agreed or agreed strongly with the item. Other items that scored higher, however, varied more in responses. For example, although the majority of teachers agreed with "The collaboration diagram is helpful," three neither agreed nor disagreed.

NLP Classifier Performance. As the gold standard for evaluating DT classifier performance, we used the manual annotations from step 3 of the data collection discussed above. Table 2 shows the distribution of the gold-standard codes, while Table 3 shows classifier performance when compared to these gold-standards.[6] The results in Table 3 were obtained by training each classifier separately on corpus C1 (footnote 1) and testing on corpus C2 (the 18 discussions collected in this study). Hyperparameter optimization was performed using cross-validation on C1 in order to find out how much contextual information before/after the target ADU to consider (i.e. context window size). This yielded an argument classifier that added a window of 2 ADUs preceding and 2 ADUs following the target ADU for embedding. Though all classifiers show respectable results, predictions for argument move and specificity are more consistent for individual class labels, as evidenced by the small difference between macro and micro F-score. The lower macro F-score for collaboration is due to poor prediction performance for the agree and challenge/probe codes.

[6]The input for the gold-standard and automated coding was identical (loosely, a spreadsheet version of the first three columns in Figure 2). A professional service (rather than ASR) performed the audio transcription and segmentation into turns (for collaboration coding). A researcher further segmented student turns into ADUs (for argumentation and specificity coding).

4 Summary and Future Directions

In this work we described the development of a classroom analytics system and reported usability results from real world classroom deployment. We conducted a survey that showed teachers found the system easy to use and the analytics (based on human-annotated labels) helpful in analyzing collaborative argumentation. Evaluation of the automated NLP classifiers showed that they are in moderate to substantial agreement with the labels provided by human annotators. The main goal of future work is to continue to enhance our neural classification methods, and to develop an end-to-end, completely automated system. To this end, we will consider several aspects: perform new data collections and improve classifier performance; incorporate Automatic Speech Recognition to perform automated transcription; develop algorithms to automatically segment turns into ADUs. In addition, we will further develop the interface by addressing teacher feedback and improving the system's ease of use. Finally, teachers will evaluate our newer versions of DT, including versions where the analytics are based on classifier outputs rather than human-annotated labels.

Acknowledgements

This work was supported by the National Science Foundation (1842334 and 1917673) and by the University of Pittsburgh's Learning Research and Development Center as well as Center for Research Computing through the resources provided. We would like to thank the teachers who participated in this study.

References

Nathaniel Blanchard, Patrick J Donnelly, Andrew M Olney, Borhan Samei, Brooke Ward, Xiaoyi Sun, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. 2016. Identifying teacher questions using automatic speech recognition in classrooms. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 191.

G Chen, SN Clarke, and LB Resnick. 2014. An analytic tool for supporting teachers' reflection on classroom talk. In *Learning and becoming in practice: The International Conference of the Learning Sciences (ICLS) 2014*. International Society of the Learning Sciences.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, June.

David Gerritsen, John Zimmerman, and Amy Ogan. 2018. Towards a framework for smart classrooms that teach instructors to teach. In *International Conference of the Learning Sciences*, volume 3.

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)*, pages 549–554.

Heather Holden and Roy Rada. 2011. Understanding the influence of perceived usability and technology self-efficacy on teachers' technology acceptance. *Journal of Research on Technology in Education*, 43(4):343–367.

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, Berlin, Germany.

Magdalene Lampert, Heather Beasley, Hala Ghousseini, Elham Kazemi, and Megan Franke. 2010. Using designed instructional activities to enable novices to manage ambitious mathematics teaching. In *Instructional explanations in the disciplines*, pages 129–141. Springer.

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, January.

Luca Lugini and Diane Litman. 2018. Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67.

Luca Lugini and Diane Litman. 2020. Contextual argument component classification for class discussions. In *Proceedings s of the 28th International Conference on Computational Linguistics*, Online, December.

Luca Lugini, Diane Litman, Amanda Godley, and Christopher Olshefski. 2018. Annotating student talk in text-based classroom discussions. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 110–116.

Huy Nguyen and Diane Litman. 2016. Context-aware Argumentative Relation Mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany.

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. The Discussion Tracker corpus of collaborative argumentation. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 1033–1043, Marseille, France, May.

Ann-Kathrin Pehmer, Alexander Gröschner, and Tina Seidel. 2015. How teacher professional development regarding classroom dialogue affects students' higher-order learning. *Teaching and Teacher Education*, 47:108–119.

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.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226.

Michiel Voet and Bram De Wever. 2017. History teachers' knowledge of inquiry methods: An analysis of cognitive processes used during a historical inquiry. *Journal of Teacher Education*, 68(3):312–329.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.